6.034 Quizzes

12/15

by entire quiz - not section

Quiz 1

	Thorough understanding	Adequate understanding	
Problem 1 Rules	≥ 41	≥ 35	43
Problem 2 Search	≥ 39	≥ 34	41
Problem 3 Ideas	≥ 8	≥ 6	10
Overall	≥ 88	≥ 75	95

5

Quiz 2

	Thorough understanding	Adequate understanding	
Problem 1 Games	≥ 35	≥ 30	20
Problem 2 Constraints	≥ 35	≥ 31	21
Problem 3 Drawings	≥ 15	≥ 10	0
Overall	≥ 85	≥ 71	42

2

Quiz 3

	Thorough understanding	Adequate understanding	
Problem 1 Nearest neighbors/Classification Trees	≥ 35	≥ 30	27
Problem 2 Neural nets	≥ 35	≥ 29	38
Problem 3 Learning	≥ 14	≥ 10	4
Overall	≥ 84	≥ 69	70

4

Quiz 4

	Thorough understanding	Adequate understanding	
Problem 1 SVMs	≥ 39	≥ 34	30
Problem 2 Boosting	≥ 41	≥ 37	35
Problem 3 Representation	≥ 8	≥ 6	4
Overall	≥ 88	≥ 77	70

3

Quiz 5

Must take, Now

6.034 Final Prep

12/15

Need to study 2, 4, 5

possibly 3 - but prob not

Do Unit 5 first (never preped)

Probability from 6.041

└ how much do I remember?

1. JPT (Joint Prob Table)
2. Axioms
3. Conditional Prob
4. Chain Rule
5. Independence
6. Conditional Ind
7. Belief Nets

(2)

$P(a) =$ 

size a / size Ω

Union

$$P(a \cup b) = P(a) + P(b) - P(a \cap b)$$

$\cup$ union or

$\cap$ and

$P(a, b) = P(a \cap b)$ ~~:: so confusing!~~ ✓ Confirmed

notation

recognize

Conditional

$$P(a|b) = \frac{P(a \cap b)}{P(b)}$$

$$P(a|b)\,P(b) = P(a \cap b)$$

using above

did I recognize that before? Or need to think

(3)

$$P(a|b)\,P(b) = P(a \wedge b) = P(b|a)\,P(a)$$

Chain Rule

$$P(a \wedge b \wedge c) = P(a, b, c)$$
$$= P(a \mid b \wedge c)\,P(b \wedge c)$$
$$= P(a \mid b \wedge c)\,P(b|c)\,P(c)$$

I forgot this

But guess learned in 6.041

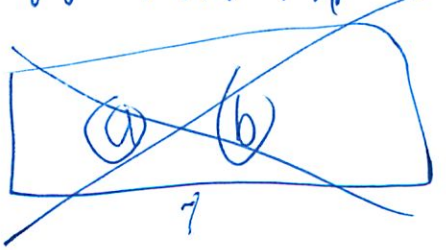
(Should not have sold book!)

I looked in 6.041 notes: called multiplication rule

Also Total Prob Thm $P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2)$

Independence

$P(a|b) = P(b)$ when ind



(4)

Conditional Ind

$$P(a \mid b \wedge z) = P(a \mid z)$$

So should be

~~$= P(a \mid b \wedge c)\ P(b \mid c)$~~

~~$= P(b \mid z)$~~

~~BH~~ $= \dfrac{P(a \wedge b \wedge z)}{P(b \wedge z)}$

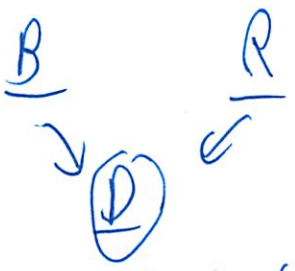
$= \dfrac{P(a \mid b \wedge z)\ \cancel{P(b \mid z)\ P(z)}}{\cancel{P(b \mid z)\ P(z)}}$

Same

So what should this normally be = to

But anyway b does not matter

(5) Dog bark example.



B	R	P(D)
T	T	.75
T	F	.5
F	T	.1
F	F	.01

$P(B) = .1$

$P(R) = .5$

Prob is cond. ind. of all non decendents

↳ they don't matter

Prob of everything together

$P(C, D, T, B, R) =$

i.e. all being true

$= P(C \wedge D \wedge T \wedge B \wedge R) =$

$= P(C \mid D \wedge T \wedge B \wedge R)\, P(D \mid T \wedge B \wedge R)\, P(T \mid B \wedge R)\, P(B \mid R) P(R)$

but some stuff cond ind

$= P(C \mid D)\, P(D \mid B \wedge R)\, P(T \mid R)\, P(B)\, P(R)$

(6)

So was the order that was listed is important?

Try otherwise

$P(B \cap R \cap T \cap (C \cap D)) =$

Then just say

$P(B)\ P(R)\ P(T|R)$

but it would have been harder to eliminate

Or would not have had the right things

basically just add its parents

↑ all above or just parents

↑ I believe

(also most are 2 level)

↑ Pretty sure

Basically net that only has things its dep on better than all items

↳ many of which its actually ind on

Figuring out what things actually depend on each is better

(7)

Lecture 2

JPT - table of all possible values

~~-~~ long + hard

Come on - get to questions

(do one now)

(not many variable)

Q: all right - but did in recitation

Oh did a few

Now study rest of stuff

Constrict values by using conditional prob

Oh he suggesting the model example

I saw earlier

Can use to tell stories

Which is more probable

Use hill climbing

Lecture 3 Nancy Kanwisher

a bit of history on brain regions

people agree on simple, broad regions

why we care

ways to investigate

Show faces to people in fMRI

Upside down very different

Strong respone on faces

Not other body parts

Specificity

- fMRI only sees millions of Neurons
- Can measure exact pt w/ probe in monkies
- Or study people w/ brain damage
- turn brain areas off w/ Transcranial Magnetic Stimulation
  - crude but works

(9)

- Genetic:
  - diff in identical twins
  - but 1-3 day infants good at face recognition
  - and monkies who never saw faces

Other specific areas:
- reading is new
- natural selection has not let area grow

- Can regions move over?

How do regions work together?

Naive Bayes – features are conditionally ind from each other

$$P(C, f_1 \cap f_2, \text{etc}) = \arg\max_C P(C) \cdot P(f_1 \cap f_2, \text{etc})$$

$$= \arg\max_C P(C) \cdot P(f_1 \mid C) \cdot P(f_2 \mid C) \cdot \text{etc}$$

(10)

Can use Bayes in Google translate

looks for most probable mix

└ split words into n-grams

P(blue | the sky is)

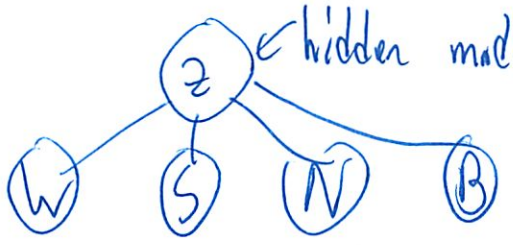
Google has 4-gram

5-6 grams for most pair of langs

Smooth a little bit so don't have 0 solutions

└ Robin Hood solution

Naive Bayes tries to build an observable world



This is not that arg max thing again

Right?

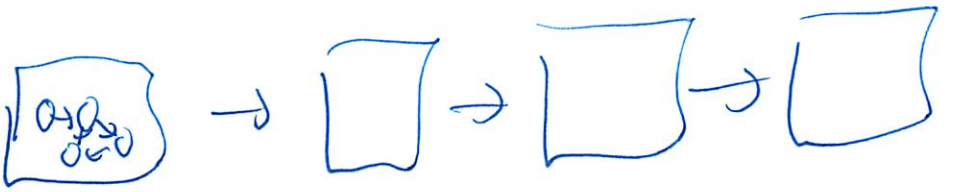
Lecture Right Way / 5 Hypotheses

What do ya think and why?

What makes us different from organisms at 200?

Chomsky Combine concepts w/o limit

Others Tying together a story



How do we solve AI problems?

1. Characterize behavior
2. Formulate computational problems
3. Propose computational solutions
4. Exploratory Implementation
5. Principles

Common sense lang
vs
Reflective lang

Cultural biases on Mac Beth story
revenge vs senseless violence
situational vs dispositional

Elaboration graph

Minskys' 6 levels of thinking

1. Self concias reflective thinking - what will others think about my revenge policy
2. Self reflective thinking - this will be revenge I don't do that
3. Reflective thinking - how I think about your thinking
4. Deliberate thinging - if I anger ya, you'll kill me
5. Learned reflex - if I kill you, your dead
6. Inate reflex

1. Story
2. Story-telling Story
   - how to generalize
3. Concept discovery
   - how to extract what is revenge from story?

Social Animal Hyp

We develop an ToM long since we are animals

Directed Perception Hyp

The mechanism that enables us humans to direct + hallucinate w/ our perceptual faculties seperate our intellegence from that of other primates

- cat drinking is diff from human drinking
- learning to recognize a jump
  - produce a video that does that

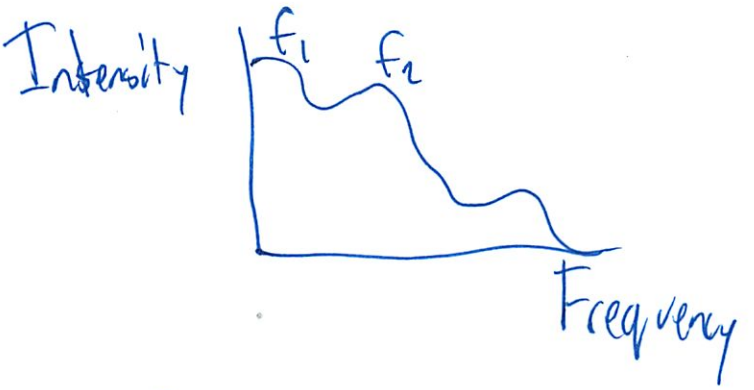
Last Lecture

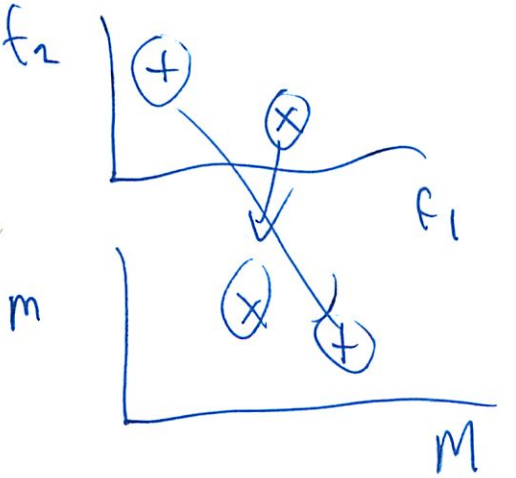
Exotic Engineering hyp

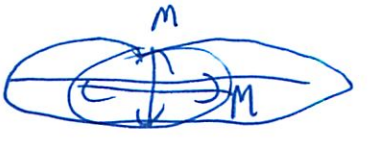
There is a kind of engineering in our heads we are nearly not aware of

Our brain is all over the place

Speech find peaks







15

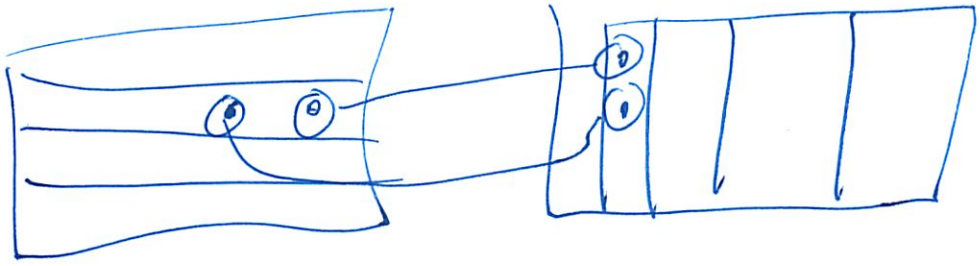
Cluster Left + Right

I don't get this

Said could be on final

Online resource is a 24 pg sci paper

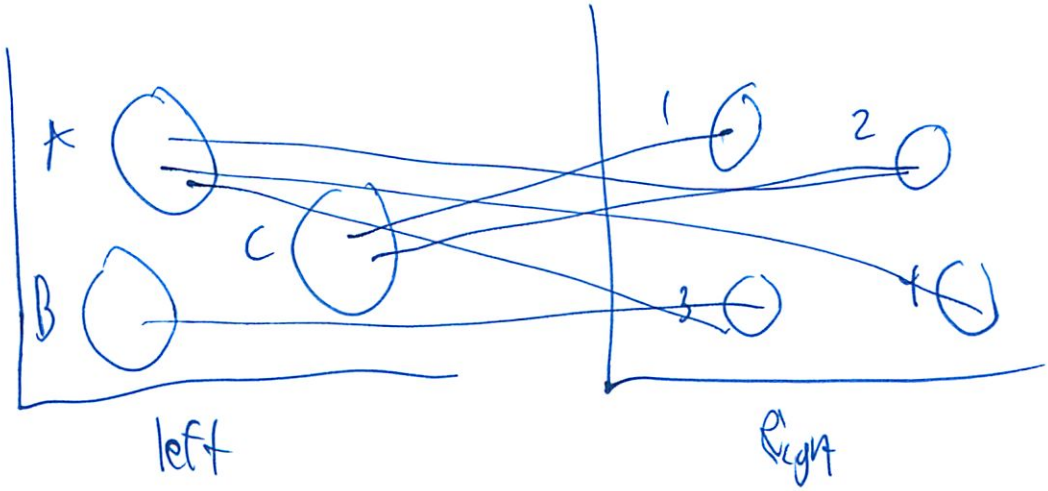
Lguess I should read it...



Paper: freq of formants

(wp: acoustic resonance of human speech

(Paper pretty useless!)



Clustering paper

Multimodal Dynamics:
Self-Supervised Learning in Perceptual and Motor Systems

by

Michael Harlan Coen

Submitted to the Department of Electrical Engineering and Computer Science on May 25, 2006 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

ABSTRACT

This thesis presents a self-supervised framework for perceptual and motor learning based upon correlations in different sensory modalities. The brain and cognitive sciences have gathered an enormous body of neurological and phenomenological evidence in the past half century demonstrating the extraordinary degree of interaction between sensory modalities during the course of ordinary perception. We develop a framework for creating artificial perceptual systems that draws on these findings, where the primary architectural motif is the cross-modal transmission of perceptual information to enhance each sensory channel individually. We present self-supervised algorithms for learning perceptual grounding, intersensory influence, and sensory-motor coordination, which derive training signals from internal cross-modal correlations rather than from external supervision. Our goal is to create systems that develop by interacting with the world around them, inspired by development in animals.

We demonstrate this framework with: (1) a system that learns the number and structure of vowels in American English by simultaneously watching and listening to someone speak. The system then cross-modally clusters the correlated auditory and visual data. It has no advance linguistic knowledge and receives no information outside of its sensory channels. This work is the first unsupervised acquisition of phonetic structure of which we are aware, outside of that done by human infants. (2) a system that learns to sing like a zebra finch, following the developmental stages of a juvenile zebra finch. It first learns the song of an adult male and then listens to its own initially nascent attempts at mimicry through an articulatory synthesizer. In acquiring the birdsong to which it was initially exposed, this system demonstrates self-supervised sensorimotor learning. It also demonstrates afferent and efferent equivalence – the system learns motor maps with the same computational framework used for learning sensory maps.

Thesis Supervisor: Whitman Richards
Title: Professor of Brain and Cognitive Sciences

Thesis Supervisor: Howard Shrobe
Title: Principal Research Scientist, EECS

We have sat around for hours and wondered how you look. If you have closed your senses upon silk, light, color, odor, character, temperament, you must be by now completely shriveled up. There are so many minor senses, all running like tributaries into the mainstream of love, nourishing it.
The Diary of Anais Nin (1943)

He plays by sense of smell.
Tommy, The Who (1969)

Chapter 1

Introduction

This thesis presents a unified framework for perceptual and motor learning based upon correlations in different sensory modalities. The brain and cognitive sciences have gathered a large body of neurological and phenomenological evidence in the past half century demonstrating the extraordinary degree of interaction between sensory modalities during the course of ordinary perception. We present a framework for artificial perceptual systems that draws on these findings, where the primary architectural motif is the cross-modal transmission of perceptual information to structure and enhance sensory channels individually. We present self-supervised algorithms for learning *perceptual grounding*, *intersensory influence*, and *sensorimotor coordination*, which derive training signals from internal cross-modal correlations rather than from external supervision. Our goal is to create perceptual and motor systems that develop by interacting with the world around them, inspired by development in animals.

Our approach is to formalize mathematically an insight in Aristotle's *De Anima* (350 B.C.E.), that *differences in the world are only detectable because different senses perceive the same world events differently*. This implies both that sensory systems need

A glossary of technical terms is contained in Appendix 1. Our usage of the word "sense" is defined in §1.5.

some way to share their different perspectives on the world and that they need some way to incorporate these shared influences into their own internal workings.

We begin with a computational methodology for *perceptual grounding*, which addresses the first question that any natural (or artificial) creature faces: *what different things in the world am I capable of sensing?* This question is deceptively simple because a formal notion of what makes things different (or the same) is non-trivial and often elusive. We will show that animals (and machines) can learn their perceptual repertoires by simultaneously correlating information from their different senses, even when they have no advance knowledge of what events these senses are individually capable of perceiving. In essence, by *cross-modally* sharing information between different senses, we demonstrate that sensory systems can be perceptually grounded by mutually bootstrapping off each other. As a demonstration of this, we present a system that learns the number (and formant structure) of vowels in American English, simply by watching and listening to someone speak and then cross-modally clustering the accumulated auditory and visual data. The system has no advance knowledge of these vowels and receives no information outside of its sensory channels. This work is the first unsupervised acquisition of phonetic structure of which we are aware, at least outside of that done by human infants, who solve this problem easily.

The second component of this thesis naturally follows perceptual grounding. Once an animal (or a machine) has learned the range of events it can detect in the world, *how does it know what it's perceiving at any given moment?* We will refer to this as *perceptual interpretation*. Note that grounding and interpretation are different things. By way of analogy to reading, one might say that *grounding* provides the dictionary and *interpretation* explains how to disambiguate among possible word meanings. More formally, grounding is an ontological process that defines what is perceptually knowable, and interpretation is an algorithmic process that describes how perceptions are categorized within a grounded system. We will present a novel framework for perceptual interpretation called *influence networks* (unrelated to a formalism know as *influence diagrams*) that blurs the distinctions between different sensory channels and allows them to influence one another while they are in the midst of perceiving. Biological perceptual

systems share cross-modal information routinely and opportunistically (Stein and Meredith 1993, Lewkowicz and Lickliter 1994, Rock 1997, Shimojo and Shams 2001, Calvert et al. 2004, Spence and Driver 2004); *intersensory influence* is an essential component of perception but one that most artificial perceptual systems lack in any meaningful way. We argue that this is among the most serious shortcomings facing them, and an engineering goal of this thesis is to propose a workable solution to this problem.

The third component of this thesis enables sensorimotor learning using the first two components, namely, perceptual grounding and interpretation. This is surprising because one might suppose that motor activity is fundamentally different than perception. However, we take the perspective that motor control can be seen as perception *backwards*. From this point of view, we imagine that – in a notion reminiscent of a Cartesian theater – an animal can "watch" the activity in its own motor cortex, as if it were a privileged form of *internal* perception. Then for any motor act, there are two associated perceptions – the *internal* one describing the generation of the act and the *external* one describing the self-observation of the act. The perceptual grounding framework described above can then *cross-modally ground* these internal and external perceptions with respect to one another. The power of this mechanism is that it can learn mimicry, an essential form of behavioral learning (see the developmental sections of Meltzoff and Prinz 2002) where one animal acquires the ability to imitate some aspect of another's activity, constrained by the capabilities and dynamics of its own sensory and motor systems. We will demonstrate sensorimotor learning in our framework with an artificial system that learns to sing like a zebra finch by first listening to a real bird sing and then by learning from its own initially uninformed attempts to mimic it.

This thesis has been motivated by surprising results about how animals process sensory information. These findings, gathered by the brain and cognitive sciences communities primarily over the past 50 years, have challenged century long held notions about how the brain works and how we experience the world in which we live. We argue that current approaches to building computers that perceive and interact with the real, human world are largely based upon developmental and structural assumptions, tracing back

several hundred years, that are no longer thought to be descriptively or biologically accurate. In particular, the notion that perceptual senses are in functional isolation – that they do not internally structure and influence each other – is no longer tenable, although we still build artificial perceptual systems as if it were.

1.1 Computational Contributions

This thesis introduces three new computational tools. The first is a mathematical model of *slices*, which are a new type of data structure for representing sensory inputs. Slices are topological manifolds that encode dynamic perceptual states and are inspired by surface models of cortical tissue (Dale et al. 1999, Fischl et al. 1999, Citti and Sarti 2003, Ratnanather et al. 2003). They can represent both symbolic and numeric data and provide a natural foundation for aggregating and correlating information. Slices represent the data in a perceptual system, but they are also *amodal*, in that they are not specific to any sensory representation. For example, we may have slices containing visual information and other slices containing auditory information, but it may not be possible to distinguish them further without additional information. In fact, we can equivalently represent either sensory or motor information within a slice. This generality will allow us to easily incorporate the learning of motor control into what is initially a perceptual framework.

The second tool is an algorithm for *cross-modal clustering*, which is an unsupervised technique for organizing slices based on their spatiotemporal correlations with other slices. These correlations exist because an event in the world is simultaneously – but differently – perceived through multiple sensory channels in an observer. The hypothesis underlying this approach is that the world has regularities – natural laws tend to correlate physical properties (Thompson 1917, Richards 1980, Mumford 2004) – and biological perceptory systems have evolved to take advantage of this. One may contrast this with mathematical approaches to clustering where some knowledge of the clusters, e.g., how many there are or their distributions, must be known a priori in order to derive them. Without knowing these parameters in advance, many algorithmic clustering techniques may not be robust (Kleinberg 2002, Still and Bialek 2004). Assuming that in many circumstances animals cannot know the parameters underlying their perceptual inputs,

how can they learn to organize their sensory perceptions? Cross-modal clustering answers this question by exploiting naturally occurring intersensory correlations.

The third tool in this thesis is a new family of models called *influence networks* (Figure 1.1). Influence networks use slices to interconnect independent perceptual systems, such as those illustrated in the classical view in Figure 1.1a, so they can influence one another during perception, as proposed in Figure 1.1b. Influence networks dynamically modify percepts within these systems to effect influence among their different components. The influence is designed to increase perceptual accuracy within individual perceptual channels by incorporating information from other co-occurring senses. More formally, influence networks are designed to move ambiguous perceptual inputs into easily recognized subsets of their representational spaces. In contrast with approaches taken in engineering what are typically called *multimodal systems*, influence networks are not intended to create high-level joint perceptions. Instead, they share sensory information across perceptual channels to increase local perceptual accuracy within the individual perceptual channels themselves. As we discuss in Chapter 6, this type of cross-modal perceptual reinforcement is ubiquitous in the animal world.

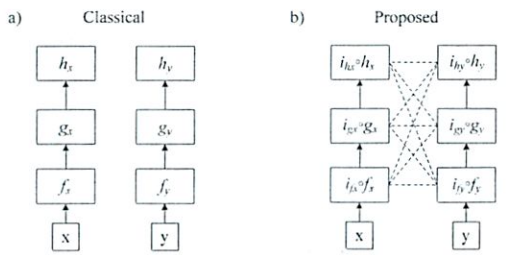


Figure 1.1– Adding an influence network to two preexisting systems. We start in (a) with two pipelined networks that independently compute separate functions. In (b), we compose on each function a corresponding *influence function*, which dynamically modifies its output based on activity at the other influence functions. The interaction among these influence functions is described by an *influence network*, which is defined in Chapter 5. The parameters describing this network can be found via unsupervised learning for a large class of perceptual systems, due to correspondences in the physical events that generate the signals they perceive and to the evolutionary incorporation of these regularities into the biological sensory systems that these computational systems model. Note influence networks are distinct from an unrelated formalism called influence diagrams.

1.2 Theoretic Contributions

The work presented here addresses several important problems. From an engineering perspective, it provides a principled, neurologically informed approach to building complex, interactive systems that can learn through their own experiences. In perceptual domains, it answers a fundamental question in mathematical clustering: *how should an unknown dataset be clustered?* The connection between clustering and perceptual grounding follows from the observation that learning to perceive is learning to organize perceptions into meaningful categories. From this perspective, asking what an animal can perceive is equivalent to asking how it should cluster its sensory inputs. This thesis presents a *self-supervised* approach to this problem, meaning our sub-systems derive feedback from one another cross-modally rather than rely on an external tutor such as a parent (or a programmer). Our approach is also highly nonparametric, in that it presumes neither that the number of clusters nor their distributions are known in advance, conditions which tend to defy other algorithmic techniques. The benefits of self-supervised learning in perceptual and motor domains are enormous because engineered approaches tend to be ad hoc and error prone; additionally, in sensorimotor learning we generally have no adequate models to specify the desired input/output behaviors for our systems. The notion of *programming by example* is nowhere truer than in the developmental mimicry widespread in animal kingdom (Meltzoff and Prinz 2002), and this work is a step in that direction for artificial sensorimotor systems.

Furthermore, this thesis suggests that not only do senses influence each other during perception, which is well established, it also proposes that *perceptual channels cooperatively structure their internal representations*. This mutual structuring is a basic feature in our approach to perceptual grounding. We argue, however, that it is not simply an epiphenomenon; rather, it is a fundamental component of perception itself, because *it provides representational compatibility for sharing information cross-modally* during higher-level perceptual processing. The inability to share perceptual data is one of the major shortcomings in current engineered approaches to building interactive systems.

Finally, within this framework, we will address three questions that are basic to developing a coherent understanding of cross-modal perception. They concern both

process and representation and raise the possibility that unifying (i.e. meta-level) principles might govern intersensory function:

1) Can the senses be perceptually grounded by bootstrapping off each other? Is shared experience sufficient for learning how to categorize sensory inputs?

2) How can seemingly different senses share information? What representational and computational restrictions does this place upon them?

3) Could the development of motor control use the same mechanism? In other words, can there be afferent and efferent equivalence in learning?

1.3 A Brief Motivation

The goal of this thesis is to propose a design for artificial systems that more accurately reflects how animal brains appear to process sensory inputs. In particular, we argue against *post-perceptual* integration, where the sensory inputs are separately processed in isolated, increasingly abstracted pipelines and then merged in a final integrative step as in Figure 1.2. Instead, we argue for *cross-modally integrated perception*, where the senses share information during perception that synergistically enhances them individually, as in Figure 1.1b. The main difficulty with the post-perceptual approach is that integration happens after the individual perceptions are generated. Integration occurs *after* each perceptual subsystem has already "decided" what it has perceived, when it is too late for intersensory influence to affect the individual, concurrent perceptions. This is due to information loss from both vector quantization and the explicit abstraction fundamental to the pipeline design. Most importantly, these approaches also preclude cooperative perceptual grounding; the bootstrapping provided by cross-modal clustering cannot occur when sensory systems are independent. These architectures are therefore also at odds with developmental approaches to building interactive systems.

Not only is the post-perceptual approach to integration biologically implausible from a scientific perspective, it is poor engineering as well. The real world is inherently multimodal in a way that most modern artificial perceptual systems do not capture or take advantage of. Isolating sensory inputs while they are being processed prevents the lateral sharing of information across perceptual channels, even though these sensory inputs are inherently linked by the physics of the world that generates them. Furthermore, we will argue that the co-evolution of senses within an individual species provided evolutionary pressure towards representational and algorithmic compatibilities essentially unknown in modern artificial perception. These issues are examined in detail in Chapters 6.

Our work is computationally motivated by Gibson (1950, 1987), who viewed perception as an external as well as an internal event, by Brooks (1986, 1991), who elevated perception onto an equal footing with symbolic reasoning, and by Richards (1988), who described how to exploit regularities in the world to make learning easier. The recursive use of a perceptual mechanism to enable sensorimotor learning in Chapter 4 is a result of our exposure to the ideas of Sussman and Abelson (1983).

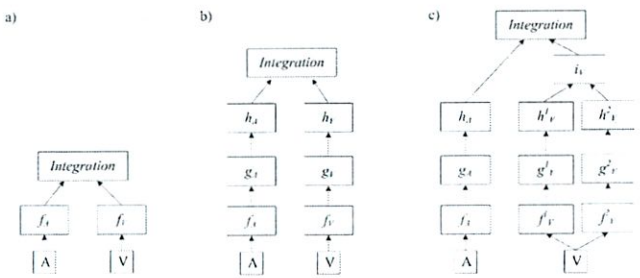


Figure 1.2 – Classical approaches to post-perceptual integration in traditional multimodal systems. Here, auditory (A) and visual (V) inputs pass through specialized unimodal processing pathways and are combined via an integration mechanism, which creates multimodal perceptions by extracting and reconciling data from the individual channels. Integration can happen earlier (a) or later (b). Hybrid architectures are also common. In (c), multiple pathways process the visual input and are pre-integrated before the final integration step; for example, the output of this preintegration step could be spatial localization derived solely through visual input. This diagram is modeled after (Stork and Hennecke 1996).

1.4 Demonstrations

The framework and its instantiation will be evaluated by a set of experiments that explore *perceptual grounding*, *perceptual interpretation*, and *sensorimotor learning*. These will be demonstrated with:

1) **Phonetic learning**: We present a system that learns the number and formant structure of vowels (monophthongs) in American English, simply by watching and listening to someone speak and then cross-modally clustering the accumulated auditory and visual data. The system has no advance knowledge of these vowels and receives no information outside of its sensory channels. This work is the first unsupervised machine acquisition of phonetic structure of which we are aware.

2) **Speechreading**: We incorporate an *influence network* into the cross-modally clustered slices obtained in Experiment 1 to increase the accuracy of perceptual classification within the slices individually. In particular, we demonstrate the ability of influence networks to move ambiguous perceptual inputs to unambiguous regions of their perceptual representational spaces.

3) **Learning birdsong**: We will demonstrate self-supervised sensorimotor learning with a system that learns to mimic a Zebra Finch. The system is directly modeled on the dynamics of how male baby finches learn birdsong from their fathers (Tchernichovski et al. 2004, Fee et al. 2004). Our system first listens to an adult finch and uses cross-modal clustering to learn *songemes*, primitive units of bird song that we propose as an avian equivalent of phonemes. It then uses a vocalization synthesizer to generate its own nascent birdsong, guided by random exploratory motor behavior. By listening to itself sing, the system organizes the motor maps generating its vocalizations by cross-modally clustering them with respect to the previously learned *songeme* maps of its parent. In this way, it learns to generate the same sounds to which it was previously exposed. Finally, we incorporate a standard hidden Markov model into this system, to model the

temporal structure and thereby combine songemes into actual birdsong. The Zebra Finch is a particularly suitable species to use for guiding this demonstration, as each bird essentially sings a single unique song accompanied by minor variations.

We note that the above examples all use real data, gathered from a real person speaking and from a real bird singing. We also present results on a number of synthetic datasets drawn from a variety of mixture distributions to provide basic insights into the algorithms and *slice* data structure work. Finally, we believe it is possible to allow the computational side of this question to inform the biological one, and we will analyze the model, in its own right and in light of these results, to explore its algorithmic and representational implications for biological perceptual systems, particularly from the perspective of how sharing information restricts the modalities individually.

1.5 What Is a "Sense?"

Although Appendix 1 contains a glossary of technical terms, one clarification is so important that it deserves special mention. We have repeatedly used the word *sense*, e.g., sense, sensory, intersensory, etc., without defining what a *sense* is. One generally thinks of a sense as the perceptual capability associated with a distinct, usually external, sensory organ. It seems quite natural to say vision is through the eyes, touch is through the skin, etc. (Notable exceptions include proprioception – the body's sense of internal state – which is somewhat more difficult to localize and vestibular perception, which occurs mainly in the inner ear but is not necessarily experienced there.) However, this coarse definition of *sense* is misleading.

Each sensory organ provides an entire class of sensory capabilities, which we will individually call *modes*. For example, we are familiar with the *bitterness* mode of taste, which is distinct from other taste modes such as *sweetness*. In the visual system, *object segmentation* is a mode that is distinct from *color perception*, which is why we can appreciate black and white photography. Most importantly, individuals may lack

particular modes *without other modes in that sense being affected* (e.g., Wolfe 1983), thus demonstrating they are phenomenologically independent. For example, people who like broccoli are insensitive to the taste of the chemical *phenylthiocarbamide* (Drayna et al. 2003); however, we would not say these people are unable to taste – they are simply missing an individual taste mode. There are a wide variety of visual agnosias that selectively affect visual experience, e.g., *simultanagnosia* is the inability to perform visual object segmentation, but we certainly would not consider a patient with this deficit to be blind, as it leaves the other visual processing modes intact.

Considering these fine grained aspects of the senses, we allow intersensory influence to happen between modes even within the same sensory system, e.g., entirely within vision, or alternatively, between modes in different sensory systems, e.g., in vision and audition. Because the framework presented here is *amodal*, i.e., not specific to any sensory system or mode, it treats both of these cases equivalently.

1.6 Roadmap

Chapter 2 sets the stage for the rest of this thesis by visiting an example stemming from the 1939 World's Fair. It intuitively makes clear what we mean by perceptual grounding and interpretation, which until now have remained somewhat abstract.

Chapter 3 presents our approach to perceptual grounding by introducing *slices*, a data structure for representing sensory information. We then define our algorithm for cross-modal clustering, which autonomously learns perceptual categories within slices by considering how the data within them co-occur. We demonstrate this approach in learning the vowel structure of American English by simultaneously watching and listening to a person speak. Finally, we examine and contrast related work in unsupervised clustering with our approach.

Chapter 4 builds upon the results in Chapter 3 to present our approach to perceptual interpretation. We incorporate the temporal dynamics of sensory perception by treating slices as *phase spaces* through which sensory inputs move during the time windows

corresponding to percept formation. We define a dynamic activation model on slices and interconnect them through an *influence network*, which allows different modes to influence each other's perceptions dynamically. We then examine using this framework to disambiguate simultaneous audio-visual speech inputs. Note that this mathematical chapter may be skipped on a cursory reading of this thesis.

Chapter 5 builds upon the previous two chapters to define our architecture for sensorimotor learning, based on a Cartesian theater. Our system simultaneously "watches" its internal motor activity while it observes the effects of its own actions externally. Cross-modal clustering then allows it to ground its motor maps using previously clustered perceptual maps. This is possible because slices can equivalently contain perceptual or motor data, and in fact, slices do not "know" what kind of data they contain. The principle example in this chapter is the acquisition of species-specific birdsong.

Chapter 6 connects the work in the computational framework presented in this thesis with a modern understanding of perception in biological systems. Doing so motivates the approach taken here and allows us to suggest how this work may reciprocally contribute towards a better computational understanding biological perception. We also examine related work in multimodal integration and examine the engineered system that motivated much of the work in this thesis. Finally, we speculate on a number of theoretical issues in Intersensory perception and examine how the work in this thesis addresses them.

Chapter 7 contains a brief summary of the contributions of this thesis and outlines future work.

Chapter 2

Setting the Stage

We begin with an example to illustrate the two fundamental problems of perception addressed in this thesis:

1) *Grounding* – how are sensory inputs categorized in a perceptual system?

2) *Interpretation* – how should sensory inputs be classified once their possible categories are known?

The example presented below concerns speechreading, but the techniques presented in later chapters for solving the problems raised here are not specific to any perceptual modality. They can be applied to range of perceptual and motor learning problems, and we will examine some of their nonperceptual applications as well.

2.1 Peterson and Barney at the World's Fair

Our example begins with the 1939 World's Fair in New York, where Gordon Peterson and Harold Barney (1952) collected samples of 76 speakers saying sustained American

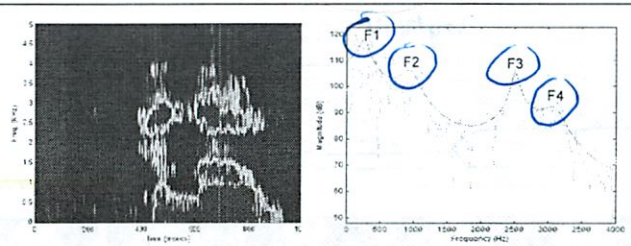


Figure 2.1— On the left is a spectrogram of the author saying, "Hello." The demarcated region (from 690-710ms) marks the middle of phoneme /ow/, corresponding to the middle of the vowel "o" in "hello." The spectrum corresponding to this 20ms window is shown on the right. A 12^{th} order linear predictive coding (LPC) model is shown overlaid, from which the formants, i.e., the spectral peaks, are estimated. In this example: F1 = 266Hz, F2 = 922Hz, and F3 = 2531Hz. Formants above F3 are generally ignored for sound classification because they tend to be speaker dependent. Notice that F2 is slightly underestimated in this example, a reflection of the heuristic nature of computational formant determination.

English vowels. They measured the fundamental frequency and first three formants (see Figure 2.1) for each sample and noticed that when plotted in various ways (Figure 2.2), different vowels fell into different regions of the formant space. This regularity gave hope that spoken language – at least vowels – could be understood through accurate estimation of formant frequencies. This early hope was dashed in part because co-articulation effects lead to considerable movement of the formants during speech (Holbrook and Fairbanks 1962). Although formant-based classifications were largely abandoned in favor of the dynamic pattern matching techniques commonly used today (Jelinek 1997), the belief persists that formants are potentially useful in speech recognition, particularly for dimensional reduction of data.

It has long been known that watching the movement of a speaker's lips helps people understand what is being said (Bender 1981, p41). The sight of someone's moving lips in an environment with significant background noise makes it easier to understand what the speaker is saying; visual cues – e.g., the sight of lips – can alter the signal-to-noise ratio of an auditory stimulus by 15-20 decibels (Sumby and Pollack 1954). The task of lip-reading has by far been the most studied problem in the computational multimodal

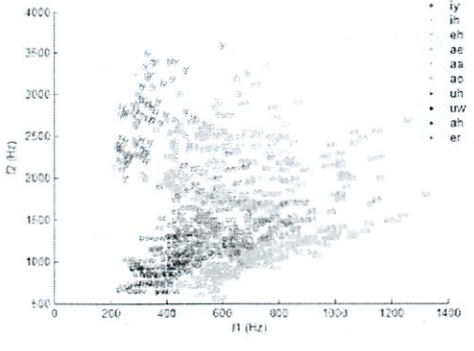


Figure 2.2 – Peterson and Barney Data. On the left is a scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories.

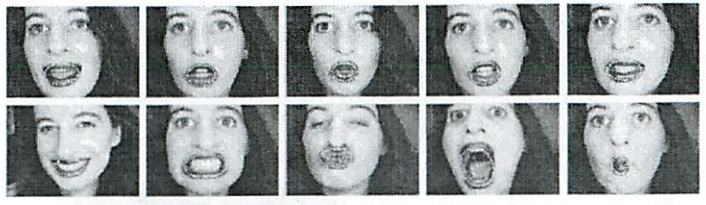


Figure 2.3 – Automatically tracking mouth positions of test subject in a video stream. Lip positions are found via a deformable template and fit to an ellipse using least squares. The upper images contains excerpts from speech segments, corresponding left to right with phonemes /eh/, /ae/, /uw/, /ah/, and /iy/. The bottom row contains non-speech mouth positions. Notice that fitting the mouth to an ellipse can be non-optimal, as is the case with the two left-most images; independently fitting the upper and lower lip curves to low-order polynomials would yield a better fit. For the purposes of this example, however, ellipses provide an adequate, distance invariant, and low-dimensional model. The author is indebted to his wife for having lips that were computationally easy to detect.

literature (e.g., Mase and Pentland 1990, Huang et al. 2003, Potamianos et al. 2004), due to the historic prominence of automatic speech recognition in computational perception. Although significant progress has been made in automatic speech recognition, state of the art performance has lagged human speech perception by up to an order of magnitude, even in highly controlled environments (Lippmann 1997). In response to this, there has been increasing interest in non-acoustic sources of speech information, of which vision has received the most attention. Information about articulator position is of particular interest, because in human speech, acoustically ambiguous sounds tend to have visually unambiguous features (Massaro and Stork 1998). For example, visual observation of tongue position and lip contours can help disambiguate unvoiced velar consonants /p/ and /k/, voiced consonants /b/ and /d/, and nasals /m/ and /n/, all of which can be difficult to distinguish on the basis of acoustic data alone.

Articulation data can also help to disambiguate vowels. Figure 2.3 contains images of a speaker voicing different sustained vowels, corresponding to those in Figure 2.2. These images are the unmodified output of a mouth tracking system written by the author, where the estimated lip contour is displayed as an ellipse and overlaid on top of the speaker's mouth. The scatterplot in Figure 2.4 shows how a speaker's mouth is represented in this way, with contour data normalized such that a resting mouth

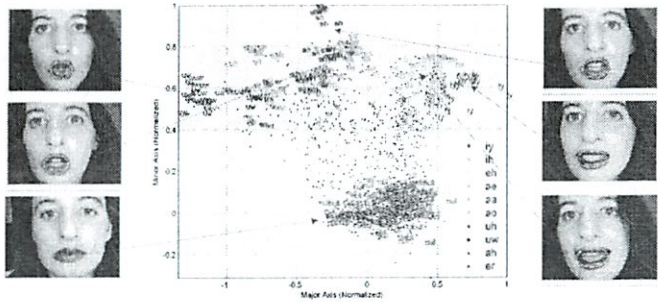


Figure 2.4 -- Modeling lip contours with an ellipse. The scatterplot shows normalized major (x) and minor (y) axes for ellipses corresponding to the same vowels as those in Figure 2.2. In this space, a closed mouth corresponds to a point labeled *null*. Other lip contours can be viewed as offsets from the null configuration and are shown here segmented by color. These data points were collected from video of this woman speaking.

configuration (referred to as *null* in the figure) corresponds with the origin, and other mouth positions are viewed as offsets from this position. For example, when the subject makes an /iy/ sound, the ellipse is elongated along its major axis, as reflected in the scatterplot.

Suppose we now consider the formant and lip contour data simultaneously, as in Figure 2.5. Because the data are conveniently labeled, the clusters within and the correspondences between the two scatterplots are obvious. We notice that the two domains can mutually disambiguate one another. For example, /er/ and /uh/ are difficult to separate acoustically with formants but are easy to distinguish visually. Conversely, /ae/ and /eh/ are visually similar but acoustically distinct. Using these complementary representations, one could imagine combining the auditory and visual information to create a simple speechreading system for vowels.

2.2 Nature Does Not Label Its Data

Given this example, it may be surprising that our interest here is not in building a speechreading system. Rather, we are concerned with a more fundamental problem: how do sensory systems learn to segment their inputs to begin with? In the color-coded plots

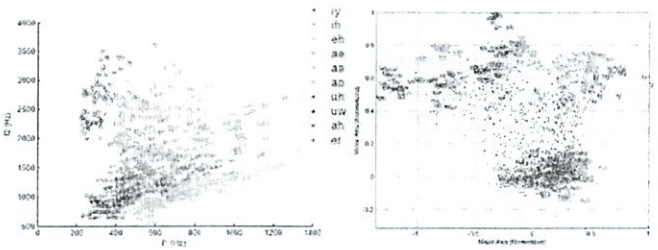


Figure 2.5 – Labeled scatterplots side-by-side. Formant data (from Peterson Barney 1952) is displayed on the left and lip contour data (from the author's wife) is show on the right. Each plot contains data corresponding to the ten listed vowels in American English.

in Figure 2.5, it is easy to see the different represented categories. However, perceptual events in the world are generally not accompanied with explicit category labels. Instead, animals are faced with data like those in Figure 2.6 and must somehow learn to make sense of them. We want to know how the categories are learned in the first place. We note this learning process is not confined to development, as perceptual correspondences are plastic and can change over time.

We would therefore like to have a general purpose way of taking data (such as shown in Figure 2.6) and deriving the kinds of correspondences and segmentations (as shown in Figure 2.5) without external supervision. This is what we mean by *perceptual grounding*

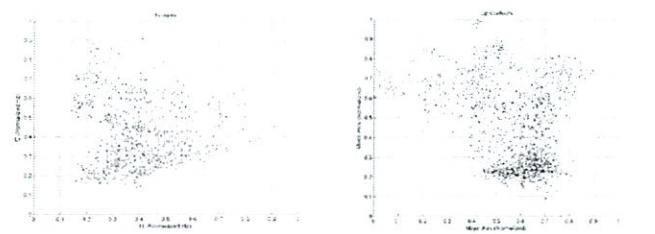


Figure 2.6 – Unlabeled data. These are the same data shown above in Figure 2.5, with the labels removed. This picture is closer to what animals actually encounter in Nature. As above, formants are displayed on the left and lip contours are on the right. Our goal is to learn the categories present in these data without supervision, so that we can automatically derive the categories and clusters such as those show above.

and our perspective here is that it is a clustering problem: animals must learn to organize their perceptions into meaningful categories. We examine below why this is a challenging problem.

2.3 Why Is This Difficult?

As we have noted above, Nature does not label its data. By this, we mean that the perceptual inputs animals receive are not generally accompanied by any meta-level data explaining what they represent. Our framework must therefore assume the learning is unsupervised, in that there are no data outside of the perceptual inputs themselves available to the learner.

From a clustering perspective, perceptual data is highly non-parametric in that both the number of clusters and their underlying distributions may be unknown. Clustering algorithms generally make strong assumptions about one or both of these. For example, the Expectation Maximization algorithm (Dempster et al. 1977) is frequently used a basis for clustering mixtures of distributions whose maximum likelihood estimation is easy to compute. This algorithm is therefore popular for clustering known finite numbers of Gaussian mixture models (e.g., Nabney 2002, Witten and Frank 2005). However, if the number of clusters is unknown, the algorithm tends to converge to a local minimum with the wrong number of clusters. Also, if the data deviate from a mixture of Gaussian (or some expected) distributions, the assignment of clusters degrades accordingly. More generally, when faced with nonparametric, distribution-free data, algorithmic clustering techniques tend not be robust (Fraley and Raftery 2002, Still and Bialek 2004).

Perceptual data are also noisy. This is due both to the enormous amount of variability in the world and to the probabilistic nature of the neuronal firings that are responsible for the perception (and sometimes the generation) of perceivable events. The brain itself introduces a great deal of uncertainty into many perceptual processes. In fact, one may perhaps view the need for high precision as the exception rather than the rule. For example, during auditory localization based on interaural time delays, highly specialized

neurons known as the *end-bulbs of Held* – among the largest neuronal structures in the brain – provide the requisite accuracy by making neuronal firings in this section of auditory cortex highly deterministic (Trussell 1999). It appears that the need for mathematical precision during perceptual processing can require extraordinary neuroanatomical specialization.

Perhaps most importantly, perceptual grounding is difficult because there is no objective mathematical definition of "coherence" or "similarity." In many approaches to clustering, each cluster is represented by a prototype that, according to some well-defined measure, is an exemplar for all other data it represents. However, in the absence of fairly strong assumptions about the data being clustered, there may be no obvious way to select this measure. In other words, it is not clear how to formally define what it means for data to be objectively similar or dissimilar. In perceptual and cognitive domains, it may also depend on why the question of similarity is being asked. Consider a classic AI conundrum, "*what constitutes a chair?*" (Winston 1970, Minsky 1974, Brooks 1987). For many purposes, it may be sufficient to respond, "*anything upon which one can sit.*" However, when decorating a home, we may prefer a slightly more sophisticated answer. Although this is a higher level distinction than the ones we examine in this thesis, the principle remains the same and reminds us why similarity can be such a difficult notion to pin down.

Finally, even if we were to formulate a satisfactory measure of similarity for static data, one might then ask how this measure would behave in a dynamic system. Many perceptual (and motor) systems are inherently dynamic – they involve processes with complex, non-linear temporal behavior (Thelen and Smith 1994), as can been seen during perceptual bistability, cross-modal influence, habituation, and priming. Thus, one may ask whether a similarity metric captures a system's temporal dynamics; in a clustering domain, the question might be posed: *do points that start out in the same cluster end up in the same cluster?* We know from Lorentz (1964) that it is possible for arbitrarily small differences to be amplified in a non-linear system. It is quite plausible that a static similarity metric might be oblivious to a system's temporal dynamics, and therefore, sensory inputs that initially seem almost identical could lead to entirely different percepts

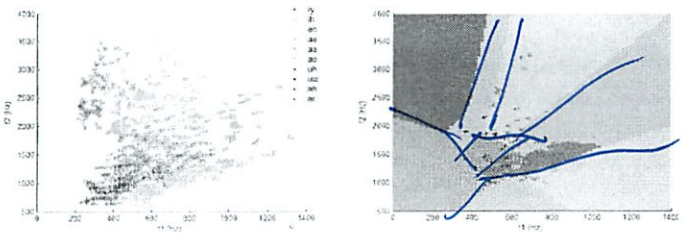


Figure 2.7 – On the left is a scatterplot of the first two formants, with different regions labeled by their corresponding vowel categories. The output of a backpropagation neural network trained on this data is shown on the right and displays decision boundaries and misclassified points. The misclassification error in this case is 19.7%. Other learning algorithms, e.g., AdaBoost using C4.5, Boosted stumps with LogitBoost, and SVM with a 5th order polynomial kernel, have all shown similarly lackluster performance, even when additional dimensions (corresponding to F0 and F3) are included (Klautau 2002). (Figure on right is derived from ibid. and used with permission.)

being generated. This issue will be raised in more detail in Chapter 4, where we will view clusters as fixed points in representational phase spaces in which perceptual inputs follow trajectories between different clusters.

In Chapter 3, we will present a framework for perceptual grounding that addresses many of the issues raised here. We show that animals (and machines) can learn how to cluster their perceptual inputs by simultaneously correlating information from their different senses, even when they have no advance knowledge of what events these senses are individually capable of perceiving. By *cross-modally* sharing information between different senses, we will demonstrate that sensory systems can be perceptually grounded by bootstrapping off each other.

2.4 Perceptual Interpretation

The previous section outlined some of the difficulties in unsupervised clustering of nonparametric sensory data. However, even if the data came already labeled and clustered, it would still be challenging to classify new data points using this information. Determining how to assign a new data point to a preexisting cluster (or category) is what we mean by *perceptual interpretation*. It is the process of deciding what a new input

actually represents. In the example here, the difficultly is due to the complexity of partitioning formant space to separate the different vowels. This 50 year old classification problem still receives attention today (e.g., Jacobs et al. 1991, de Sa and Ballard 1998, Clarkson and Moreno 1999) and Klautau (2002) has surveyed modern machine learning algorithms applied to it, an example of which is shown on the right in Figure 2.7.

A common way to distinguish classification algorithms is by visualizing the different spaces of possible decision boundaries they are capable of learning. If one closely examines the Peterson and Barney dataset (Figure 2.8), there are many pairs of points that are nearly identical in any formant space but correspond to different vowels in the actual data, at least according to the speaker's intention. It is difficult to imagine any accurate partitioning that would simultaneously avoid overfitting. There are many factors that contribute to this, including the information loss of formant analysis (i.e., incomplete data is being classified), computational errors in estimating the formants, lack of

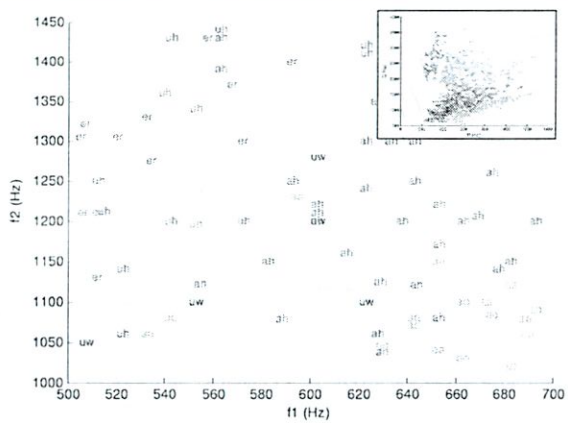


Figure 2.8 – Focusing on one of many ambiguous regions in the Peterson-Barney dataset. Due to a confluence of factors described in the text, the data in these regions are not obviously separable.

differentiation in vowel pronunciation in different dialects of American English, variations in prosody, and individual anatomical differences in the speakers' vocal tracts. It is worth pointing out the latter three of these for the most part exist independently of how data is extracted from the speech signal and may present difficulties regardless of how the signal is processed.

The curse of dimensionality (Bellman 1961) is a statement about exponential growth in hypervolume as a function of a space's dimension. Of its many ramifications, the most important here is that many low dimensional phenomena that we are intuitively familiar with do not exist in higher dimensions. For example, the natural clustering of uniformly distributed random points in a two dimensional space becomes extremely unlikely in higher dimensions; in other words, random points are relatively far apart in high dimensions. In fact, transforming nonseparable samples into higher dimensions is a general heuristic for improving separation with many classification schemes. There is a flip-side to this high dimensional curse for us: ***low dimensional spaces are crowded***. It can be difficult to separate classes in these spaces because of their tendency to overlap. However, blaming low dimensionality for this problem is like the proverbial cursing of darkness. Cortical architectures make extensive use of low dimensional spaces, e.g., throughout visual, auditory, and somatosensory processing (Amari 1980, Swindale 1996, Dale et al. 1999, Fischl et al. 1999, Kaas and Hackett 2000, Kardar and Zee 2002, Bednar et al. 2004), and this was a primary motivating factor in the development of Self Organizing Maps (Kohonen 1984). In these crowded low-dimensional spaces, approaches that try to implicitly or explicitly refine decision boundaries such as those in Figure 2.8 (e.g., de Sa 1994) are likely to meet with limited success because there may be no good decision boundaries to find; perhaps in these domains, decision boundaries are the wrong way to think about the problem.

Rather than trying to improve classification boundaries directly, one could instead look for a way to move ambiguous inputs into easily classified subsets of their representational spaces. This is the essence of the *influence network* approach presented in Chapter 4 and is our proposed solution to the problem of perceptual interpretation. The goal is to use cross-modal information to "move" sensory inputs within their own state spaces to make

them easier to classify. Thus, we take the view that perceptual interpretation is inherently a dynamic – rather than static – process that occurs during some window of time. This approach relaxes the requirement that the training data be separable in the traditional machine learning sense; unclassifiable subspaces are not a problem if we can determine how to move out of them by relying on other modalities, which are experiencing the same sensory events from their unique perspectives. We will show that this approach is not only biologically plausible, it is also computationally efficient in that it allows us to use lower dimensional representations for modeling sensory and motor data.

How do you do it

1. ~~Rules + Sets~~ Characterize behavior
2. Formulate Computational problems
3. Propose " solutions
4. Implement explatory Systems
5. Crystalize the principles

Strong Story Hyp

That was his story telling MacBeth/ Croatia thing
One of his papers is online

Revised exam for other Part 3

- Near Miss - which is quiz 3 here
- And some multiple choice
  - That may be harder this year
- Mutiple Choice I can look up
  - open book

(learned about it in OH - find that pg)

lets ya evaluate much faster

Anytime algorithm can stop it at any pt in time and it will give ya best more found so far

(17)

How do you do it

1. ~~Rules + Se~~ Characterize behavior
2. Formulate Computational problems
3. Propose " solutions
4. Implement explatory Systems
5. Crystalize the principles

Strong Story Myp

That was his story telling Macbeth/ Croatia thing
One of his papers is online

Revised exam for after Part 3

- Near Miss - which is quiz 3 here
- And some multiple choice
  - That may be harder this year
- Mutiple Choice I can look up
  - open book

(18)

Unit 2 Review

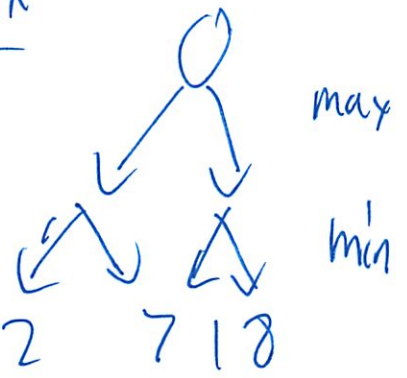
Games + Constraints

Games

Need to think ahead of opponent
└ several moves

Scoring polynomial

Minimax



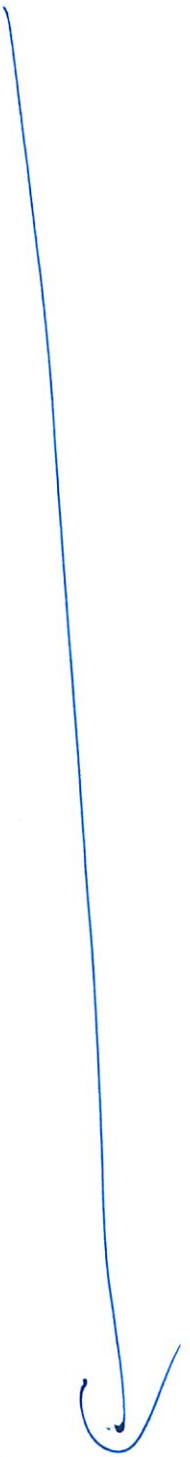
α, β very complicated add on
(learned about it in OH - find that pg)
lets ya evaluate much faster

Anytime algorithm can stop it at any pt in time and it will give ya best move found so far

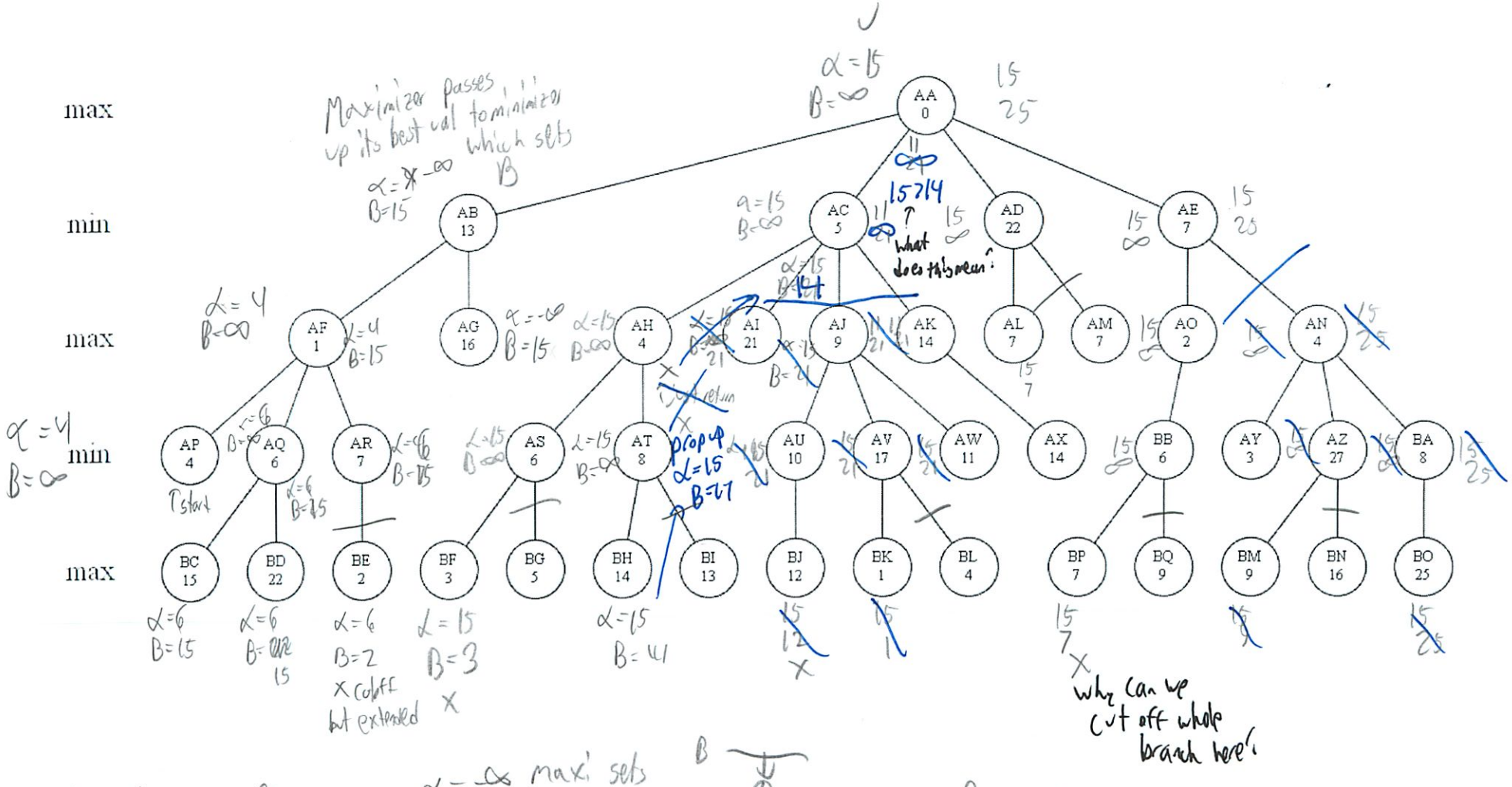
(19)

Ok α, β in details

I should try this exact problem

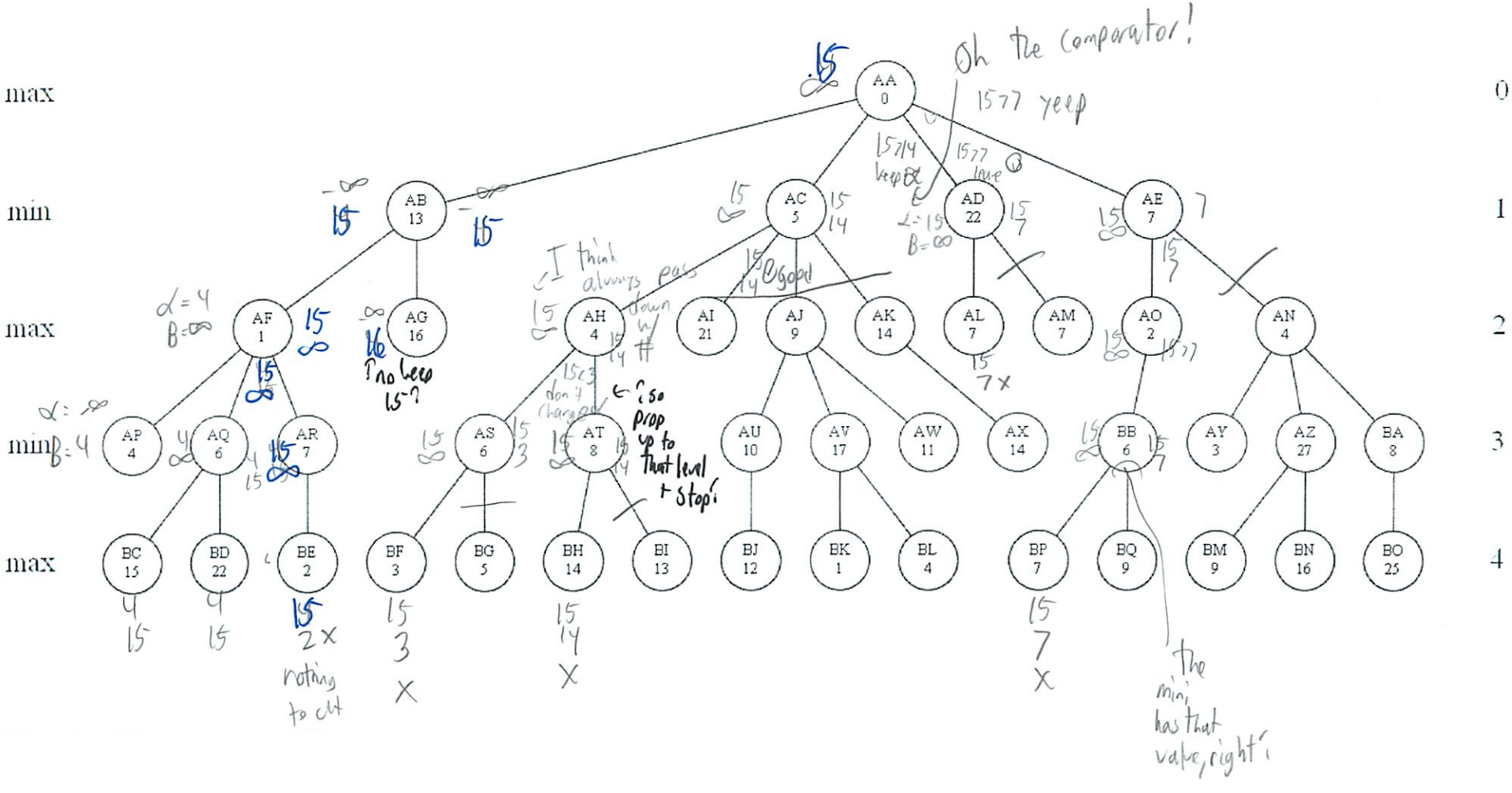


Tear off sheet



21

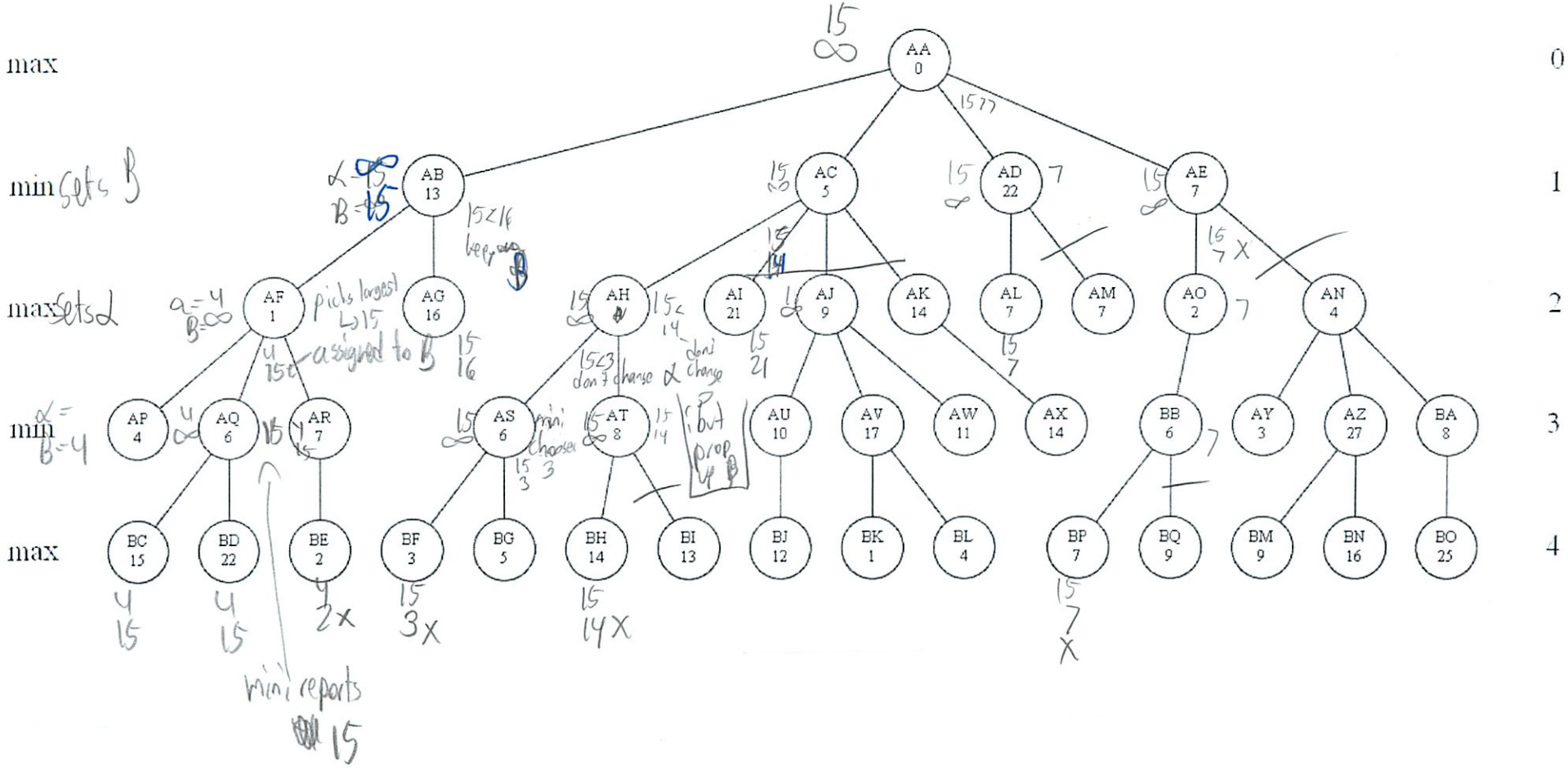
Ok try again



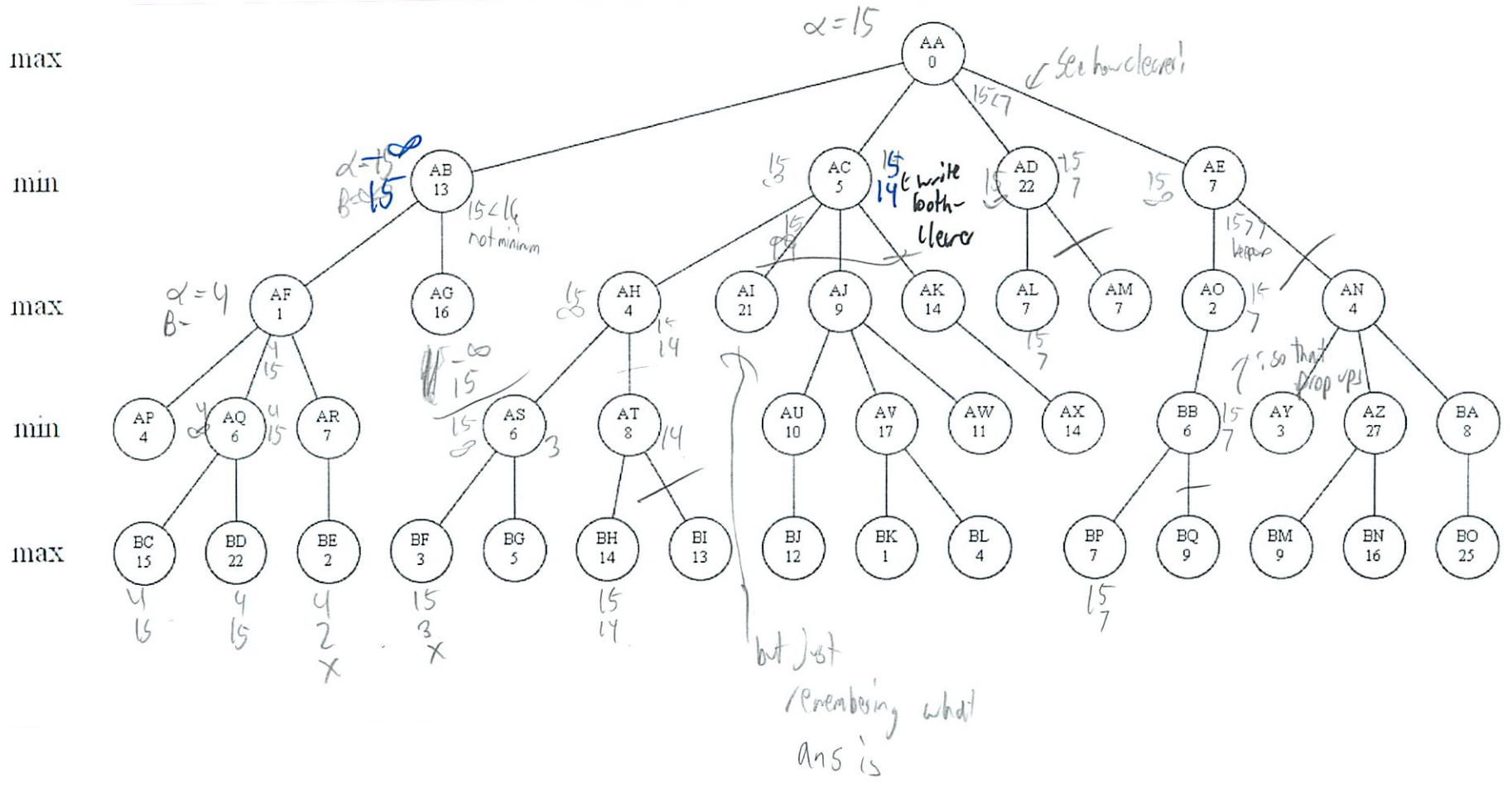
14

(22)

One more time



I think I kinda got it



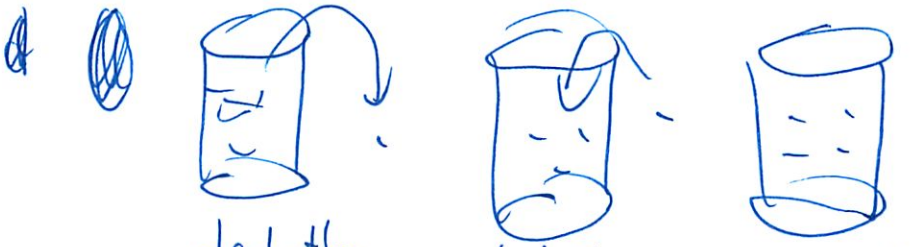
Try to ask
still kinda confused

Explore other issues:

Lash in oth

Line labeling

world has vertexes

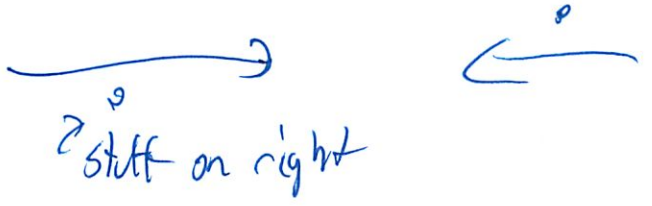


deduction
many to few

induction
few to many

abduction
usually true, so assume to be true

1. 3 Faces → Tri-hedral vertices
2. General Position
3. Assume can see convex edge + or concave edge —



↗ stuff on right

- specific case of constraint prop.

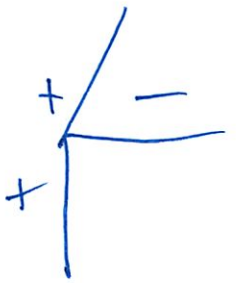
(25)

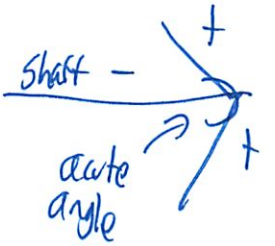
On our exam

Variables - the 4 corners

domains 3 possibilities

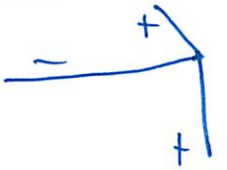
1. Temp assign UL to H



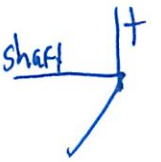


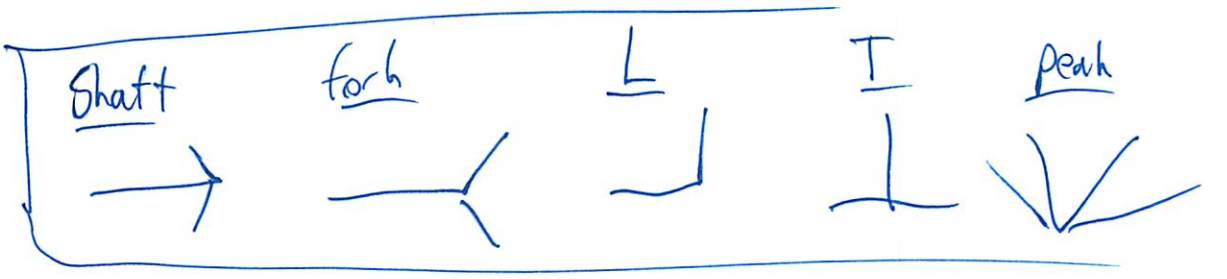
2. Go to UR

−3 fit



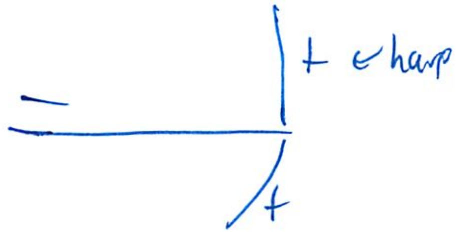
3. Br



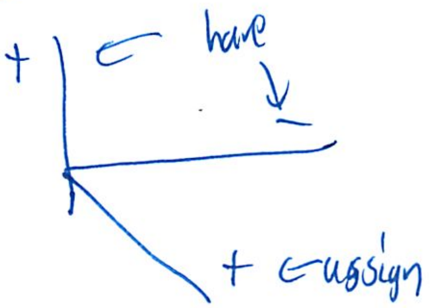


(26)

Could assign A again



4. BL



✓ works

~~So Think of Physically~~

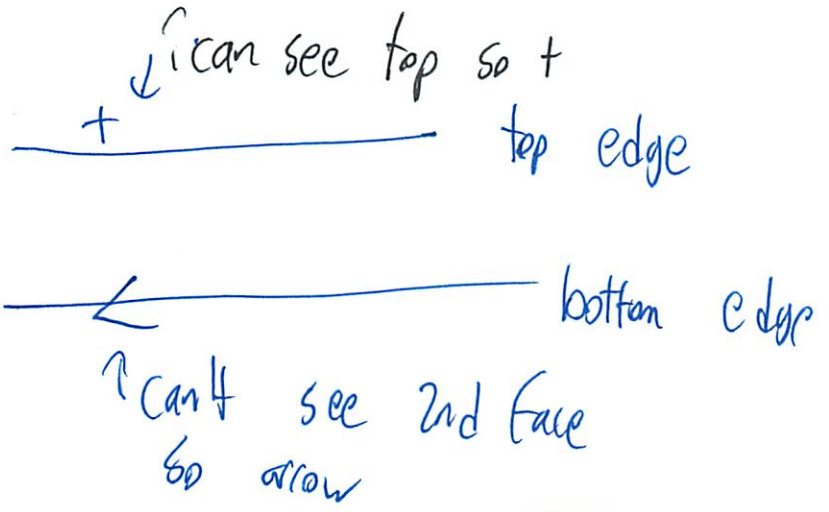
+ or · → ↓ stuff on right

both say where stuff is

So leaning head on table looking off

When sit back in chair

27



So Shapes are often constraint problems
- like before

Waltz More complex system

Shadows

Cracks

C C C C

3 types of lighting

28

Constraint Propagation

This is that 700 problem
all those different versions

Lecture example: assigning colors to states

Full monty: backtracking

Bit pain in ass - instead domain-reducing algor,

Terms

Variable V can have assignments

Values x, y can be assignments

Domain D bag of values

Constraint C limit values in typical pair of domains

I remember making a comprehensive list of diff options somewhere

DFS w/ Constraint prop

(long algo description)

Ways to consider

1. No check
2. Assignments
3. Neighbors only
4. Domains reduced to 1 value
5. Neighbors of neighbors
6. Check everything

Airline scheduling ? is map coloring

---

3 types here

1. BT
2. ~~BT~~ BT w/ FC w/ singleton domains
3. BT w/ FC w/ reduced "
   - like Arc consistency 2

---

singleton domains ← is this actually called forward checking

When 1 assigned ~~to~~ | (well modified cross out - normally check all neighbors when assigning node)
- for all neighbors cross out | don't yet assign when only 1 left

(30)

Propagation though singleton domains

if during FC we reduce a domain size to 1 then assign that single 1 and repeat forward checking (or crossing out?) for that one

So on exam trick w/ prop for any reduced domain
- which meant problem was solved instantly
- so not much actual DFS to do

These problems seem simple - but aren't
Or they would be as easy as Quiz 1 search if I did them right!

Its also hard to visualize

Think these are range of choice
didn't clearly sep FC from prop in my head before

31

Also the notation for this stuff is hard
Lneed to write as expected
Asked in OH

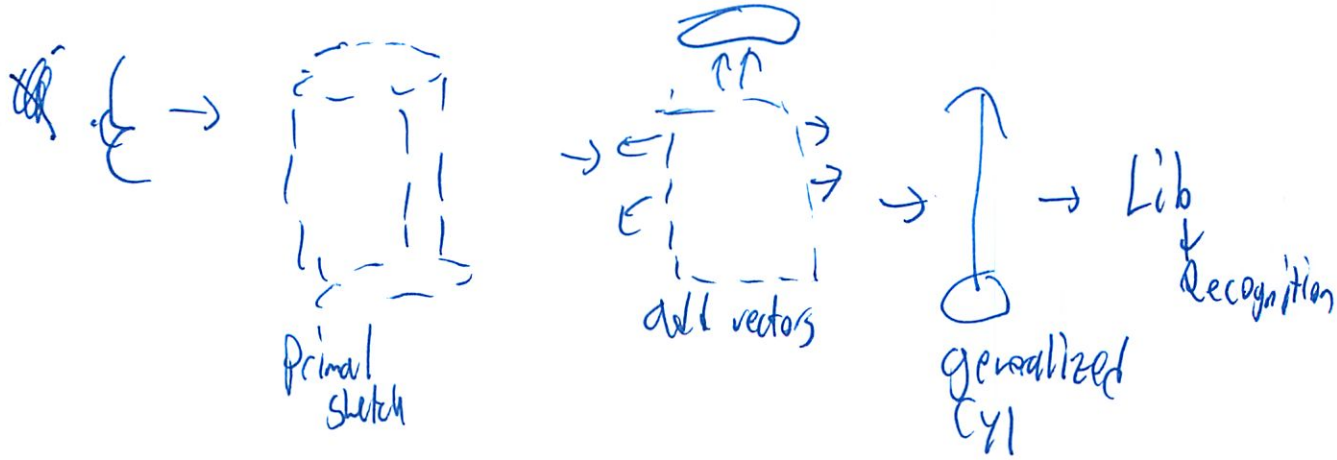
Don't write propagation checking one
if free code (from P-set) tries it → write
if prop code " " → Don't write

Consider constraints before writing
Don't write things eliminated w/ FC

Again hard to write w/ changes over time

Lecture Recognition

Marr's Model



Equation w/ lines

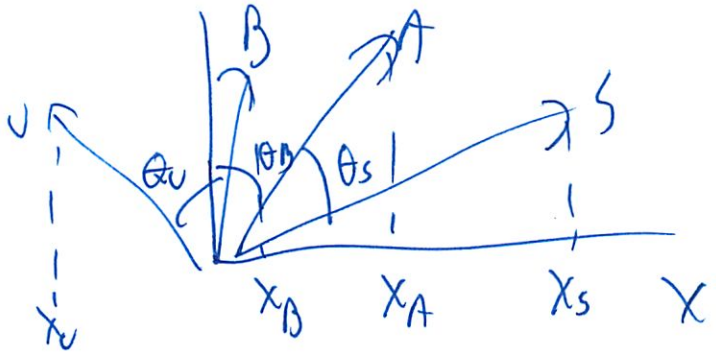
$$\begin{aligned} X_u &= \alpha X_A + \beta X_B + \gamma X_C + T \\ X_v &= \alpha X_A + \beta X_B + \gamma X_C + T \\ X_w &= \alpha X_A + \beta X_B + \gamma X_C + T \\ X_w &= \alpha X_A + \beta X_B + \gamma X_C + T \end{aligned}$$

all the same

4 eq 4 unknowns

α, β, γ, T

Rotate



$$\begin{aligned} X_A &= X_S \cos\theta_A - Y_S \sin\theta_A \\ X_B &= X_S \cos\theta_B - Y_S \sin\theta_B \\ X_u &= X_S \cos\theta_u - Y_S \sin\theta_u \end{aligned}$$

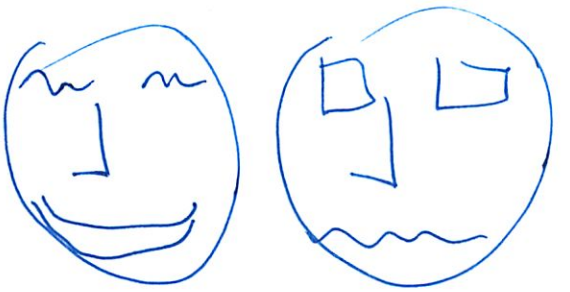
Are we going to have to do anything w/ this?
Seems too complicated

3rd approach

Scan over img looking for correlation

$$\max \sum_x \sum_y f(x,y)\, g(x+X, y+Y)$$

Have img



Not

Goldilox - pieces can't be to big or small

(34)

So Classification trees, etc Not on exam

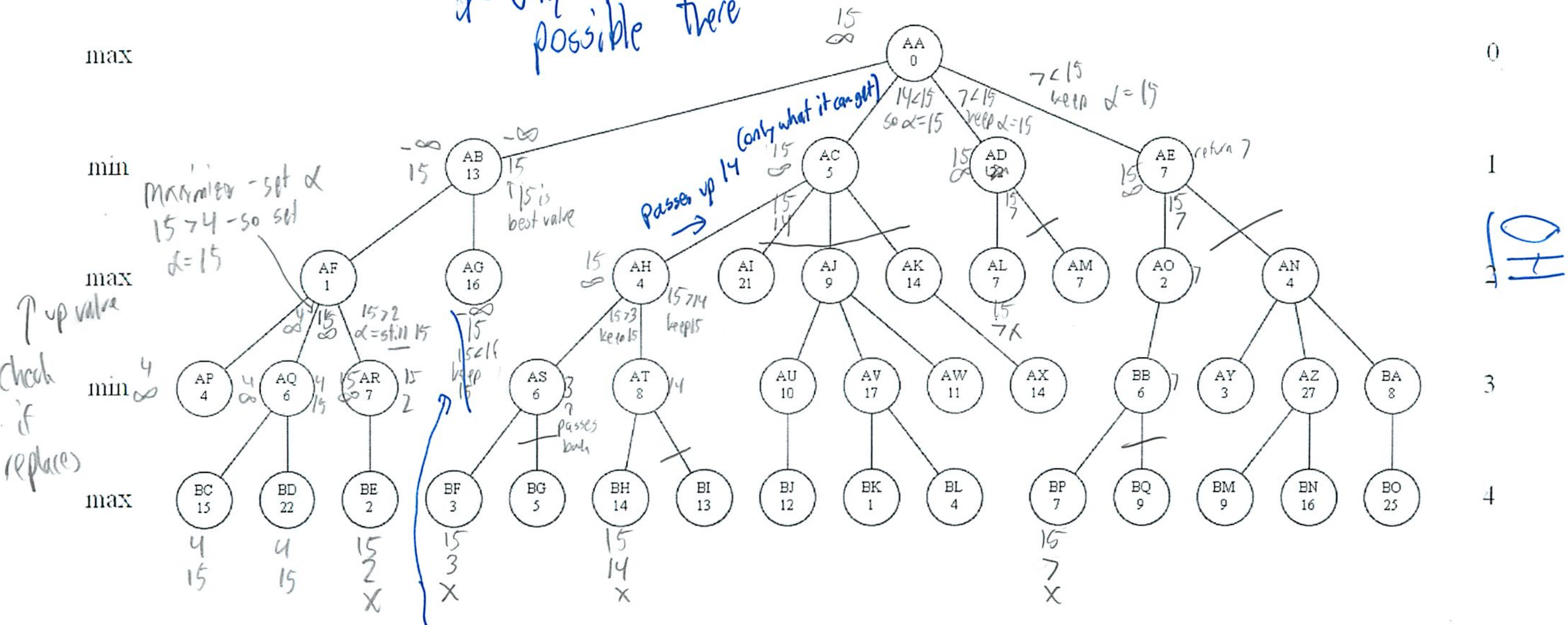
Now need to pratice

- did a bunch of pratice exams

12/16 OH w/ Erek

on 2,B search

* Only pass up values possible there



for leaf just pass up value not doing α,β notation windows

Maxi sets α
mini β

12/16

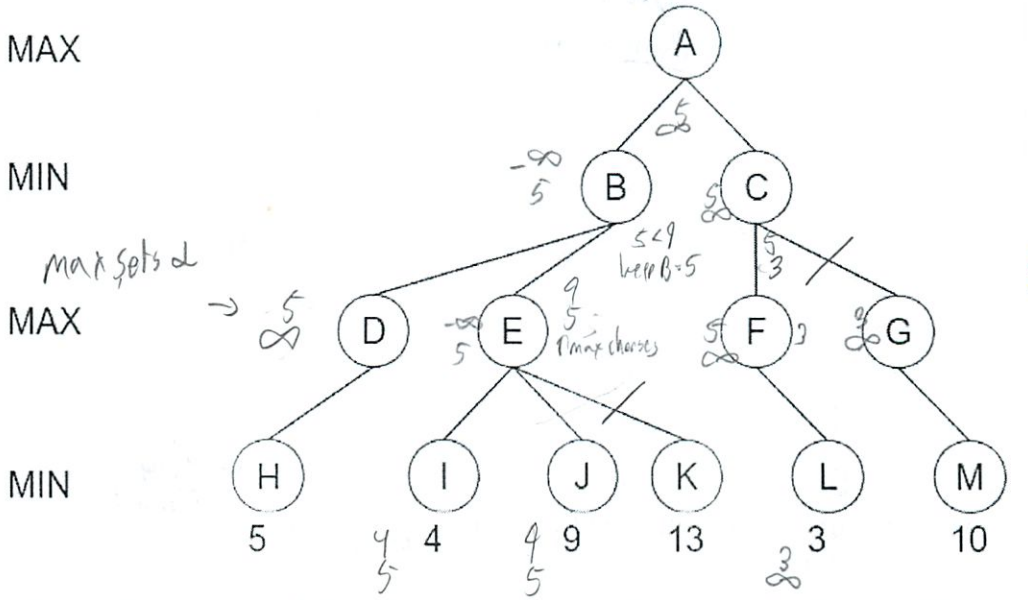
(2)

Part B: Alpha Beta (35 points)

B1: Straight-forward Alpha Beta(15 points)

* Always same as minmax

Perform Alpha Beta search on the following tree.
- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



* think about minmax

Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

# of evaluations: 9 **List:** H I J K

Next Move: _____ **Moving towards:** _____

Optimal

best to worst →

sel best things lst so cut off early

max (like here) greatest → least not guarented

(4)

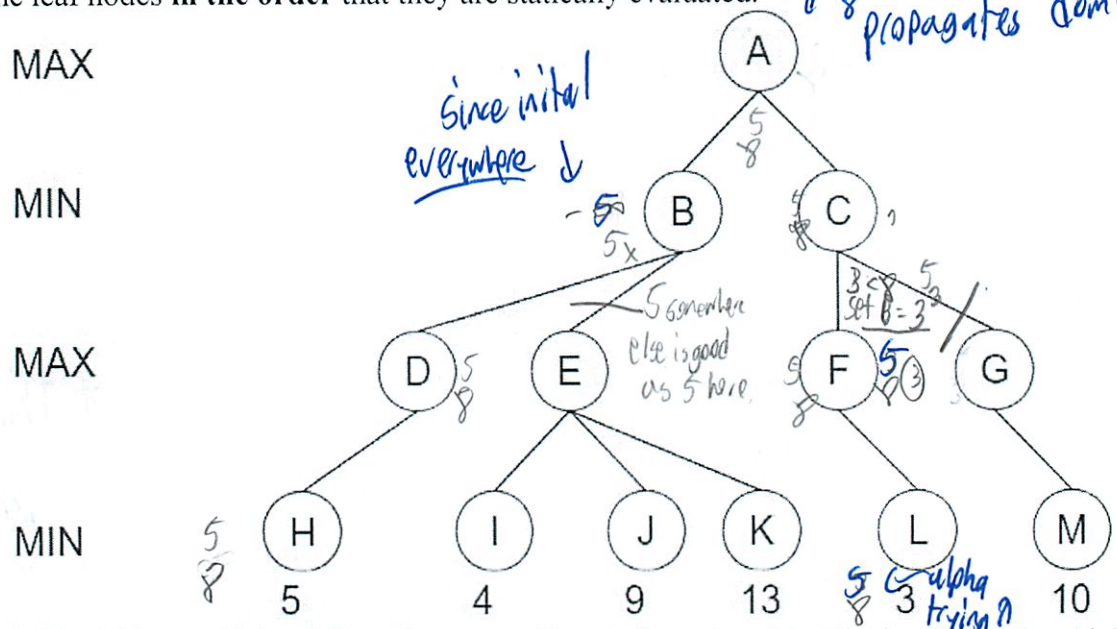
like a tree can't see

outcome can differ from minimax

B2: Preset Alpha-Beta (15 points)

Perform alpha-beta search, using **initial values** of **alpha = 5 and beta = 8**.

- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

# of evaluations: 2 **List:** H L

Next Move: _____ **Moving towards:** _____

Technically don't go down tree - could do to M but would have to prune I, J

B3: Alpha-Beta Properties (5 points)

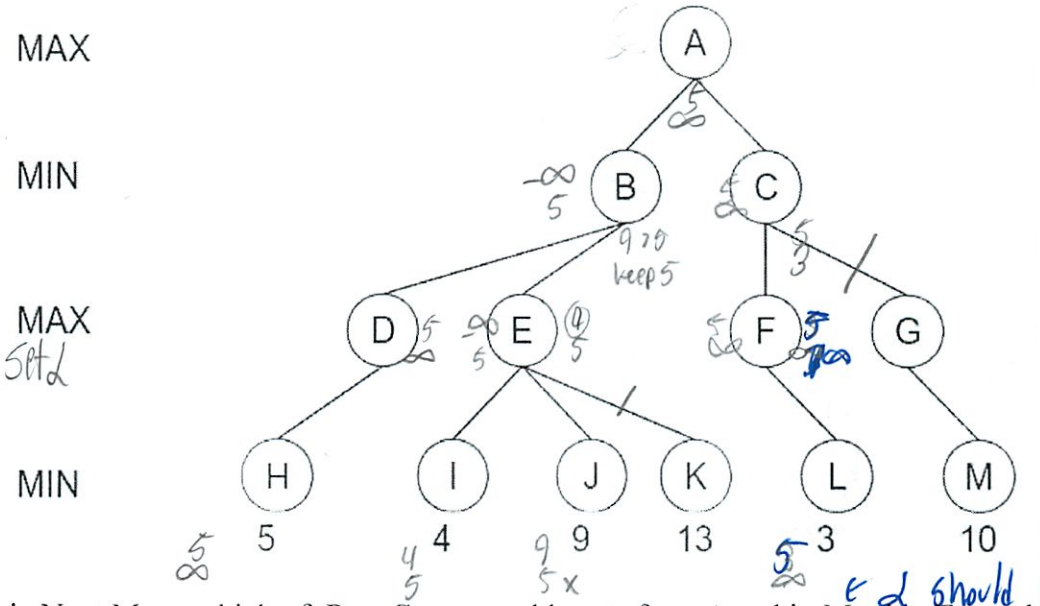
If you were able to maximally prune a tree while performing Alpha-Beta search, approaximately how many static evaluations would you end up doing for a tree of depth d and branching factor b?

(3)

Part B: Alpha Beta (35 points)

B1: Straight-forward Alpha Beta(15 points)

Perform Alpha Beta search on the following tree.
- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

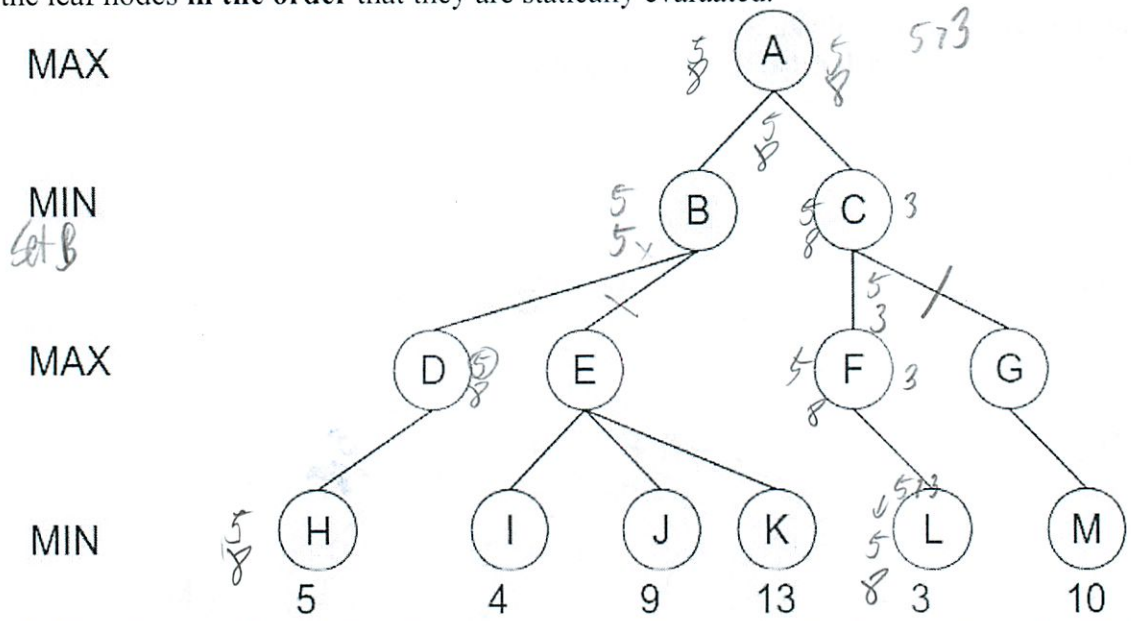
# of evaluations:_____ **List:** H I J L

Next Move: _____ **Moving towards:** _____

B2: Preset Alpha-Beta (15 points)

Perform alpha-beta search, using **initial values** of **alpha = 5 and beta = 8**.

- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

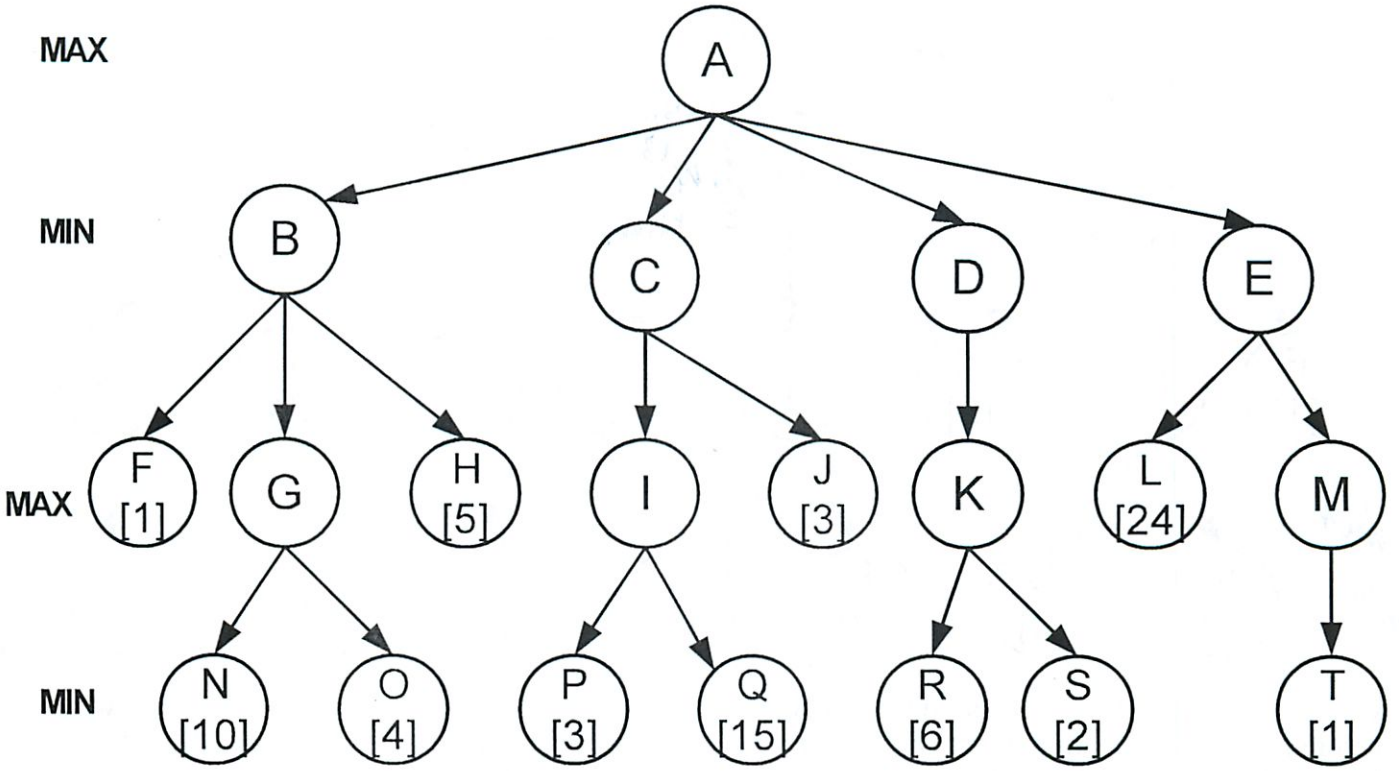
# of evaluations:____ **List:_____**

Next Move: _____ **Moving towards: _____**

B3: Alpha-Beta Properties (5 points)

If you were able to maximally prune a tree while performing Alpha-Beta search, approaximately how many static evaluations would you end up doing for a tree of depth *d* and branching factor *b*?

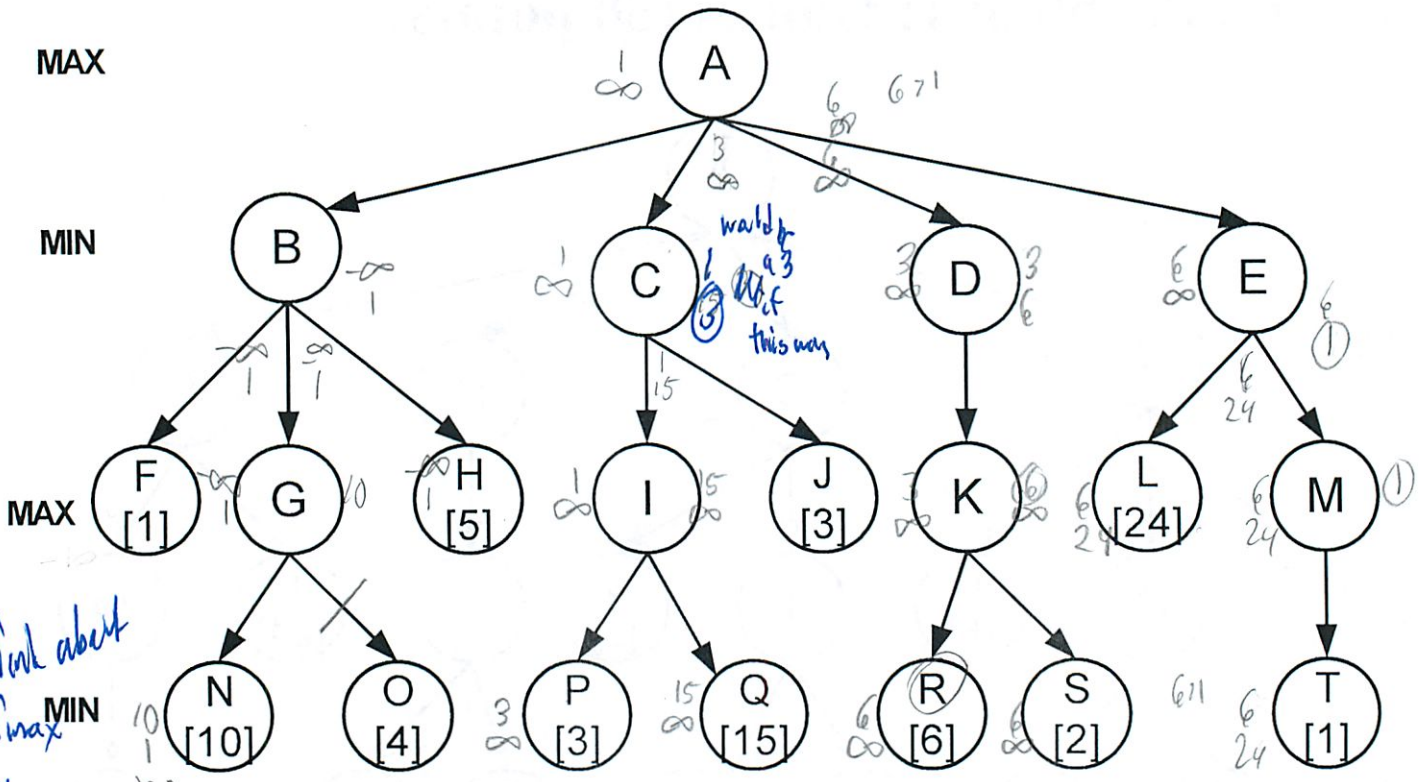
Quiz 2 Problem 1: Games (50 points)



A: Using minimax only, no alpha-beta,, indicate the values of the following nodes. (10 pts)

A =	G =
B =	I =
C =	K =
D =	M =
E =	

B: Using minimax only, what is the best next move from A? **(Indicate a letter)** (4 pts)



C: Trace the steps of Alpha beta pruning on the same tree above. Note that alpha, betas are updated before pruning occurs, if in doubt consult the reference implementation given on the tear-off sheet. **List the leaf nodes in the order that they are statically evaluated.** (10 pts)

F N H P Q J R S L T

What are the final Alpha Beta values at node E (4pts)

Alpha = 6	Beta = 1

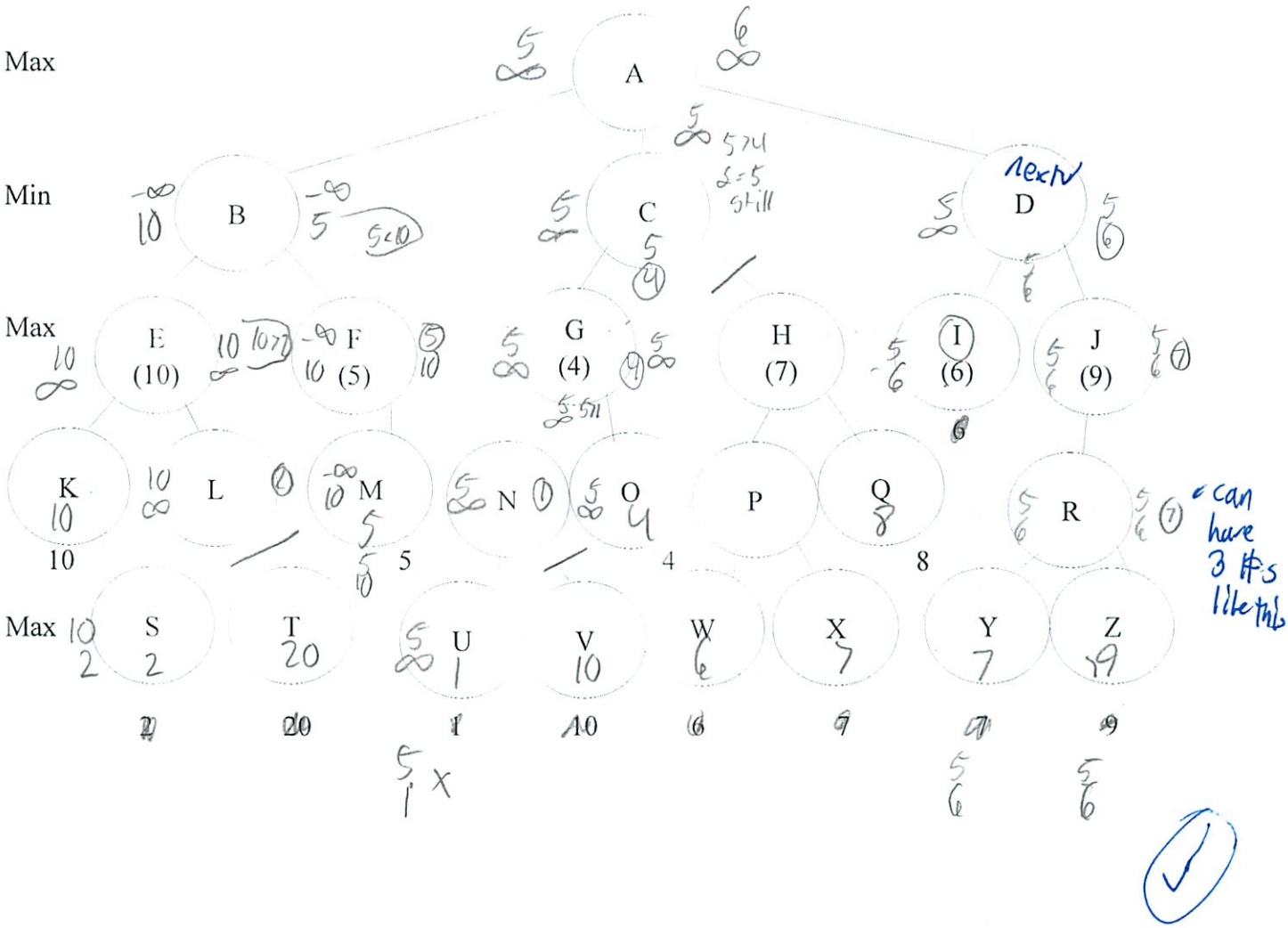
What are the final Alpha Beta values at node A (4pts)

Alpha = 6	Beta = ∞

Quiz 2, Question 1, Games (50 points)

You are playing a new Sim game called Obamaquest, the Legend of the Lost International Credibility. In this game, you play a charismatic incoming president who must make a choice on various issues in order to save your country. After each of your turns, the outgoing president will attempt to perform the most meddlesome acts possible to make it less likely that you will succeed. You realize quickly that you can model this game using a simple Game Tree from 6.034, as shown below.

Static values are shown underneath leaf nodes. Ignore the numbers in parentheses for now.



Part A (15 points)

First, you decide to perform a simple minimax algorithm on the tree.

Which direction will the maximizer choose to go at node A?

What is the minimax value of node A?

Which static evaluations did you perform? (write the nodes you statically evaluated, in order).

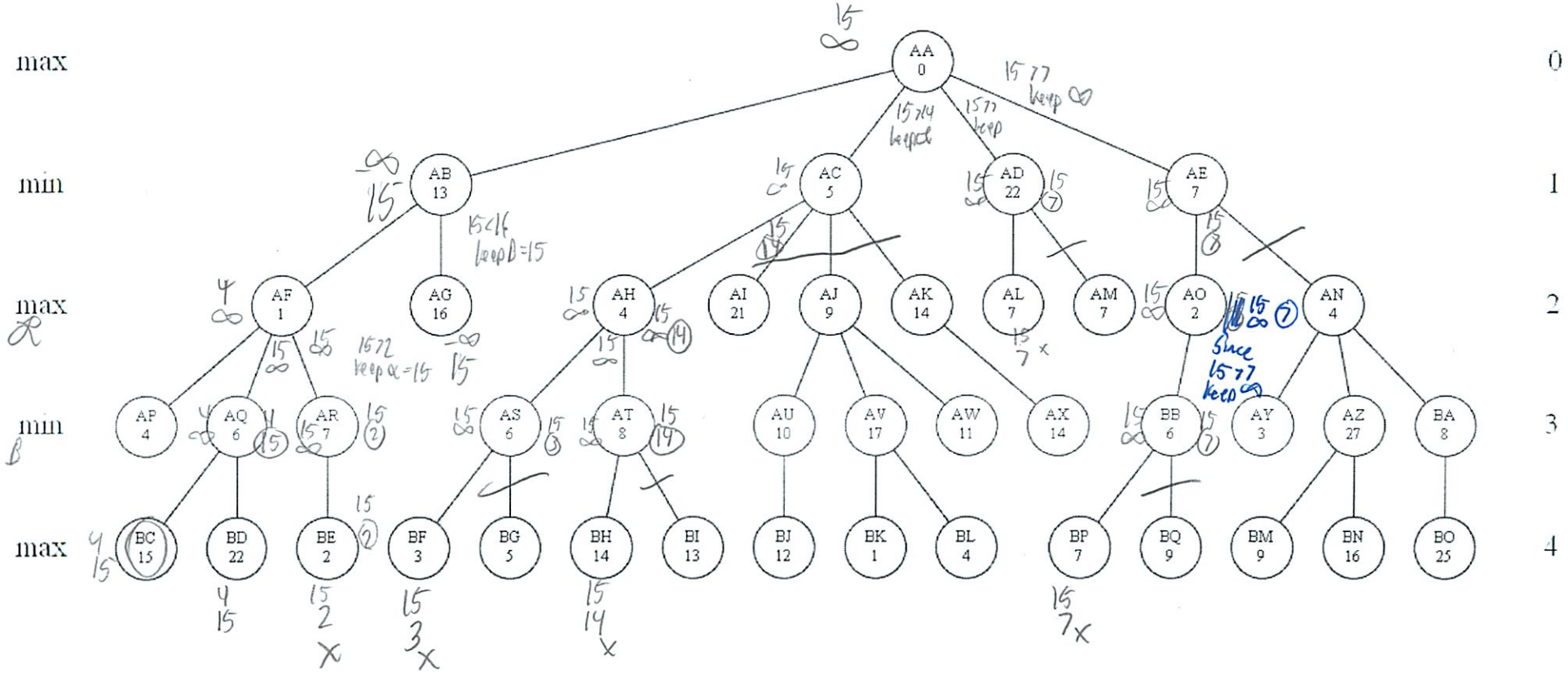
Part B (25 points)

Minimax was taking too many static evaluations, so you use alpha-beta instead.

This time what direction will the maximizer choose to go at node A?

Which static evaluations did you perform? (write the nodes you statically evaluated, in order).

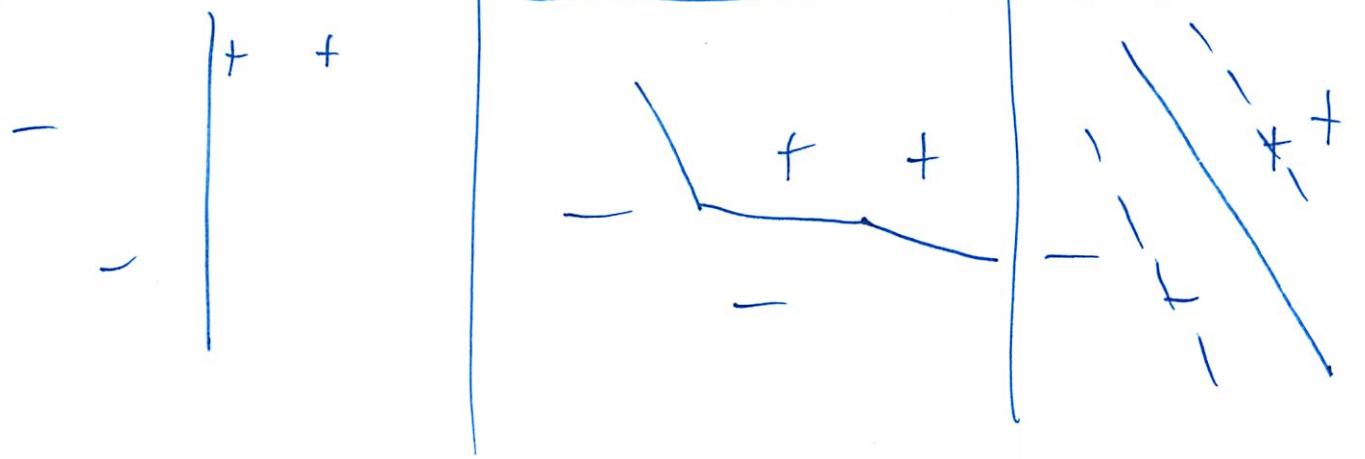
14



$\alpha \geq \beta$ x cutoff

Unit 4 SVMs + Boosting

Classification Trees | Nearest Neighbor | SVMs

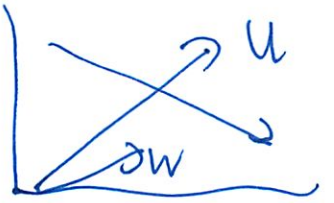


Look for widest street

(darn this was also technically hard...)

Points on street are called support vectors

1. Decision Rule



$||\bar{w}||\,||\bar{u}||\cos\theta = \bar{w}\cdot\bar{u}$

? magnitude of

$\bar{w}\cdot\bar{u} \geq c$ then u should be ⊕

$\bar{w}\cdot\bar{u} + b > 0$ then ⊕ $\quad b = -c$

Constraints

$$\bar{w} \cdot \bar{x}_+ + b \geq 1$$
$$\bar{w} \cdot x_- + b \leq 1$$

$$Y = \begin{cases} 1 \ (+) & y_+ \\ -1 \ (-) & y_- \end{cases}$$

So

$$\boxed{y_i(\bar{w} \cdot x_2 + b) - 1 \geq 0}$$

Find widest street + gutters

$$(\bar{x}_+ - \bar{x}_-) \cdot \frac{\bar{w}}{\|\bar{w}\|} = \text{width}$$

width of street $= \frac{2}{\|\bar{w}\|}$

↑ maximize

$$L = \frac{1}{2}\|w\|^2 - \sum \alpha_i [y_i(\bar{w} \cdot \bar{x}_i + b) - 1]$$

$$\frac{\partial L}{\partial \bar{w}} = \bar{w} - \sum \alpha_i y_i \bar{x}_i = 0$$

↑ differentiating vector

(37)

$$w = \sum_i \alpha_i y_i x_i$$

(etc)

I should just look on review sheet how actually used it

And on quiz where made mistake

Prep was on 12/6

↳ 10 days ago

$$\|\bar{w}\| = \sqrt{x^2 + y^2}$$

Lagrangian - function that summarizes dynamics of a system

Decision rule $\sum \alpha_i y \bar{x}_i \bar{v} + b \geq 0$

Can swap in a kernal function

$$\bar{z} = \phi(\bar{x})$$

$$\phi[\bar{x}_i] \cdot \phi(x_j) = k(\bar{x}_i, \bar{x_j})$$

Don't need ϕ, just need k

One such

$$k = e^{\frac{\|x_i - x_j\|^2}{\sigma}}$$

↑ constant - distance each sample has influenced

α = how much we are constraining the width of the road

Solve for

$W_1 x + W_2 y + b = 0$ for dotted line

↑ pts on line → eqn for line

$= 1$ for ⊕ SVM (solid line)

$= -1$ for ⊖ "

Can plug in x, y + solve

$$\sum \alpha_+ = \sum \alpha_-$$

$$\bar{w} = \sum \alpha_\oplus \bar{x}_+ - \sum \alpha_\ominus \bar{x}_-$$

Shortcut d is half of perpindicular bisector

$$d = \frac{1}{\|w\|}$$

ϕ = transform

Solving that actual example as in recitation

List of different kernels

1. Linear

   Like Perceptrons

2. Polynomial

   $(\vec{u} \cdot \vec{v} + b)^n \quad n \geq 1$

   - parabolic, linear, or hyperbolic

   - not circles!

3. Radial / Gaussian

   - contour circles around ⊕ ⊖ points
   - when σ^2 is large - get wider gaussian

Can combine several to get "perfect" fit
↳ ie overfit

4. Sigmodial (tan h) kernal

Allows for combos of linear decision boundries
↳ like neural nets' 2nd neuron

$$k = \tanh\left(k\vec{u}\cdot\vec{v} + b\right)$$

$$= \frac{e^{k\vec{u}\cdot\vec{v}+b} + 1}{e^{k\vec{u}\cdot\vec{v}+b} - 1}$$

5. Linear Combo of kernals

~~left~~ left blank

So now look at what I got wrong on actual exam

Visual linear kernal ✓

Finding values close
↳ def need to review!

41

More complex - fell apart

~~Highorder~~

Higher order kernals ✓

└ what I just wrote

Proposed transformation - totally bomb

Ada boost - close

- Some errors in formula
- need to review + pratice
- but I get it conceptually

don't run out of time!

Problem 3s - need to study on all of them

Review Tutorial before exam

$(V_1 \cdot V_2 + 1) \rightarrow$ linear

$(V_1 \cdot V_2 + 1)^2 \rightarrow$ quadratic

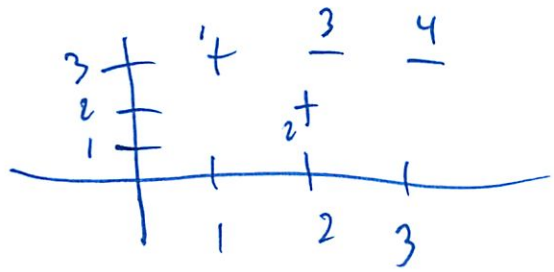
$(V_1 \cdot V_2 + 1)3 \rightarrow$ more degrees of freedom

Small σ = small circle

large σ = large "

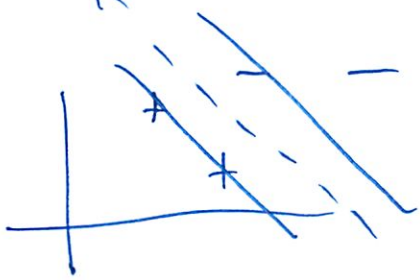
(42)

2006 Final



1. Visually.

✓ got this



(Swap solid + dotted lines)

2. Now numerically

this is what I need help on

$w =$ ← width of road

$\alpha_1 =$

$\alpha_2 =$

$\alpha_3 =$

$\alpha_4 = 0$ ← not SV

Eqn

$$\sum \alpha_i y_i = 0$$

$$\sum \alpha_i y_i x_i = w$$

$$y_i(\bar{w}\bar{x}_i + b) = 1$$

y_i: $+1$, -1

(43) Now do

$$\alpha_1 + \alpha_2 - \alpha_3 = 0$$

$$\alpha_1 x_1 + \alpha_2 x_2 - \alpha_3 x_3 = w$$

known vectors of the pts

Sub in for α_3

$$\alpha_3 = \alpha_1 + \alpha_2$$

$$\alpha_1 x_1 + \alpha_2 x_2 - \alpha_1 x_3 - \alpha_2 x_3 = w$$

$$\alpha_1 (x_1 - x_3) + \alpha_2 (x_2 - x_3) = w$$

Now

$$w x_1 + b = 1$$

$$w x_2 + b = 1$$

$$w x_3 + b = -1$$

Now plug in one for another

$$\alpha_1 (x_1 x_1 - x_1 x_3) + \alpha_2 (x_1 x_2 - x_1 x_3) + b = 1$$

$$\alpha_1 (x_2 x_1 - x_2 x_3) + \alpha_2 (x_2 x_2 - x_2 x_3) + b = 1$$

$$\alpha_1 (x_3 x_1 - x_3 x_3) + \alpha_2 (x_3 x_2 - x_3 x_3) + b = -1$$

(44)

Now solve for 3 unknowns

$\alpha_1 = \alpha_2$ Note here (but not always)

So we never finished that; we should now

Do we have numbers for them

Yes - but are vectors

$X_1 = (1,3)$

$X_2 = (2,2)$

$X_3 = (2,3)$

$X_4 = (3,3)$

But I don't remember vector math on exam

It was

(This must be wrong recitation I am looking at)

In Recitation notes, the kernel fn

$\alpha_A k(X_A, x) + \alpha_B k(X_B, x) + b$

45

So what does it mean for my problem?

$$\alpha_1\left(\begin{bmatrix}1\\3\end{bmatrix}\begin{bmatrix}1\\3\end{bmatrix}-\begin{bmatrix}1\\3\end{bmatrix}\begin{bmatrix}2\\3\end{bmatrix}\right)+\alpha_1\left(\begin{bmatrix}1\\3\end{bmatrix}\begin{bmatrix}2\\2\end{bmatrix}-\begin{bmatrix}1\\3\end{bmatrix}\begin{bmatrix}2\\3\end{bmatrix}+b\right)=1$$

So just upper + lower row

$$\alpha_1(1-2)+\alpha_1(2-2)+b=1$$

$$-1\alpha_1+b=1$$

$$0\alpha_1+-3\alpha_1+b=1$$

$$\alpha_1\left(\begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}1\\3\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}2\\3\end{bmatrix}\right)+\alpha_1\left(\begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}2\\2\end{bmatrix}-\begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}2\\3\end{bmatrix}+b\right)=1$$

$$\alpha(-2)+\alpha(0)+b=1$$

$$\alpha(0)+-2\alpha+b=1$$

$$-2\alpha+b=1$$

? It does not add up!

I think I am doing this wrong

it's not this complex!

(46)

Escntially

$$W_1 x + W_2 y + b = 0$$ dotted

↑ quiz asks ↗ ya to solve

1 ⊕

-1 ⊖

$$\sum \alpha_\oplus = \sum \alpha_\ominus$$

$$w = \sum \alpha_\oplus \bar{x}_+ - \sum \alpha_- \bar{x}_-$$

Can find w first

2D = distance bw lines

$$\sqrt{y^2 + y^2}$$

$$\sqrt{(\Delta x)^2 + (\Delta y)^2}$$

$$D = \frac{1}{\|w\|}$$

$$w = \sqrt{2k^2}$$

$$\uparrow \sqrt{w_1^2 + w_2^2}$$

(47)

Here $\bar{w} = \begin{bmatrix} -1/4 \\ 1/4 \end{bmatrix}$ $b = 1/4$

How did they find these things?

Now use to solve

$$\begin{bmatrix} -1/4 \\ 1/4 \end{bmatrix} = \alpha \begin{bmatrix} -1 \\ 2 \end{bmatrix} - \alpha \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

?

since

$\alpha_+ = \alpha_-$

$$= \begin{bmatrix} -4 \\ 4 \end{bmatrix} \alpha$$

$$\alpha = \frac{1}{16}$$

Conceptually what is w?

w is the weight vector

(still don't fully get)

(48)

$$\|w\| = \sqrt{\sum w_i^2}$$

↑ want

such that $y_i(w_i x_i + b) \geq 1$

$w_1 = c \quad w_2 = -c$

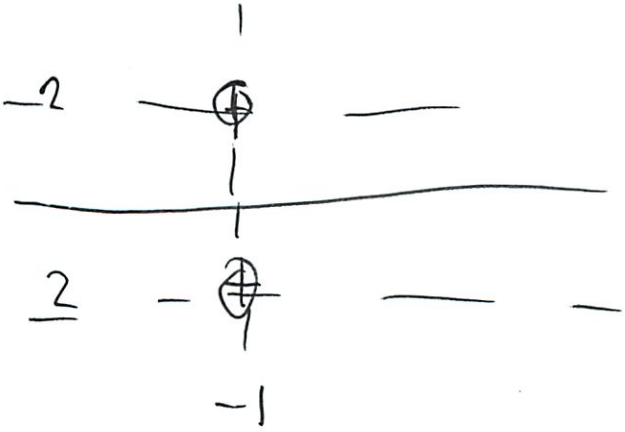
$$\frac{2}{\sqrt{\sum w_i^2}} = \frac{2}{\sqrt{c^2 + c^2}} = \frac{2}{\sqrt{2c^2}} = 2\sqrt{2}$$

still can't find good example!

So I think I am getting it...

So start w/ w

Let me look at this exam



So w is vector to middle $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$

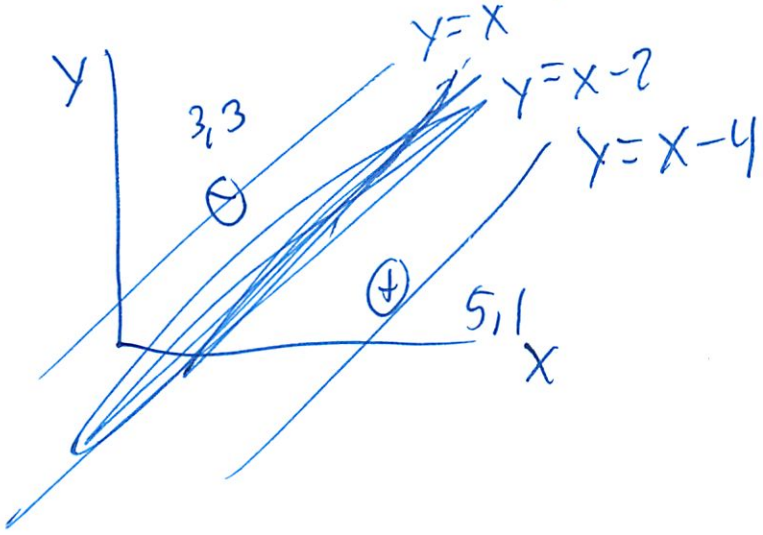
(49)

Or vector of solid black line:

$d = \frac{1}{\|w\|}$ = half of perpendicular bisector

So the solid black line

Or other fact - from parent



$y \leq x - 2$

↑ everything less than is ⊕

~~Free~~

$1x - 1y - 2 \geq 0$

Ohhh w_1 w_2 b

(50)

So for exam eqn of line is $y = 0x + 0$

So $0x - 1y + 0 = 0$

$\begin{bmatrix} 0 \\ -1 \end{bmatrix}$

I answered $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ sols have $\begin{bmatrix} 0 \\ 1/2 \end{bmatrix}$

But I remember reading lots of w/bs that are multiples of each other!

Then others

$y_+ = 0x + 2$

$y_- = 0x - 2$

So what do w/ this

$0x - y_+ + 2 = 1$

this somehow is 1

Or

$0x - y + 2 = 1$

$0x - x - 2 = -1$

(51)

(I wish they gave solutions... or that I asked in OH)

So the way did this packet problem (going back)

was diff in recitation

margin $= 2\sqrt{2}$

$y = x - 4$ (+)

$y = x - 2$ line

$y = x$ (−)

$x - y - 2 \geq 0$

^No can't just read off line

$\sqrt{w_x^2 + w_y^2} = \sqrt{2}$

$Cx - Cy - 2C \geq 0$

$w_1 = C \quad w_2 = -C$

$\frac{2}{\sqrt{C^2 + (-C)^2}} = \frac{2}{\sqrt{2C^2}}$

(52)

Set = t_0

$$\frac{2}{\sqrt{2c^2}} = 2\sqrt{2}$$

$$\frac{4}{2c^2} = 8$$

$$c = \frac{1}{2}$$

$w_1 = \frac{1}{2}$

$w_2 = -\frac{1}{2}$

$b = -1$

So is that what I forgot on exam

distance here is 4

$$\frac{2}{\sqrt{2c^2}} = 4$$ (right?)

$$\frac{4}{2c^2} = 16$$

$$4 = 32c^2$$

(53)

$\frac{4}{32} = c^2 = \frac{2}{16} = \frac{1}{8}$

$\sqrt{\frac{1}{8}} =$ what again?

have II -84

$\frac{\sqrt{2}}{4}$

hmm - that does not seem right -

- and no place to look for answers

Oh did I skip ahead too fast

$W_1 = 0$

$W_2 = -C$

So $\frac{2}{\sqrt{0^2 + c^2}} = \frac{2}{c}$

$\frac{2}{c} = 4$

$2 = 4c$

$\frac{1}{2} = c$

So $v = \begin{bmatrix} 0 \\ -1/2 \end{bmatrix}$ $b = 0$

but sign is still off

(54)

Am I writing sign right

$y \geq 0$

↑ So don't move y
just read like that

$0x + 1y + 0 \geq 0$
↑ b

That would make more sense

Now plug in αs

$\alpha_3 = \alpha_4$ others 0

~~$+b$~~ ← b not here I think

$$\alpha_3 (1) \begin{bmatrix} -1 \\ 2 \end{bmatrix} + \alpha_4 (-1) \begin{bmatrix} -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ +\frac{1}{2} \end{bmatrix}$$

(1) ↑ ⊕; $\begin{bmatrix} -1 \\ 2 \end{bmatrix}$ ↑ coords; (−1) ↑ ⊖; result ↑ $\vec{w}$

$-\alpha_3 + \alpha_4 = 0$ ← if $\alpha_3 = \alpha_4$ - not useful

$2\alpha_3 + 2\alpha_4 = +\frac{1}{2}$

$4\alpha_3 = +\frac{1}{2}$

$8\alpha_3 = +1$ $\alpha_3 = +\frac{1}{8} = \alpha_4$

✓ Sign error but much more wrong
think grader was too lenient
— made a larger mistake

55

Oh Sign error was transfered wrong

Think I figured it out finally

↳ this always takes me a bunch of pages!

Bread

Now see how much I remember

Do exam problem from memory

$y = 0$

$y + 2 = 1$

$y - 2 = -1$

$y \leq 0$

is +

$0x + 1y + 0 \leq 0$

7b

0c

1c

(56)

Distance = 4

This is 2 (believed constant) $\frac{2}{\sqrt{0^2+c^2}} = \frac{2}{c^2}$

$\frac{2}{c^2} = 4$

$2 = 4c^2$

$\frac{1}{2} = c^2$

Calc $c = \frac{\sqrt{2}}{2}$

Still something wrong I think...

Oh I did calc wrong!

$\frac{2}{c} = 4$

$2 = 4c$

$\frac{1}{2} = c$

So $w = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$ $b = 0$

(57)

Then

$$\alpha_3(1)\begin{bmatrix}-1\\2\end{bmatrix} + \alpha_4(-1)\begin{bmatrix}-1\\-2\end{bmatrix} = \begin{bmatrix}0\\1/2\end{bmatrix}$$

$$\alpha_3 = \alpha_4$$

$$-\alpha_3 + \alpha_4 = 0$$

$$2\alpha + 2\alpha = \frac{1}{2}$$

$$4\alpha = \frac{1}{2}$$

$$8\alpha = 1$$

$$\alpha = \frac{1}{8}$$ woot!

Now do b-side

(-1,-2)

w_1 b unchanged I believe

(58)

$$\alpha_3 (1)\begin{bmatrix}-1\\2\end{bmatrix} + \alpha_4 (-1)\begin{pmatrix}-1\\-2\end{pmatrix} + \alpha_5 (-1)\begin{pmatrix}1\\-2\end{pmatrix} = \begin{pmatrix}0\\1/2\end{pmatrix}$$

$\alpha_3 = \alpha_4 + \alpha_5$

"so then how to solve that"

$$-\alpha_3 + \alpha_4 - \alpha_5 = 0$$

$$2\alpha_3 + 2\alpha_4 + 2\alpha_5 = \frac{1}{2}$$

"do $\alpha_4 = \alpha_5$"

try to work w/o try stuff

$$-\alpha_4 - \alpha_5 + \alpha_4 - \alpha_5 = 0$$

$$-2\alpha_5 = 0$$

~~can be any~~ must be 0

$$2\alpha_4 + 2\alpha_5 + 2\alpha_4 + 2\alpha_5 = \frac{1}{2}$$

$$4\alpha_4 + 4\alpha_5 = \frac{1}{2}$$

I think $\alpha_4 = \alpha_5$

So $\frac{1}{8}$ $\frac{1}{16}$ $\frac{1}{16}$

(59)

What's the other rule

$$\sum a_i y_i = 0$$

$\uparrow +-$

So that just means

$$\alpha_3 - \alpha_4 - \alpha_5 = 0$$

$$\alpha_3 = \alpha_4 + \alpha_5$$

$\uparrow P$

but can't say they are evenly split

Answer

$\alpha_3 = \frac{1}{8}$ $\alpha_7 = 0$

$\alpha_4 = \frac{1}{8}$

Ah so it did not change

L the new pt not constraining solution

So my result that it was 0 was right

Should have trusted it

Lnext time I know

(40)

A4 Scale data a 10 times

So do it fully at

Lon exam guessed since short on time

~~Quiz~~ $y \geq 0$

$\frac{2}{c} = 40$

$2 = 40c$

$\frac{2}{40} = c = \frac{1}{20}$

$w = \begin{bmatrix} 0 \\ 1/20 \end{bmatrix}$ $b = 0$ (1)

trying anyway

$\alpha_3 (1)\begin{bmatrix} -10 \\ 20 \end{bmatrix} + \alpha_4 (-1)\begin{pmatrix} -10 \\ -20 \end{pmatrix} + \alpha_7 (-1)\begin{pmatrix} 10 \\ 20 \end{pmatrix} = \begin{bmatrix} 0 \\ \frac{1}{20} \end{bmatrix}$

$-10\alpha_3 + 10\alpha_4 - 10\alpha_7 = 0$

$\alpha_3 = \alpha_4 + \alpha_7$

$-10(\alpha_4 + \alpha_7) + 10\alpha_4 - 10\alpha_7 = 0$

$-20\alpha_7 = 0 \rightarrow \alpha_7 = 0$

(61)

$$20\alpha_3 + 20\alpha_4 + 20\alpha_7 = \frac{1}{20}$$

$$20\alpha_4 + 20\alpha_7 + 20\alpha_4 + 20\alpha_7 = \frac{1}{20}$$

$\alpha_7 = 0$

$$40\alpha_4 = \frac{1}{20}$$

$$800\alpha_4 = 1$$

$$\alpha_4 = \frac{1}{800}$$ ✓

$$\alpha_3 = \alpha_4 + \alpha_7$$

$$\alpha_3 = \frac{1}{800} + 0$$

$$\alpha_3 = \frac{1}{800}$$ ✓ ✓ Perfect

Ok try some other problems

✓ Done basic linear kernal problems

Now kernal functions

So what was ans to my quiz?

Well first correctly identified other kernals that would work

(62)

But B2

12|st dat pt for above
Will it classify correctly

I said No - overfitting

But ans = yes

So it could be that 1. Yes since retrain learn
2. No since no retrain

So I think I missed the retrain

We are overfitting to extream
but retrain 1st

B4 (N. Give a transformation $\phi(v)$? vector whose components $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$
that will make points linearly separable

$q = v_1 v_2$

So essentially $\pm$) $-3 \cdot 3 = -3$
$3 \cdot -3 = -3$
$-$) $3 \cdot 3 = 3$
$-3 \cdot -3 = 3$
} flipped

(63)

So why not $-U_1 U_2$

Variety of solutions possible

But I didn't know what to put down

$k(v,v) = \phi(v) \cdot \phi(v) = U_1 U_2 V_1 V_2$

is this always true

↳ Yes I think I wrote it on quiz

I see how got this

↳ but how to think of it earlier

I was thinking curve

Not just multiply values together

ie another sol $\phi(v) = |U_2 - U_1|$

From original lecture

$\bar{z} = \phi(x)$

$k(x_i, x_j) = \phi(x_i)\phi(x_j)$

(64)

Basic kernal

dot product $x \cdot y$

Often gv k if have ϕ

or ϕ " k

just $k = \phi(v)\,\phi(v)$

But what abat if

$$k(\vec{v}, \vec{v}) = -\exp\left(-\frac{\|\vec{v} - \vec{v}\|^2}{2\sigma^2}\right)$$

How seperate this to ϕ ?

Actually $k(\vec{v}; \vec{v}) = \vec{v} \cdot \vec{v} + b$ for linear

So $k(v, v) = \phi(v) \cdot \phi(v)$ the dot product in transformed space

But this seems like they would not ask

Basically ϕ = transform to other dimension

And k is what actually gives ya the values

Boosting

(need a lot of pratice w/ this)

Trying to combine a bunch of weak (crappy) classifiers into a strong classifier $H(x)$ $h_1(x), h_2(x) \ldots$

$$H(x) = \text{signex}\left(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \ldots\right)$$

$$= \text{signex} \sum_{i=1}^{S} a_i h_i(x)$$

$$H(x), h_i(x) = \begin{cases} 1 \oplus \\ -1 \ominus \end{cases}$$

Each classifier is weighted

Weights must = 1

Pick stump classifier so error rate $< \frac{1}{2}$

↳ or flip

Errors

$$E^s = \sum w_{i,wrong}$$

$$(1 - E^s) = \text{correct}$$

$$E^s < \frac{1}{2}$$

so $E^s < 1 - E^s$

$$\frac{1 - E^s}{E^s} > 1$$

$$\alpha_s = \frac{1}{2} \ln \frac{1 - E^s}{E^s}$$

Steps Oh also the pre steps of picking classifiers

1. Initalize $w_i = \frac{1}{n}$
2. Pick classifier w/ lowest error rate
3. Compute $\alpha_s = \frac{1}{2} \ln \frac{1 - E^s}{E^s}$

(67)

Update weights

Correct

$$W_i^{s+1} = \frac{1}{2}\frac{1}{1-E^s} W_i^s$$

Incorrect

$$W_i^{s+1} = \frac{1}{2}\frac{1}{E^s} W_i^s$$

3. Terminate if $s > T$

↳ basically terminate after certain # of runs

4. Output final classifier

$$H(x) = \text{sign}\left(\sum_i a_i h_i(x)\right)$$

Sounds easy - can I do it?

Try my exam again

Elfs and Magic

↑ goal

(68)

Ok try choosing classifiers
└ how could I get this wrong
└ pretty stupid

As I was thinking it over I made same mistake
Just need to think clearly - since know

A2] I figured this out

A3] Should we use all of them
Oh forgot to write those which strictly worse
(lots of oversight on this exam!)
So try now

c, b, g, e, f ← forgot one
anything w/ P (I don't see that one...)
can it be a composite
review mega Q
(where I did not take notes)

It mentions a pair - confused, email in

(69)

Ohh I see -j !!

(~~very~~ need to look hard)

So let me try again

f, c, e, g, b ✓

8

Now actual boosting

1	1/10	1/14
2	1/10	1/14
3	1/10	1/14
4	1/10	1/14
5	~~1/10~~	1/14
6	1/10	1/14
7	1/10	1/14
8	1/10 ×	7/14

---

h d

ε 1/10 only 8 total!

α $\frac{1}{2}\ln\left(\frac{9/10}{1/10}\right)$ can leave five times

(20)

Use calc - did not have on exam
& use normal formula - no tricks

So for corrects $\frac{1}{2}\frac{1}{7/8}w_i = \frac{4}{7}w_i \rightarrow \frac{1}{14}$

incorrect $\frac{1}{2}\frac{1}{1/8}w_i = 4w_i \rightarrow \frac{1}{2}\ \frac{7}{14}$

So next classifier
Something w/ 3 (9, i, j)
and does not include 8 (i, j)
Use reverse alpha order

$E = \frac{3}{14}$

$\frac{1}{2}\ln\left(\frac{11/14}{3/11}\right) = \frac{1}{2}\ln\left(\frac{11}{3}\right)$

(for use later - when put together)

Correct $\frac{1}{2}\cdot\frac{1}{11/14}\cdot w_i = \frac{7}{11}w_i \rightarrow 1/22$
$7/22$

inc $\frac{1}{2}\cdot\frac{1}{3/11}\cdot w_i = \frac{7}{3}w_i \rightarrow \frac{1}{6}$ ✓

(remember this is 5, 6, 7)

(71)

~~But error is 2~~

Next I w/ 2, 3, 6 ✓

$$\frac{1}{22} + \frac{1}{22} + \frac{1}{6} = \frac{17}{66}$$

(did not have calc to do right!)

$$\frac{1}{2} \ln\left(\frac{49/66}{17/66}\right) = \frac{1}{2} \ln \frac{49}{17}$$

Sunk by my calculator
└ well only 2 pts

$$H(x) = \left(\frac{1}{2} \ln(7)_D + \frac{1}{2} \ln\left(\frac{11}{3}\right)_J + \frac{1}{2} \ln\left(\frac{49}{17}\right)_I\right)$$ Signed

Then did not have enough time/calc ability to try
└ So what would we actually do again?

So for #1 Link

$$\frac{1}{2} \ln(7)(1) + \frac{1}{2} \ln\left(\frac{11}{3}\right)(1) + \frac{1}{2} \ln\left(\frac{49}{17}\right)(+1)$$

Opps caught something graders didn't
I = Magic yes

Don't need to calc -all 1 so DEIF ✓ correct

(12)

This seems slow

#2 Answer

1 1 -1

So now calc
"or see 2 are 1 so good"
what are each factor

D = .97 J = .64 H = .52

So .97 + .64 − .52 = (+)

#3 1 1 1

can look at mistakes

#4 1 1 1

#5 1 -1 1 still (+) ✓

#6 actually, this should be wrong -1 -1 So now = -.19 (−) ✓

#7 1 -1 1 (+) X Wrong

(13)

What are we returning 1 on?

Not really clear — 1 if ☺, -1 if ☹

But what does this mean?

+1 if true

-1 if wrong

I was doing + if correct

Lah but we don't know if correct

Only in training

So

#6 Fran

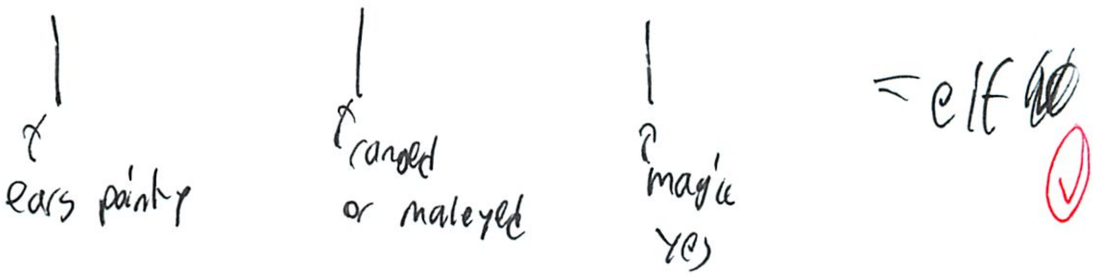
-1 (✗ cass not pointy (right should be -1 since NOT an elf)) | 1 (✗ fighting is ranged (incorrect)) | 1 (? magic is yes) = 1 elf

X but not ✓

(takes a while!)

(74)

So our new position



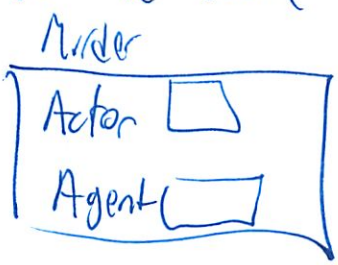
Now do some other boosting qu

(✓) Done

Now read about other stuff in chap

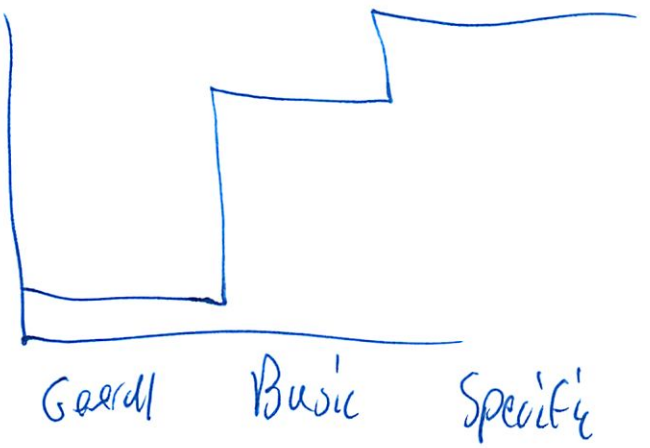
Semantic Nets

Looking at a frame



Sequence

1. Classification



Musical Int
↑
Piano
↑
Bosendopfer

2. Transitions

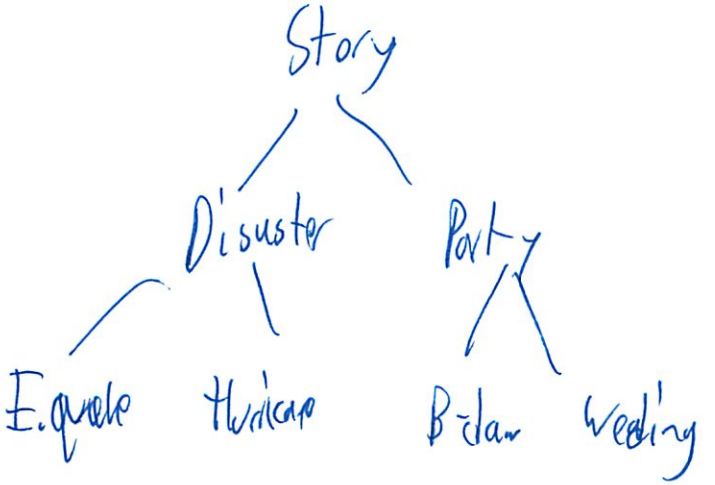
What exam has on

↑, ↓, ◁, A, D

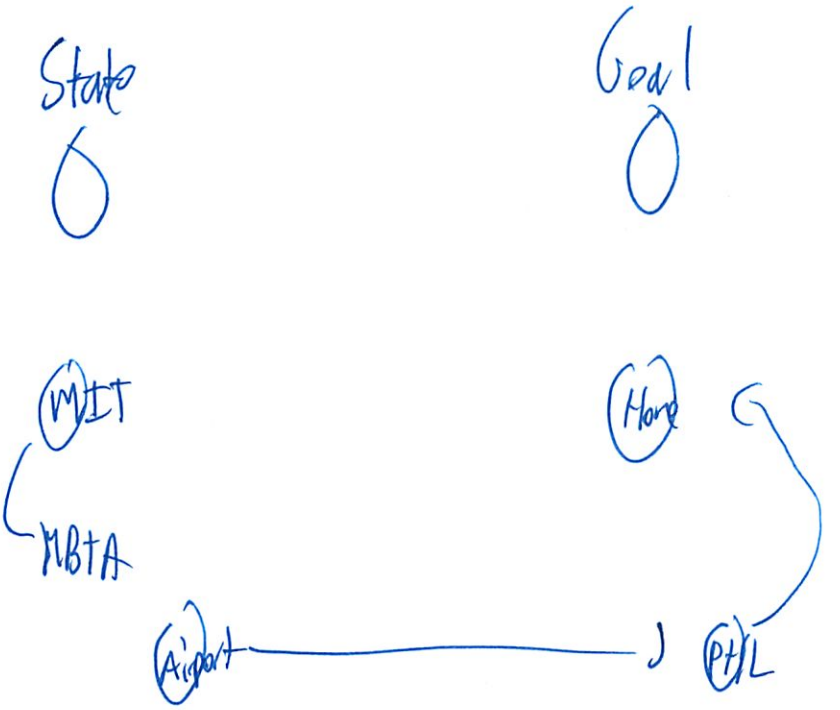
3. Trajectory

Propositions

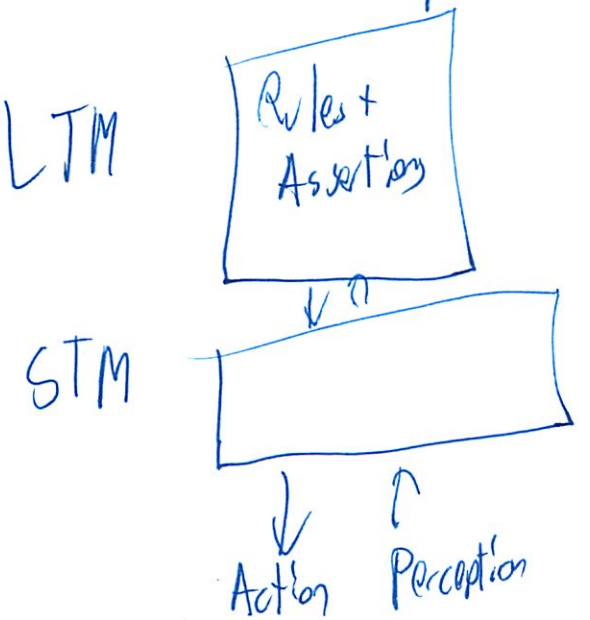
Frame of stories



General problem Solver



Soar State Operator and result



(77)

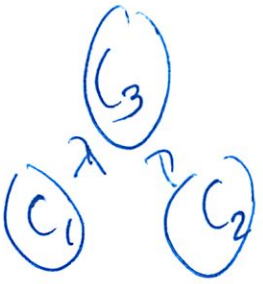
Emotion machine

↑ Philopsory

Subsumption Architure

	V	R	A
Seek	V	R	A
Explore	V	R	A
Wander	V	R	A
Avoid	V	R	A

Chomsky



Developmental

White rooms
Spin cat

6.034 Restudy

12/20

Ok exam tomorrow
Re pratice key stuff

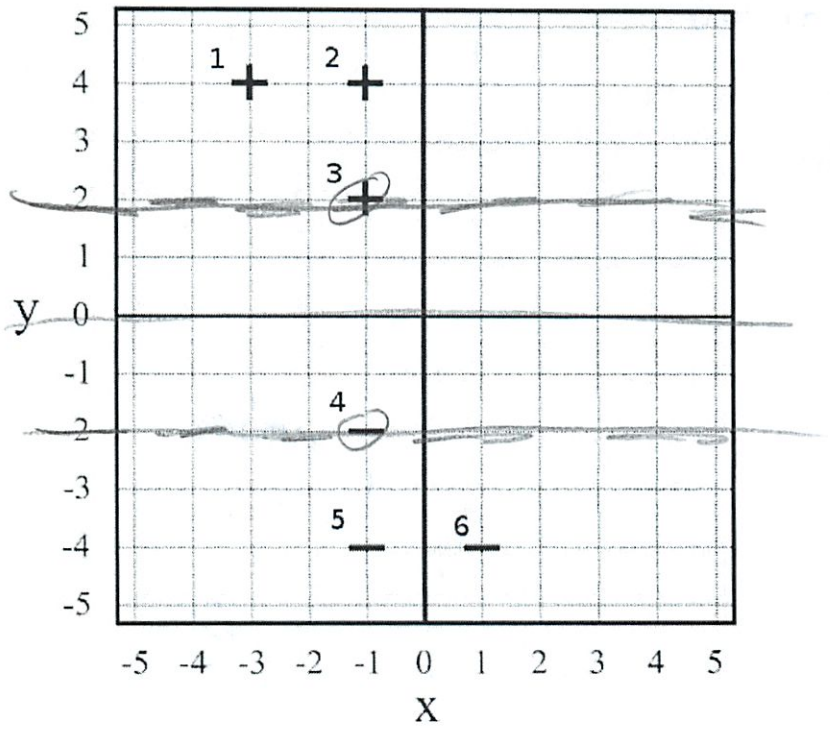
SVMs
Boosting
L, B

Do examples from my exams

Problem 1: SVMs (45 points)

Part A: Linear kernels (22 points)

Your good friend, Khan Fusion, tells you about a binary classification technique known as a support vector machine, which appears to be all the rage among computer scientists these days. Unfortunately, Khan is a busy man, and so he could only spend 50 minutes giving you a whirlwind tour to the complex world of SVMs. Eager to see if you're on the right track, you create the following toy classification problem and solve it using an SVM to get a feeling for how it works:



A1 (6 points)

On the picture above, draw with a **heavy solid line** the optimal decision boundary (that is, where the classifier outputs exactly 0) found by an SVM using a linear kernel. Draw with a **heavy dashed line** the edges of the "street" produced by the SVM (that is, where the classifier outputs exactly +1 or -1.) **Circle all support vectors.**

(3)

A2(6 points)

Give the values that the SVM finds for the following variables; note that **w** is a vector:

w = $\begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$ $\alpha_1 = 0$ $\alpha_3 = \frac{1}{8}$ $\alpha_5 = 0$

$\alpha_2 = 0$ $\alpha_4 = \frac{1}{8}$ $\alpha_6 = 0$

b = 0

Show your work (for partial credit):

$y \leq 0$ do ≤ 0 is (-)

$y \geq 0$ is (+)

Write like

$0x + 1y + 0 \leq 0$

always $\leq$

So

$w = \begin{bmatrix} 0c \\ 1c \end{bmatrix}$ $b = 0c$

Now distance b/w = 4

$\frac{2}{\sqrt{w}} = 4$

$\frac{2}{\sqrt{0^2 + c^2}} = \frac{2}{c} = 4$

$2 = 4c$

$c = \frac{1}{2}$

$w = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$

$b = 0$

Now this w/ the vectors

$(1)\begin{pmatrix} -1 \\ 2 \end{pmatrix}\alpha_3 + (-1)\begin{pmatrix} -1 \\ -2 \end{pmatrix}\alpha_4 = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}$ ✓

$\alpha_3 = \alpha_4$

$-\alpha_3 + \alpha_4 = 0$

$\alpha_4 = \alpha_3$ ← mirrored

$2\alpha_3 + 2\alpha_4 = \frac{1}{2}$

$4\alpha_3 = \frac{1}{2}$

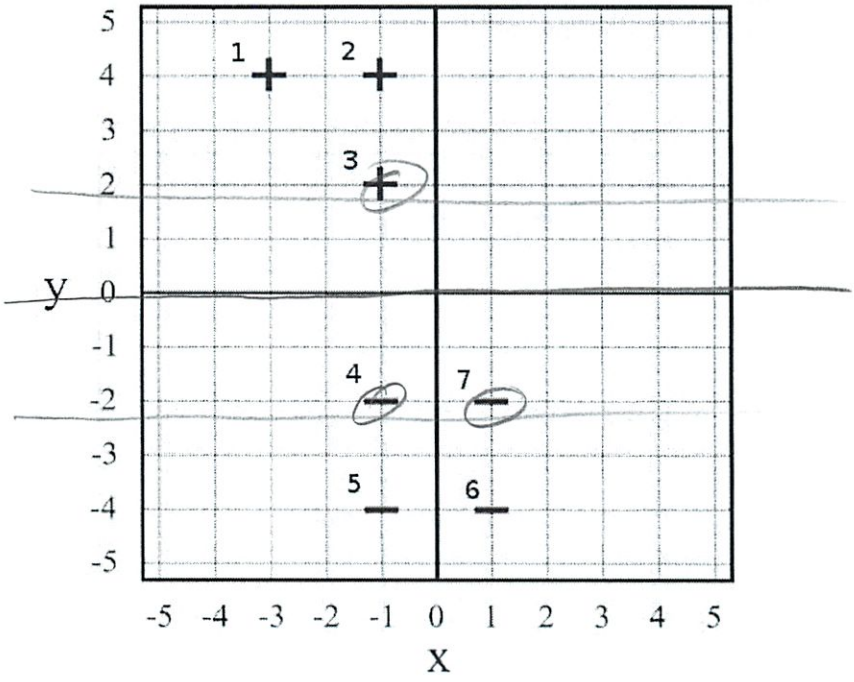
$\alpha_3 = \alpha_4 = \frac{1}{8}$ (C)

lots of mistakes

but can look up how to do

A3(5 points)

You now add a seventh data point to your graph as follows:



A3a Draw the decision boundary with a **heavy solid line.**

A3b Now what values will the SVM find for each of the following variables?

$\mathbf{w} = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$ $b = 0$

$\alpha_1 = 0$ $\alpha_2 = 0$ $\alpha_3 = \frac{1}{8}$ $\alpha_4 = \frac{1}{8}$ $\alpha_5 = 0$ $\alpha_6 = 0$ $\alpha_7 = 0$

$0 + y + 0 \leq 0$ is 0

$w = \begin{bmatrix} 0 \\ c \end{bmatrix}$ $b = 0c$

$4 = \frac{2}{\sqrt{0^2 + c^2}}$

$4c = 2$

$c = \frac{1}{2}$

$\begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}$ $b = 0$

$(1)\begin{bmatrix} -1 \\ 2 \end{bmatrix}\alpha_3 + (-1)\begin{bmatrix} -1 \\ -2 \end{bmatrix}\alpha_4 + (-1)\begin{pmatrix} 1 \\ -2 \end{pmatrix}\alpha_7 = \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix}$

$\alpha_3 = \alpha_4 + \alpha_7$

$-\alpha_3 + \alpha_4 - \alpha_7 = 0$

$2\alpha_3 + 2\alpha_4 + 2\alpha_7 = \frac{1}{2}$

$-(\alpha_4 + \alpha_7) + \alpha_4 - \alpha_7 = 0$

$-2\alpha_7 = 0$

$\alpha_7 = 0$

$2\alpha_3 + 2\alpha_4 = \frac{1}{2}$

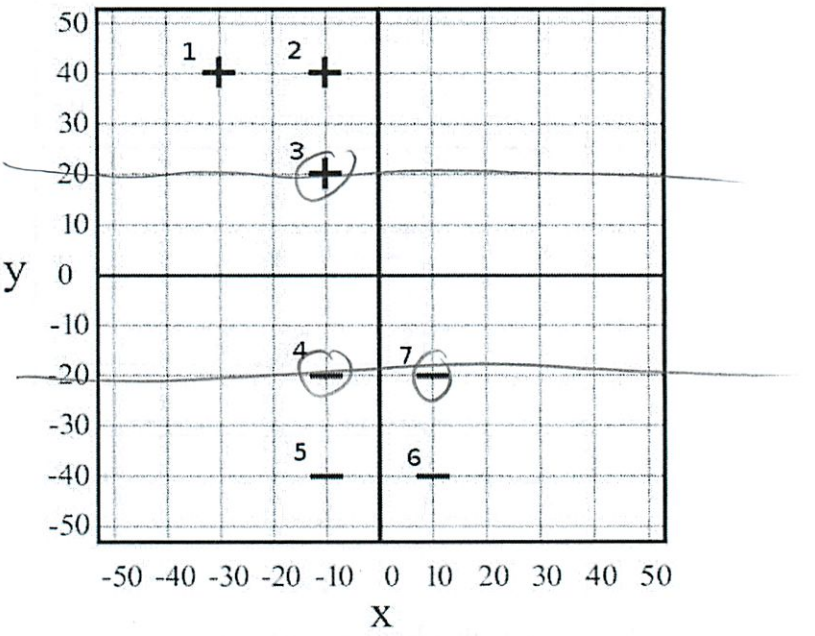
$4\alpha = \frac{1}{2}$

$\alpha = \frac{1}{8}$

(5)

A4(5 points)

You want to see how the SVM will react to changing the scale of your problem, so you decide to scale your data by 10 along both dimensions:



What values will the SVM find for each of the following variables?

$\mathbf{w} = \begin{bmatrix} 0 \\ 1/20 \end{bmatrix}$ $\quad b = 0$

$\alpha_1 = 0$ $\quad \alpha_2 = 0$ $\quad \alpha_3 = \frac{1}{800}$ $\quad \alpha_4 = \frac{1}{800}$ $\quad \alpha_5 = 0$ $\quad \alpha_6 = 0$ $\quad \alpha_7 = 0$

~~$y \geq 0$~~

$y \leq 0$ ↑ is format!

$\begin{bmatrix} 0 \\ c \end{bmatrix}$ $b = 0c$

$40 = \frac{2}{\sqrt{0^2+c^2}}$

$40c = 2$

$c = \frac{1}{20}$

$\begin{bmatrix} 0 \\ 1/20 \end{bmatrix}$ 0

$(1)\begin{pmatrix} -10 \\ 20 \end{pmatrix}\alpha_3 + (-1)\begin{pmatrix} -10 \\ -20 \end{pmatrix}\alpha_4 + (-1)\begin{pmatrix} 10 \\ -20 \end{pmatrix}\alpha_7 = \begin{bmatrix} 0 \\ \frac{1}{20} \end{bmatrix}$ ← b means nothing here!

$-10\alpha_3 + 10\alpha_4 - 10\alpha_7 = 0$

$20\alpha_3 + 20\alpha_4 + 20\alpha_7 = \frac{1}{20}$

$\alpha_3 = \alpha_4 + \alpha_7$

$-10(\alpha_4 + \alpha_7) + 10\alpha_4 - 10\alpha_7 = 0$

$-20\alpha_7 = 0$

$\alpha_7 = 0$

$20\alpha_3 + 20\alpha_4 = \frac{1}{20}$

$40\alpha_3 = \frac{1}{20}$

$\alpha_3 = \frac{1}{20 \cdot 40} = \frac{1}{800}$

5

Part B: Higher order kernels (23 points)

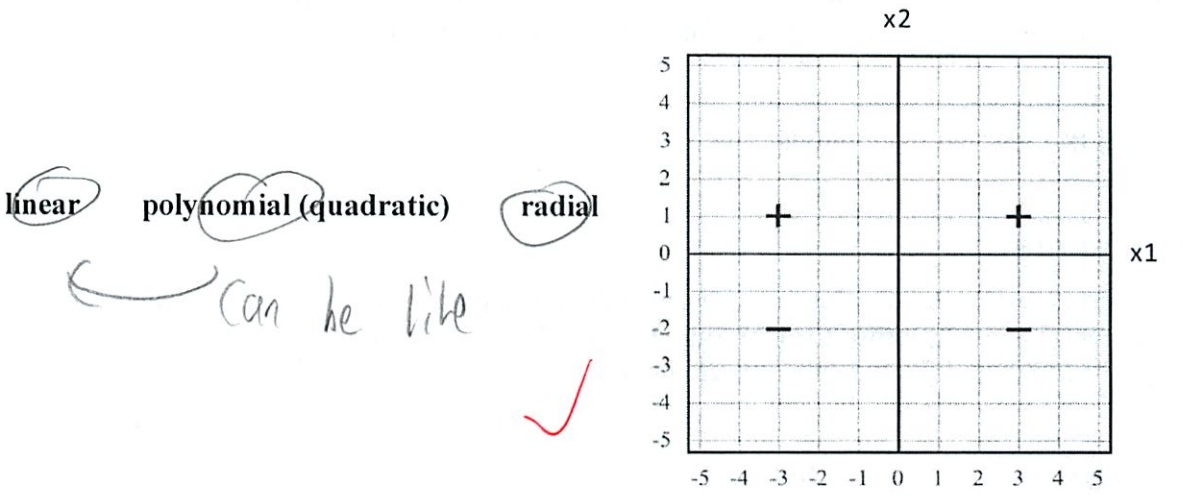
You're pretty satisfied with your understanding of SVMs at this point, so you decide to use them to tackle a slightly harder problem, certainly worthy of an MIT student's attention: given two input **integers** x and y, can we learn to predict whether their sum, x + y, will be evenly divisible by 2?

(Note: for the purpose of this problem, we consider 0 to be evenly divisible by 2.)

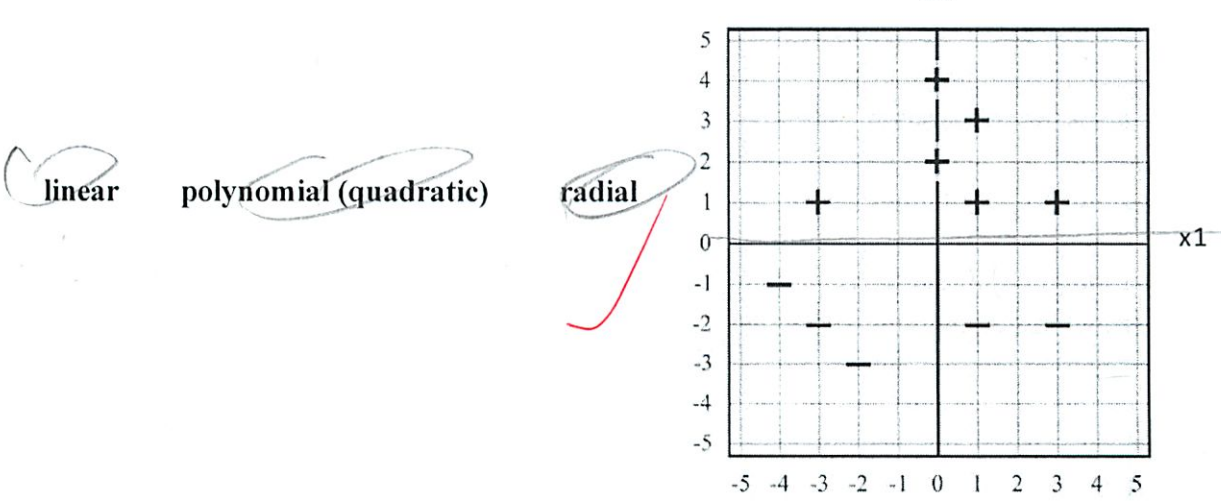
B1 (9 points)

You slowly start to populate your training set by inserting data points which you manually classify.

For each of the following graphs, **circle ALL kernels that can perfectly classify the training data.** Your three options are: **linear** kernels, **quadratic polynomial** kernels, and **radial basis function** kernels.

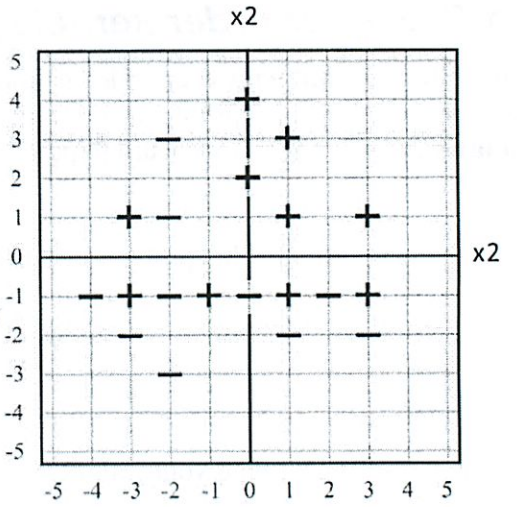


linear **polynomial (quadratic)** **radial**



linear **polynomial (quadratic)** **radial**

(7)



linear **polynomial (quadratic)** **radial**

L parabolic
linear
hyperbolic
Not circles

Overfit

B2 (4 points)

You heave a sigh of victory as you finish adding the 121st **distinct** data point to the graph above. You then run your learning algorithm with a radial basis function kernel using suitable parameters. Then, you test your classifier with two random integers between -5 and 5. Will your classifier always be able to classify this data point correctly? (**circle one**)

YES **NO**

B3 (4 points)

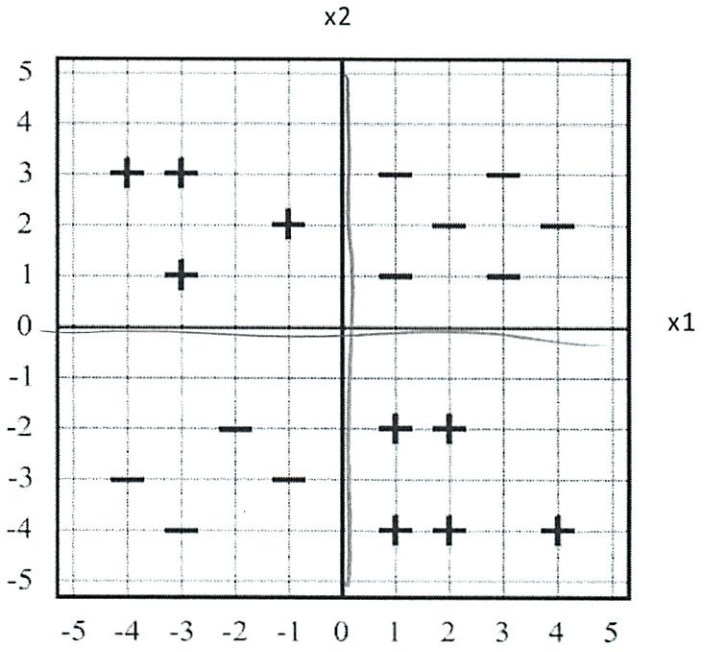
You decide to test your classifier from B2 on random data points outside the range where x and y are between -5 and +5, **without retraining your SVM**. Will your classifier always be able to classify these data points correctly? (**circle one**)

I remember ars

YES **NO**

B4 (6 points)

Confused by your results above, you decide that you want to get a more intuitive sense of how kernels work, so you turn your attention to a simpler problem:



Give a transformation $\varphi(\mathbf{u})$ (where $\mathbf{u}$ is a vector whose components are u_1 and u_2) that will make these data points linearly separable, and compute the kernel $\mathbf{K(u, v)}$ associated with the transformation.

$\varphi(\mathbf{u})$: $u_1 \cdot u_2$ ✓

$\mathbf{K(u, v)}$: $u_1 \cdot u_2 \cdot v_1 \cdot v_2$ ✓

Problem 2: Adaboost (45 points)

Your friend Ben has spent the last month playing a new video game instead of attending 6.034. This weekend his girlfriend Brittney is visiting, and they want to make a character for her, but don't know what sort of character it should be. She has given him a description of her ideal character, and Ben has asked you to use your newfound knowledge of Adaboost to build a classifier that determines if she should play an elf or a non-elf.

Below is the table that Ben has compiled of training data for your classifiers. Note that +1 in the Elf column means that the person should be an elf; -1 means the person should not be an elf.

	Person	Elf	Fighting Style	Ears	Magic	Size
1	Link	+1	Ranged	Pointy	Yes	Medium
2	Arwen	+1	Melee	Pointy	No	Medium
3	Legolas	+1	Ranged	Pointy	No	Medium
4	Dobby	+1	Ranged	Pointy	Yes	Small
5	Christmas Elf	+1	None	Pointy	Yes	Small
6	Fran	-1	Ranged	Round	Yes	Medium
7	Green Arrow	-1	Ranged	Round	No	Medium
8	Gizmo	-1	None	Pointy	No	Small

Part A: Choosing Classifiers (18 points)

A1 (10 points)

Fill in the following chart by indicating which rows, by number, have the wrong Elf value, given the indicated test.

This job sucks

Classifier	Test	Misclassified
a	Fighting Style = Melee is elf	1, 3, 4, 5,
b	Fighting Style = Ranged	2, 5, 6, 7
c	Fighting Style = None	1, 2, 3, 4, 8
d	Ears = Pointy	8
e	Magic = No is elf	1, 4, 5, 7, 8
f	Size = Medium	4, 5, 6, 7
g	True	6, 7, 8

(10)

A2 (4 points)

You notice that you could add two more good, single test, weak classifiers (fewer than half of the data points are misclassified). Fill in the tests below, given the data points they misclassify.

We have not used **h** in the table to avoid confusion with the weak classifiers that are added up to make the strong classifier, H.

Classifier	Test	Misclassified
i	Magic = Yes	2, 3, 6
j	Fighting ≠ None	5, 6, 7

A3 (4 points)

Ben thinks we should use all 9 of these classifiers in boosting, just to be safe. Do you agree? Circle one:

No can eliminate duplicate

YES (**NO**)

If you circled NO, list the classifier(s) (by letter) that we will NOT need to use during boosting:

c e g b f

(11)

Part B: Running Adaboost (27 points)

B1 (18 points)

Now that you know which weak classifiers you'll use (either all 9 or some subset, depending on your answer to part A3), you're ready to run Adaboost. Fill in the table below for the first three rounds of Adaboost. Break ties by **REVERSE alphabetical order** of the classifier.

	Round 1	Round 2	Round 3
w1	1/8	1/14	1/22
w2	1/8	1/14	1/22
w3	1/8	1/14	1/22
w4	1/8	1/14	1/22
w5	1/8	1/14 X	1/6
w6	1/8	1/14 X	1/6
w7	1/8	1/14 X	1/6
w8	1/8 X	1/2	7/22
h	d	j	i
ε	1/8	3/14	$\frac{2}{22} + \frac{1}{6} = \frac{17}{66}$
α	$\frac{1}{2}\ln\frac{7/8}{1/8}$	$\frac{1}{2}\ln\frac{11/14}{3/14}$	$\frac{1}{2}\ln\frac{49/66}{17/66}$

You can use the space below to show your work:

Correct $\frac{1}{2} \cdot \frac{1}{7/8} \cdot w_i = \frac{4}{7} \cdot \frac{1}{8} = \frac{1}{14}$

incorrect $\frac{1}{2} \cdot \frac{1}{1/8} \cdot \frac{1}{8} = \frac{1}{2}$

Correct $\frac{1}{2} \cdot \frac{1}{11/14} \cdot w_i \quad \frac{1}{14} \to \frac{1}{22}$

incorrect $\frac{1}{2} \cdot \frac{1}{3/14} \cdot \frac{1}{14} = \frac{1}{6}$ $\quad \frac{2}{22}$

(12)

B2 (4 points)

What is the final classifier produced by these three rounds of Adaboost?

H(x) = Sign($\frac{1}{2}\ln(7)$ (d) + $\frac{1}{2}\ln(\frac{11}{3})$ (j) + $\frac{1}{2}\ln(\frac{49}{17})$ (i))
.972 (d) + .6469 (j) + .529 (i)

B3 (3 points)

Does your classifier correctly classify all of the training data? Circle one:

YES **NO** (circled)

Such a pain to do
(any shortcuts

If you circled NO, list any training data points that your Adaboost classifier misclassifies:

#6 ← remember
Lverify .972(-1) + .6469(1) + .529(1)
= (+) when should be (−)

B4 (2 points)

Below is the description of Brittney's ideal character. According to your classifier, should her character be an elf?

	Elf	Fighting Style	Ears	Magic	Size
Brittney	??	Melee	Pointy	Yes	Medium

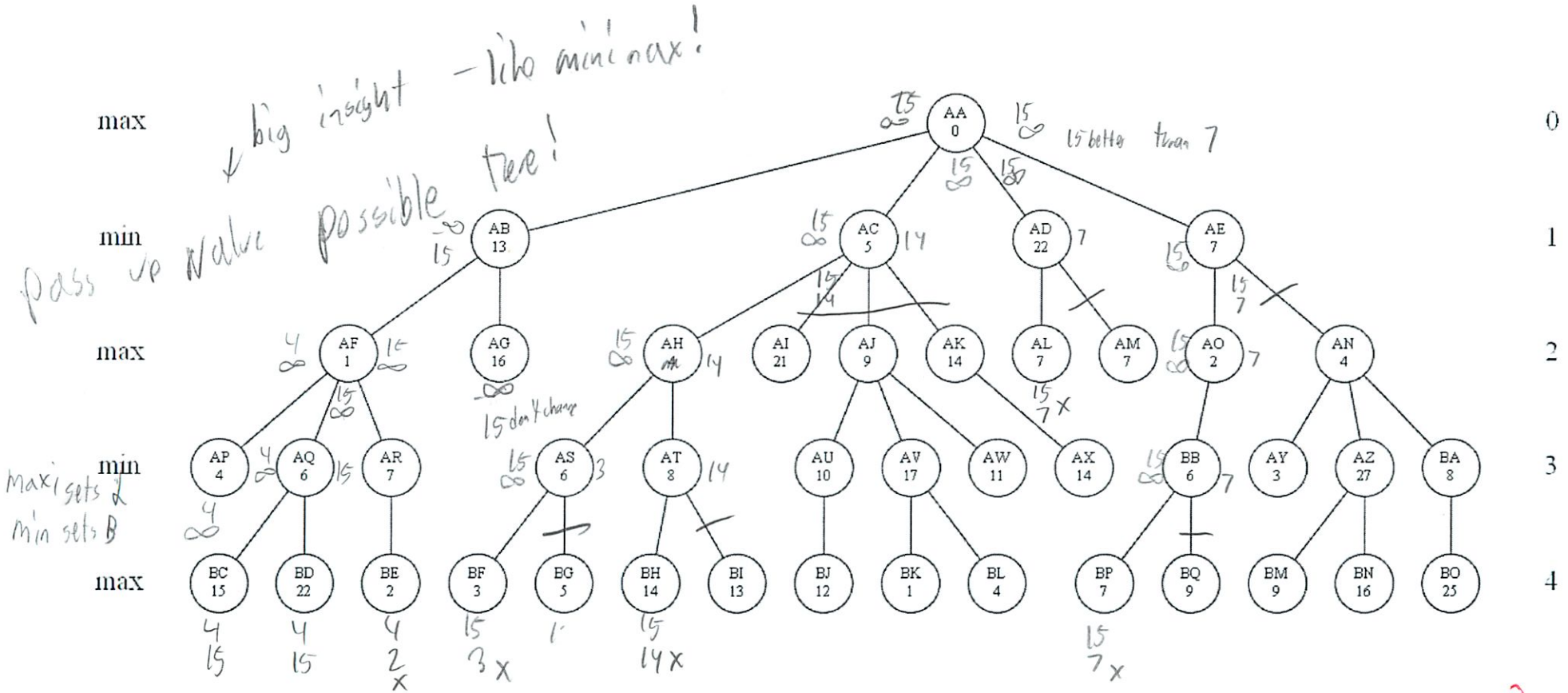
.972(1) + .6469(1) + .529(1)

Circle one:

ELF (circled) **NOT AN ELF**

Clearly

✓


big insight - like minimax!
pass up value possible tree!
max
min
max
min
max
0
1
2
3
4
AA 0
AB 13
AC 5
AD 22
AE 7
AF 1
AG 16
AH
AI 21
AJ 9
AK 14
AL 7
AM 7
AO 2
AN 4
AP 4
AQ 6
AR 7
AS 6
AT 8
AU 10
AV 17
AW 11
AX 14
BB 6
AY 3
AZ 27
BA 8
BC 15
BD 22
BE 2
BF 3
BG 5
BH 14
BI 13
BJ 12
BK 1
BL 4
BP 7
BQ 9
BM 9
BN 16
BO 25
15 better than 7
15 don't change
maxi sets α
mini sets β
α ≥ β cutoff
α ≤ β is fine!

6.034 Final Examination
December 15, 2008

Name	
EMail	

Quiz number	Maximum	Score	Grader
1	100		
2	100		
3	100		
4	100		
5	100		

There are 33 pages in this final, including this one. Additional pages of tear-off sheets are provided at the end with duplicate drawings and data. As always, open book, open notes, open just about everything.

Quiz 1, Question 1, Rules (50 points)

Before swimming in an unknown river, you want to figure out which animals are dangerous. You have a set of rules and assertions, given below.

Rules:

```
P0:   IF( '(?x) is a mammal',
          THEN( '(?x) is not a crocodile' ) )
P1:   IF( AND( '(?x) is not a crocodile',
               '(?x) lives underwater' ),
          THEN( '(?x) is a manatee'
P2:   IF( AND( '(?x) is a mammal',
               '(?x) lives underwater' ),
          THEN( '(?x) is a hippo' ) )
P3:   IF( OR( '(?x) is a crocodile',
              '(?x) is a hippo' ),
          THEN( '(?x) is dangerous' ) )
P4:   IF( '(?x) is not a crocodile',
          THEN( '(?x) is safe' ) )
```

Assertions:

```
A0:   ('Spike is a mammal')
A1:   ('Fido is a mammal')
A2:   ('Fido lives underwater')
A3:   ('Rover is a crocodile')
```

Part A: Forward Chaining (30 points)

You may make the following assumptions about forward chaining:

- Assume rule-ordering conflict resolution
- New assertions are added to the bottom of the dataset
- If a particular rule matches assertions in the dataset in more than one way, the matches are considered in the order corresponding to the top-to-bottom order of the matched assertions. Thus, if a particular rule has an antecedent that matches both A1 and A2, the match with A1 is considered first.

Run forward chaining on the rules and assertions provided. For the first two iterations, fill out the table below, noting the rules matched, fired, and new assertions added to the data set.

	Matched	Fired	New Assertions Added to Data Set
1			
2			

Which animals (Fido, Spike, Rover) are determined to be dangerous?

Which animals (Fido, Spike, Rover) are determined to be safe?

Would changing the order of the rules affect the final decision of which animals are dangerous or safe?

Part B: Backwards Chaining (20 points)

Make the following assumptions about backwards chaining:

- When working on a hypothesis, the backward chainer tries to find a matching assertion in the dataset. If no matching assertion is found, the backward chainer tries to find a rule with a matching consequent. In case none are found, then the backward chainer assumes the hypothesis is false.
- The backward chainer never alters the dataset, so it can derive the same result multiple times.
- Rules are tried in the order they appear.
- Antecedents are tried in the order they appear.

Evaluate the hypothesis 'Spike is dangerous' using backwards chaining. Draw a goal tree in the space below.

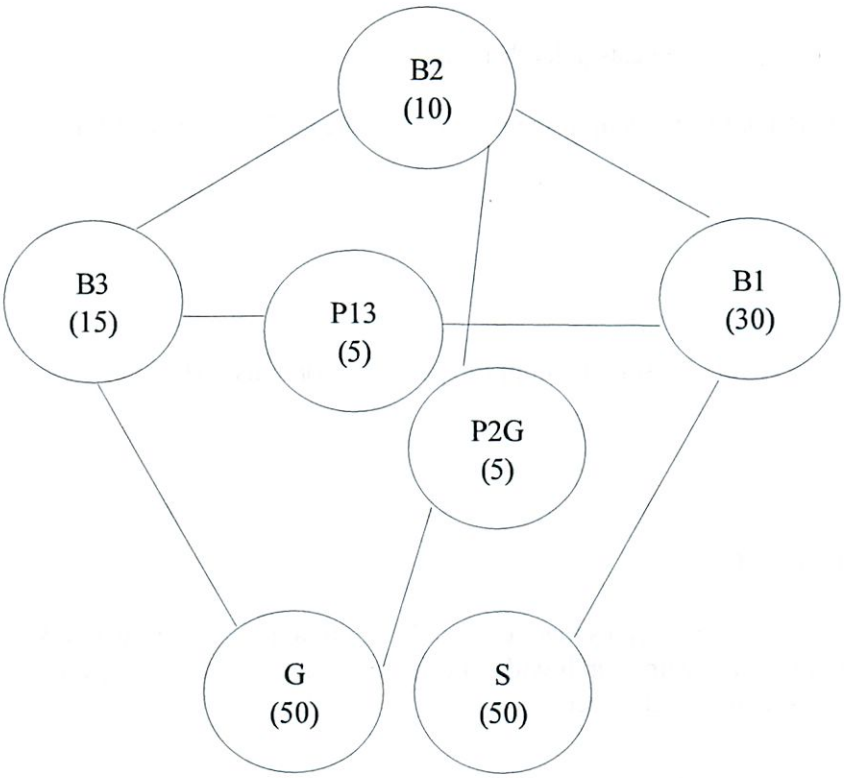
Is Spike dangerous?

Quiz 1, Question 2, Search (50 points)

Part A

Your 6.034 Tas have invented a new game based on baseball called Blurnsball. As many people know, baseball is incredibly boring, so to jazz it up, they included several rules variants, including a variant for running the bases. In Blurnsball, it is legal to run across the pitcher's mound and to the opposite side (thus making it legal to run from 1st base to the pitcher's mound to 3rd base or from 2nd base to the pitcher's mound to home). They have hired you as a consultant to use 6.034 Search techniques in order to analyze the new rules variant. See the graph below for a diagram of the new set-up.

Break all ties in lexicographic order, treating the path from start to finish as a character string. Thus, S-B1-B2-B3 comes before S-B1-P13-B3.



Part A1 (10 points)

The TAs are impatient and demand that you perform a hill-climbing search using the heuristic distances to the goal, provided in parentheses in the diagram.

What path do you find from the starting node S to the goal node G? Do not test a path to see if it reaches the goal until that path reaches the front of the search queue.

How many paths do you extend? Be sure to count the path that contains just S.

Part A2 (10 points)

Something seems funky, so Sam suggests a depth-first search.

What path do you find? Do not test a path to see if it reaches the goal until that path reaches the front of the search queue.

How many paths do you extend? Be sure to count the path that contains just S.

Part A3 (10 points)

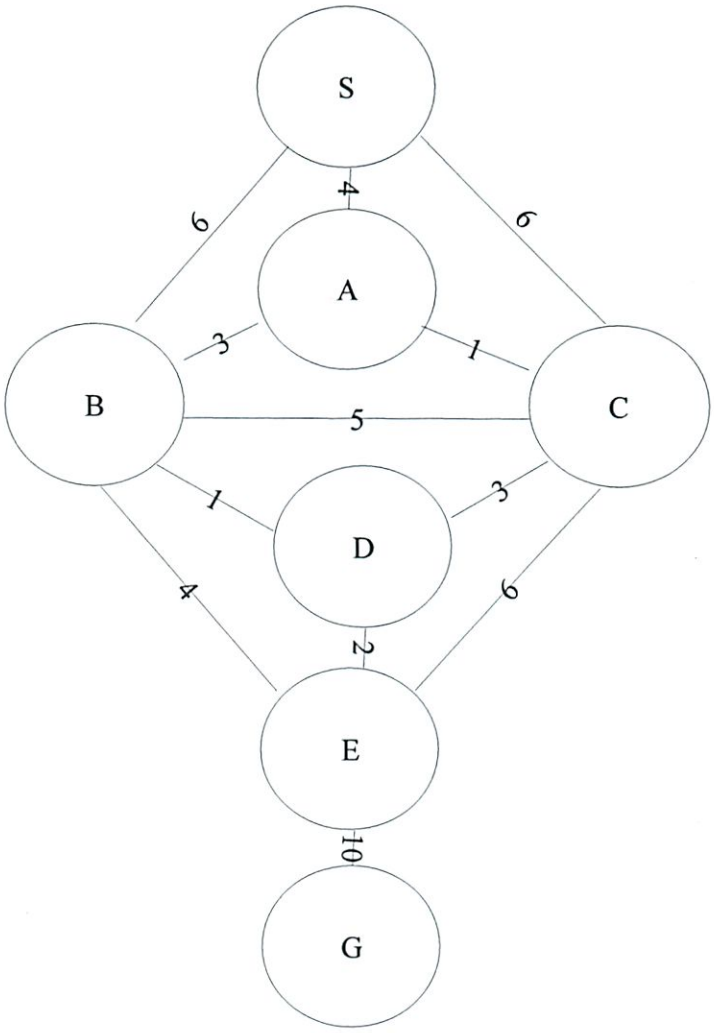
Mark is bored by your answer and decides that you should use Beam Search instead to keep the options from getting too large. Perform beam search with a beam width of 2. Sort after each extension and keep the two paths with the best heuristic distance.

What path do you find this time? Be sure to use the tie breaker on the previous page in case ties need breaking.

How many paths do you extend? Be sure to count the path that contains just S.

Part B

Alex has been chatting online for weeks with his new online girlfriend, Eliza. Stephanie and Maria are skeptical of this "Eliza" and insist that Alex meet with her in person, so when Alex asks Eliza out, Stephanie and Maria decide to come along to make sure the girlfriend is legit. They have to travel along the following streets, from S to G, and Alex wants to make sure he's on time to meet Eliza, so he insists that they take the shortest path.



Part B1 (10 points)

Using Branch and Bound with an Extended List, what is the final path from S to G? Be sure to show your goal tree for partial credit.

What nodes are in your extended list?

Part B2 (10 points)

Stephanie and Maria each suggest a heuristic:

Stephanie's heuristic:
S—25, A—16, B—3, C—15, D—12, E—0, G—0

Maria's heuristic:
S—4, A—16, B—16, C—15, D—14, E—10, G—0

Which of the following searches will find the same path you found using branch and bound with an extended list?

Circle those that work; cross out those that do not.

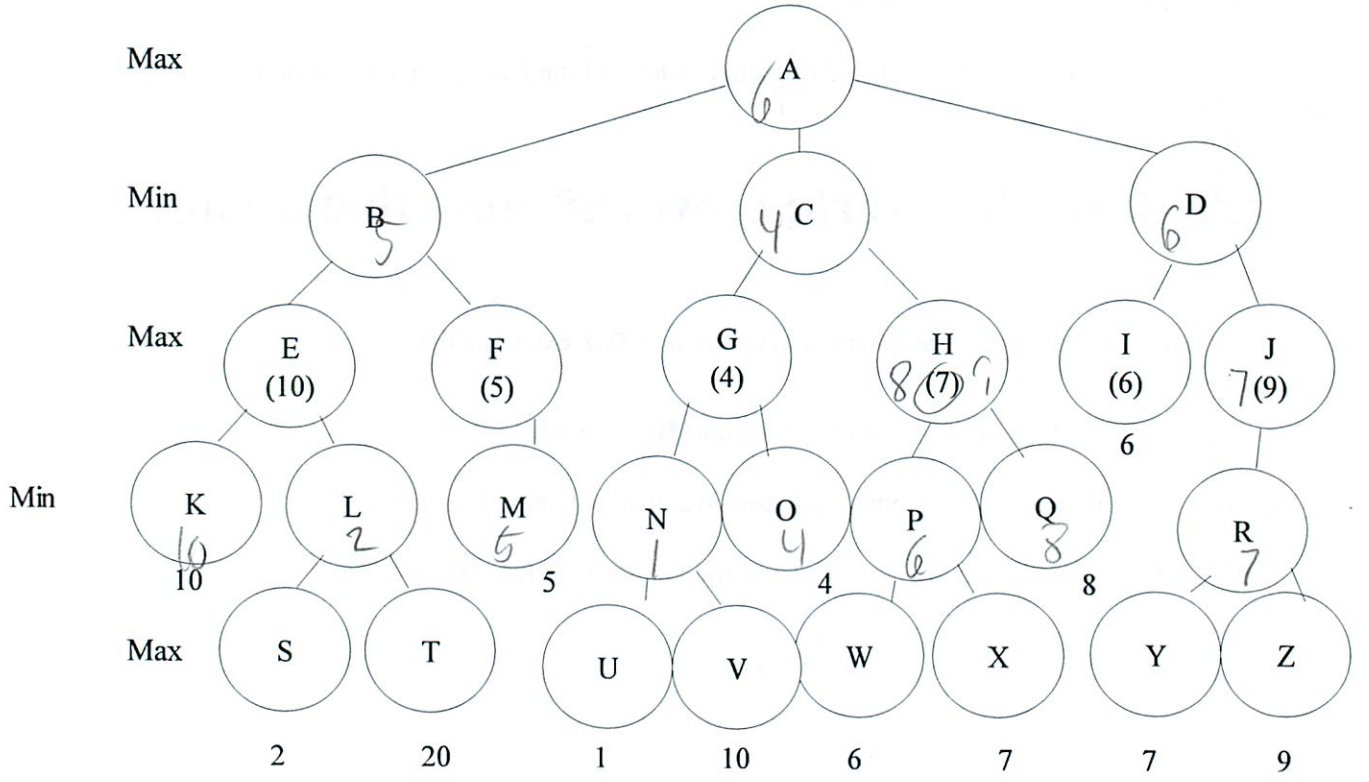
- Branch and bound with Stephanie's heuristic and **no** extended list.
- Branch and bound with Maria's heuristic and **no** extended list.
- Branch and bound with Stephanie's heuristic and an extended list (aka A*)
- Branch and bound with Maria's heuristic and an extended list (aka A*)

12/15

Quiz 2, Question 1, Games (50 points)

You are playing a new Sim game called Obamaquest, the Legend of the Lost International Credibility. In this game, you play a charismatic incoming president who must make a choice on various issues in order to save your country. After each of your turns, the outgoing president will attempt to perform the most meddlesome acts possible to make it less likely that you will succeed. You realize quickly that you can model this game using a simple Game Tree from 6.034, as shown below.

Static values are shown underneath leaf nodes. Ignore the numbers in parentheses for now.



had to read

Part A (15 points)

First, you decide to perform a simple minimax algorithm on the tree.

Which direction will the maximizer choose to go at node A?

D

What is the minimax value of node A?

6

Which static evaluations did you perform? (write the nodes you statically evaluated, in order).

k, S, t, M, U, V, W, X, Y, Z

Part B (25 points)

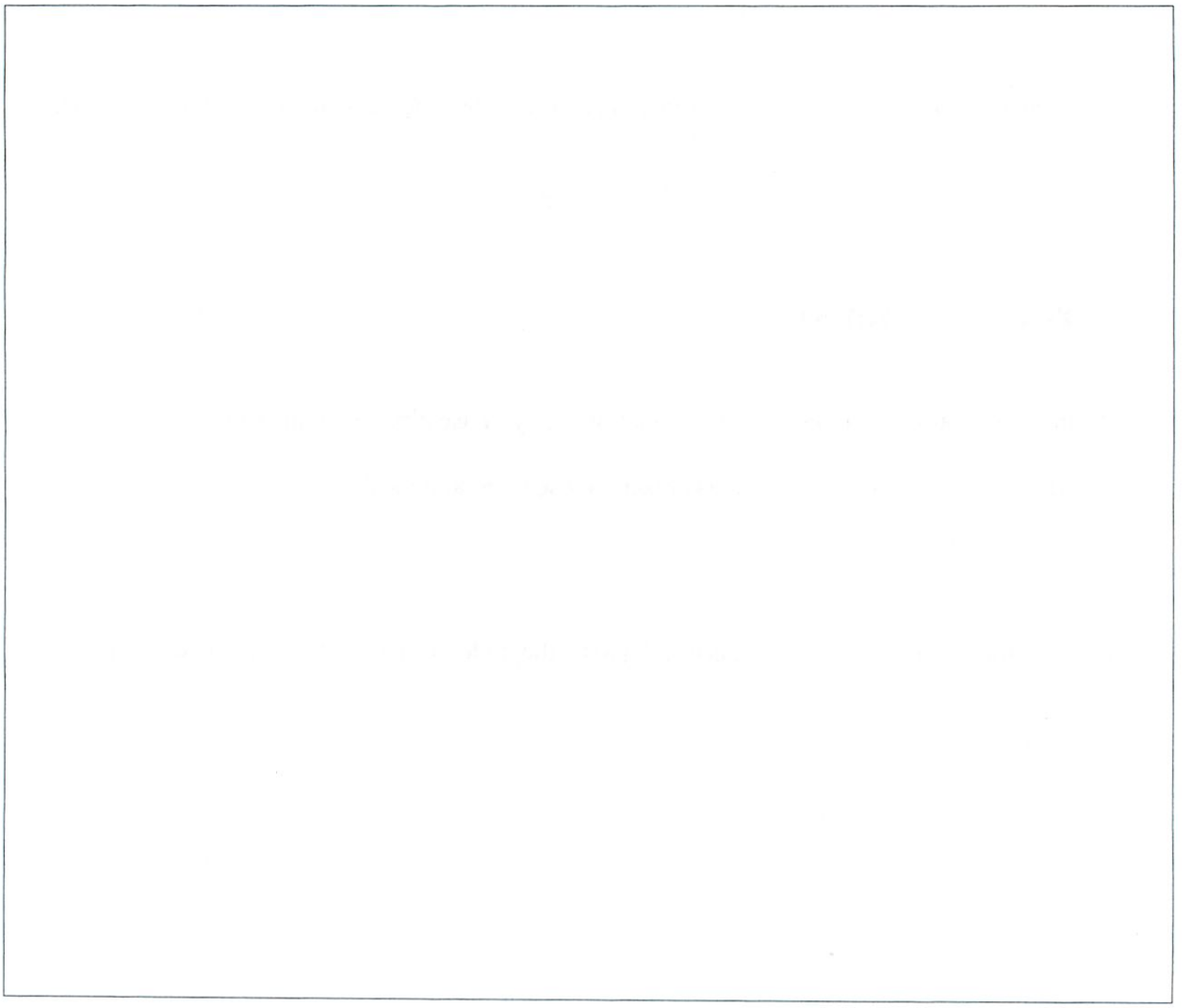
Minimax was taking too many static evaluations, so you use alpha-beta instead.

This time what direction will the maximizer choose to go at node A?

Which static evaluations did you perform? (write the nodes you statically evaluated, in order).

Part C (10 points)

The end result is still rather depressing in number of required static evaluations, so you decide to perform progressive deepening up to the second level and then reorder the tree to try for a more optimal pruning, using the static values for E,F,G,H,I, and J found in parentheses inside each circle. Draw your new ordering below. **You need not not draw any of the nodes below E, F, G, H, I, and J.**



Quiz 2, Question 2, Constraints (50 points)

Four 6.034 TAs (**Mark, Mike, Rob, Sam**) are trying to write eight questions for a final exam (somewhat like this one):

1. **C**onstraints
2. **O**ptimal Search
3. **G**ames
4. **R**ules
5. **I**D-Trees
6. **N**eural Nets
7. **S**VMs
8. **B**oosting

Some questions are harder to write than others, so they should be distributed about equally. And some problems are so similar to others that for variety, we don't want those written by the same TA. This leads to some constraints:

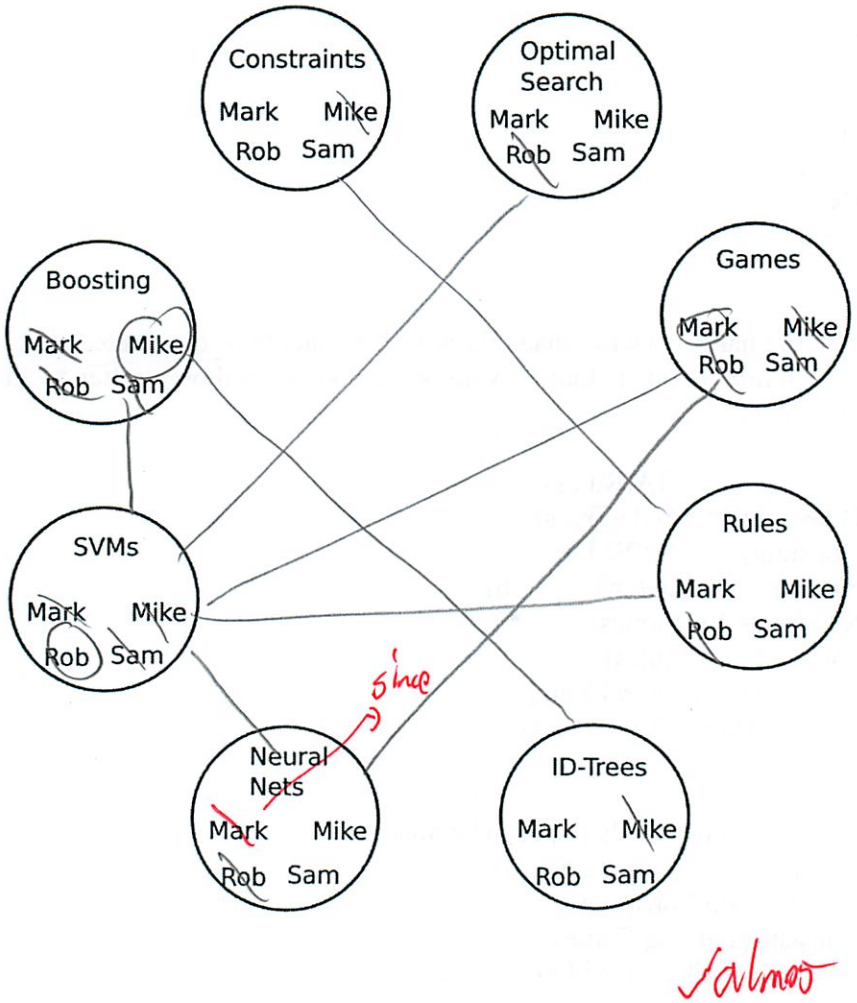
- TA(Constraints) != TA(Rules)
- TA(Boosting) != TA(ID-Trees)
- TA(Boosting) != TA(SVMs)
- TA(SVMs) != TA(Optimal Search)
- TA(SVMs) != TA(Games)
- TA(SVMs) != TA(Rules)
- TA(SVMs) != TA(Neural Nets)
- TA(Neural Nets) != TA(Games)

Also, there are some demands that have to be honored.

- Mike insists on doing Boosting (starting)
- Mike will not do Constraints
- Mark insists on doing Games
- Only Rob and Mike are willing to do SVMs.

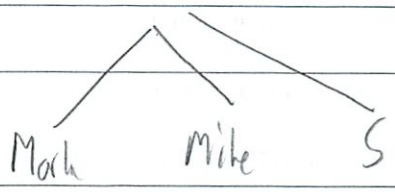
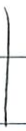
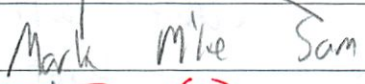
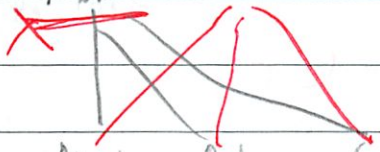
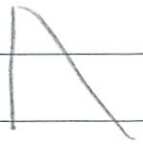
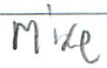
Part A (10 points)

Draw lines between variables with a != constraint, and use the TAs' demands to reduce domains by crossing out names. Continue to reduce domains using the constraints while possible.



Part B (20 points)

Starting with your domains reduced in Part A, find a solution using depth-first search only, using no constraint propagation , checking constraints at assignments only. Consider TAs in alphabetical order: Mark, Mike, Rob, Sam. Please feel free to abbreviate unambiguously. **Draw your search tree on the next page.**

Constraints	Mark Rob Sam
Optimal Search	
	
Games	
	
Rules	
	
ID-Trees	
	
Neural Nets	
SVMs	
Boosting	

only extend when yet

Almost

Cause Mark did constraints

it was an easy problem but I should have seen this

but do we extend here? last test under ✓ yes all for 1st and since all dep on mark

Part C (15 points)

You consider two more advanced constraint propagation algorithms that

- propagate choices to neighbors only
- propagate choices to neighbors and continue through any domains reduced to size one

Do these algorithms find the same result as the DFS? Which domains do these algorithms reduce during the course of their runs? Circle the answers.

Neighbors only

Same result as DFS? (YES) NO

Domains reduced: C O G R I N S B

Neighbors and any domain reduced to size one

Same result as DFS? (YES) NO

Domains reduced: C O G R I N S B

Part D (5 points)

One TA thinks the assignments made in this problem are unfair. Suggest a way to ensure the test questions are more evenly distributed across the four TAs.

that will climbin. randomly assigning next problem
or assigning least busy person

FC w/ prop Singleton

C

O

G

R

DD

NN

SVMs

Boost

M

|

Mark

Q

'wait already reduced domains won't do much unless switch D/V - but here 1 to many

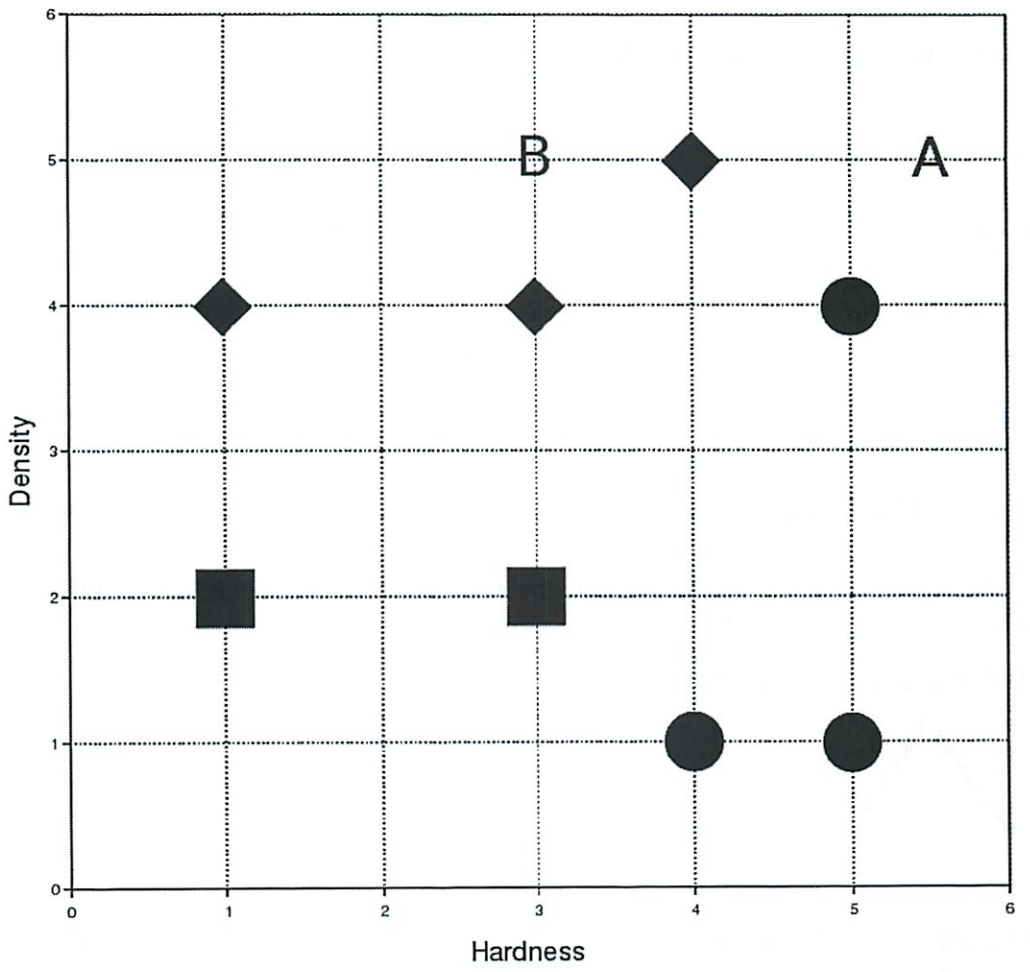
S

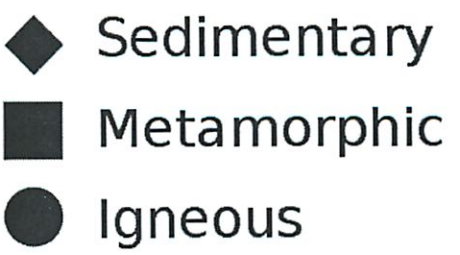
- assigning 1 does not eliminate the other!

Quiz 3, Question 1, NNand ID trees (50 points)

Part A: Nearest Neighbors

On the following graph, draw the decision boundaries produced by 1-Nearest Neighbor. Ignore the letters A and B.





How is Sample A classified by 1-NN?

By 3-NN?

How is Sample B classified by 1-NN?

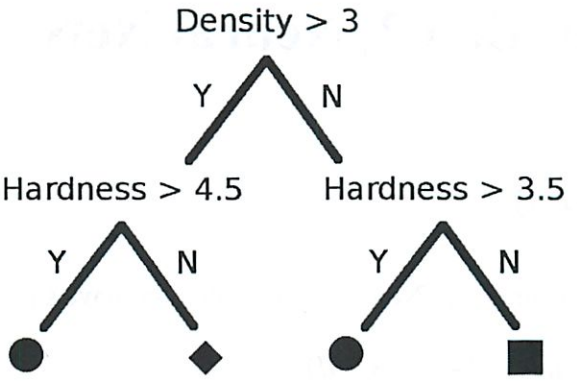
By 3-NN?

Part B: ID Trees

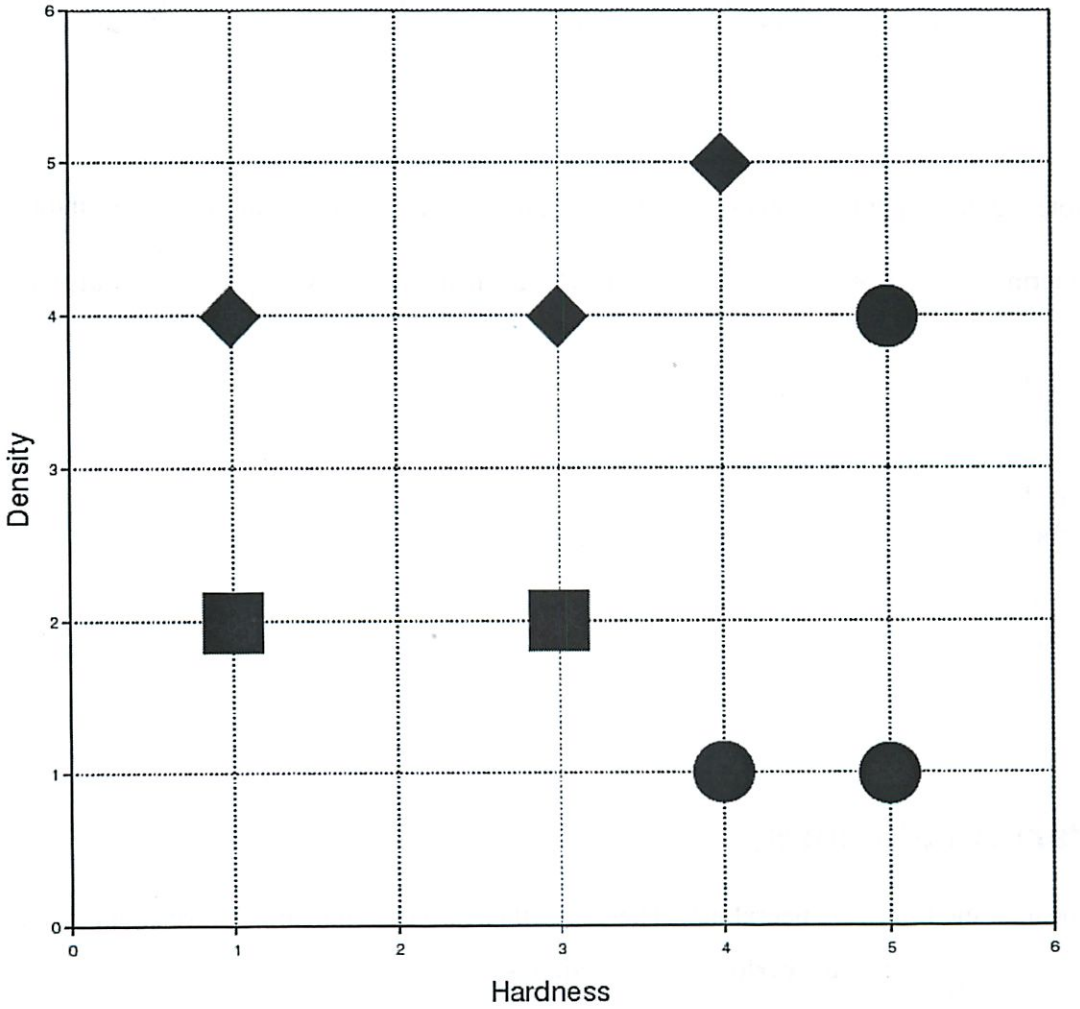
Using the same data as in Part A, calculate the disorder of the following ID Tree tests. Your answers may contain logarithms.

Density > 3  	**Disorder:**
Hardness > 3.5  	**Disorder:**

Suppose you created the following ID Tree.



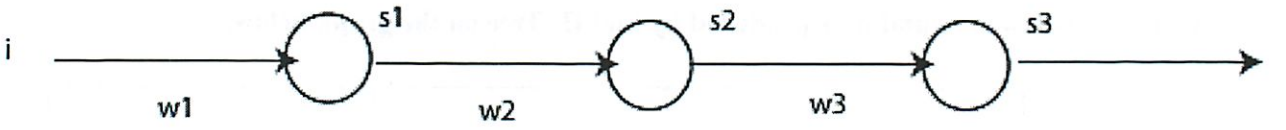
Draw the decision boundaries produced by that ID Tree on the graph below.



Quiz 3, Question 2, Neural Nets (50 points)

Part A (21 points)

Consider the following neural net. Note that all the neurons have sigmoid units, $s(z)=\frac{1}{1+e^{-z}}$ and the performance function is $P=-\frac{1}{2}(s_3-d)^2$



Note that the input to the net is i and the outputs of the sigmoid units are $s_1, s_2,$ and s_3.

In terms of $i, w_1, w_2, w_3, i, s_1, s_2, s_3,$ and d calculate the following partial derivatives:

$\frac{\partial P}{\partial w_3}=$

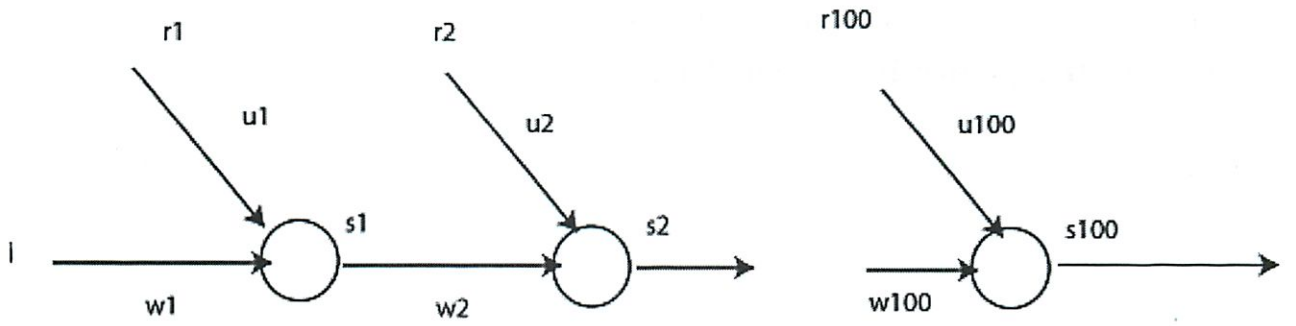
$\frac{\partial P}{\partial w_2}=$

$\frac{\partial P}{\partial w_1}=$

Part B (24 points)

Consider the following neural net. There are 100 neurons. All neurons have sigmoid units, $s(z)=\frac{1}{1+e^{-z}}$ and the performance function is $P=-\frac{1}{2}(s_{100}-d)^2$

The inputs to the network are i and r_i .The outputs of the sigmoid units in the network are s_i



At a certain time, t, $i=\gamma$ and

$r_1=r_2=\cdots=r_{100}=\rho$

$s_1=s_2=\cdots=s_{100}=\sigma$

$u_1=u_2=\cdots=u_{100}=1$

$w_1=w_2=\cdots=w_{100}=\omega$

Calculate $\frac{\partial P}{\partial w_1}$ in terms of γ ρ σ ω and d .

$\frac{\partial P}{\partial w_1}=$

Part C (5 points)

Finally, exhibit an equation that relates σ , ρ and ω .

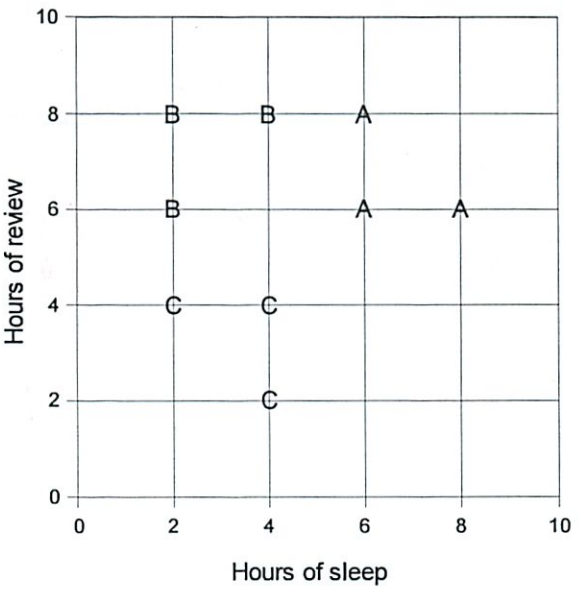
Do not attempt to solve the equation.

$\sigma =$

Quiz 4, Question 1, Support Vectors (50 points)

Miriam has a lot of finals. She wants to know how best to use her time, so naturally she collects some data to train a support vector machine.

She finds that nine of her friends have already taken the class that her next final is in, so she asks them how long they spent reviewing for the class's final exam, how many hours of sleep they got the night before the final, and what grade they got (A, B, or C). The results appear on this graph:



3 types

The problem she runs into is that SVMs can typically only distinguish two classes, and her data contains three classes of grades. However, she finds a way to make a three-class SVM:

- Divide the problem among three ordinary, two-class SVM classifiers, called $h_A(\mathbf{x})$, $h_B(\mathbf{x})$, and $h_C(\mathbf{x})$.

Sounds annoying

- Each SVM treats one of the classes (A, B, and C respectively) as +, and treats the other two classes as –.

- The final classifier outputs A, B, or C based on which of the three sub-classifiers outputs the highest value. For example, if h_A outputs -5, h_B outputs -3, and h_C outputs -2, the overall result should be C. If h_A outputs 2, h_B outputs -1, and h_C outputs -3, the overall result should be A.

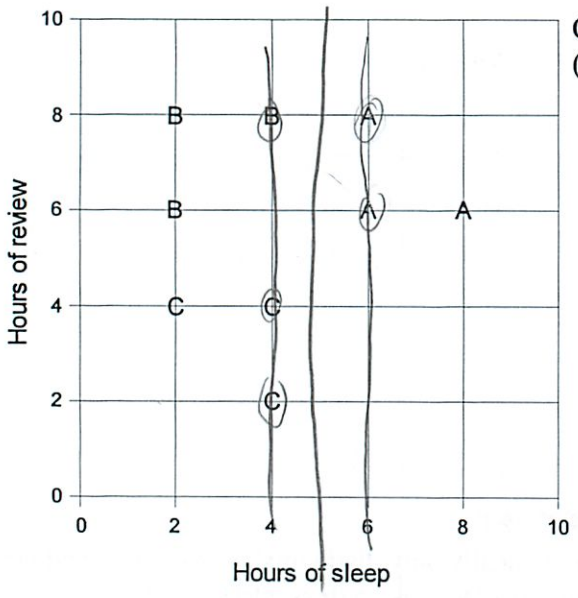
Not abs value

Part A (30 points)

The three SVM classifiers are linear, so their output will be defined by the equation $h = \mathbf{w} \cdot \mathbf{x} + b$. On the next three graphs:

- Draw the "street" that separates the data by drawing a dotted line at the $h = 0$ boundary, and solid lines at $h = 1$ and $h = -1$.
- Write the values of $\mathbf{w}$ and b for that classifier.

hmm how to write eqn



Classifier A: $h_A = \mathbf{w}_A \cdot \mathbf{x} + b_A$
(This classifier separates A's from other grades.)

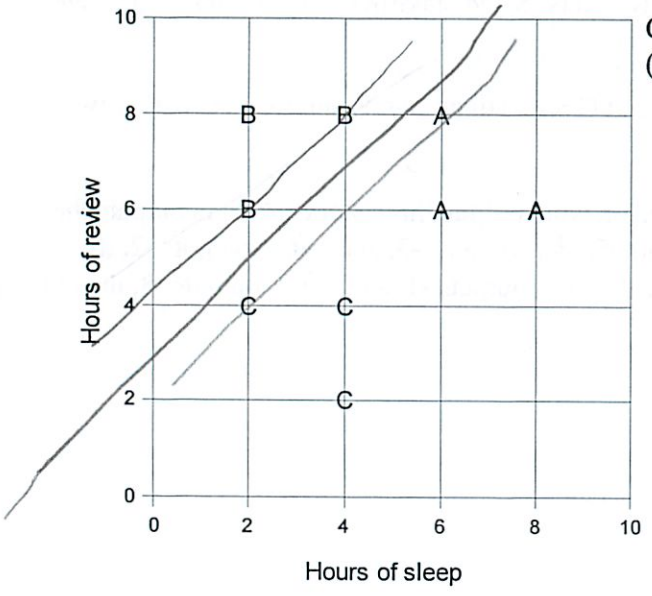
AG B,C⊖

$\mathbf{w}_A =$ $\begin{bmatrix} 5 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 2 \\ 0 \end{bmatrix}$ ← $x - 5 \leq 0$ is ⊖

$b_A =$ $\begin{bmatrix} -5c \end{bmatrix}$ $\begin{bmatrix} -10 \end{bmatrix}$ −5 they have $c = 1$

$\frac{2}{\sqrt{c^2}} = 1$

$2 = c$



Classifier B: $h_B = \mathbf{w}_B \cdot \mathbf{x} + b_B$
(This classifier separates B's from other grades.)

BG A,C⊖

$\mathbf{w}_B =$ $x + 3$ ← y is involved, duh!

$b_B =$ $\begin{bmatrix} c \\ 0 \end{bmatrix}$ $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ $\begin{bmatrix} \sqrt{2} \\ 0 \end{bmatrix}$

$\begin{bmatrix} 3c \end{bmatrix}$ −3 $3\sqrt{2}$

↑ notice what is +/−

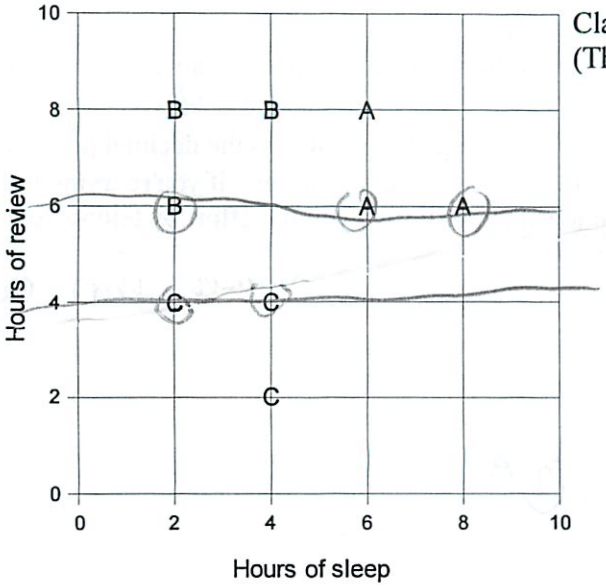
$\frac{2}{\sqrt{c^2 + c^2}} = 1$

$2 = \sqrt{2c^2}$

$4 = 2c^2$

$2 = c^2$

$c = \sqrt{2}$



Classifier C: $h_C = \mathbf{w}_C \cdot \mathbf{x} + b_C$
(This classifier separates C's from other grades.)

$\mathbf{w}_C =$

$b_C =$

Part B (6 points)

After running the three SVMs, Miriam's computer needs to determine which class wins overall. Remember that this is determined by which classifier outputs the *highest* value.

To determine whether classifier A beats classifier B, one can subtract their equations, giving a new equation in terms of the vector $\mathbf{x}$. Calculate these equations for all pairs of classifiers – this will help to analyze which classifier wins at each point.

$h_A(\mathbf{x}) - h_B(\mathbf{x}) =$

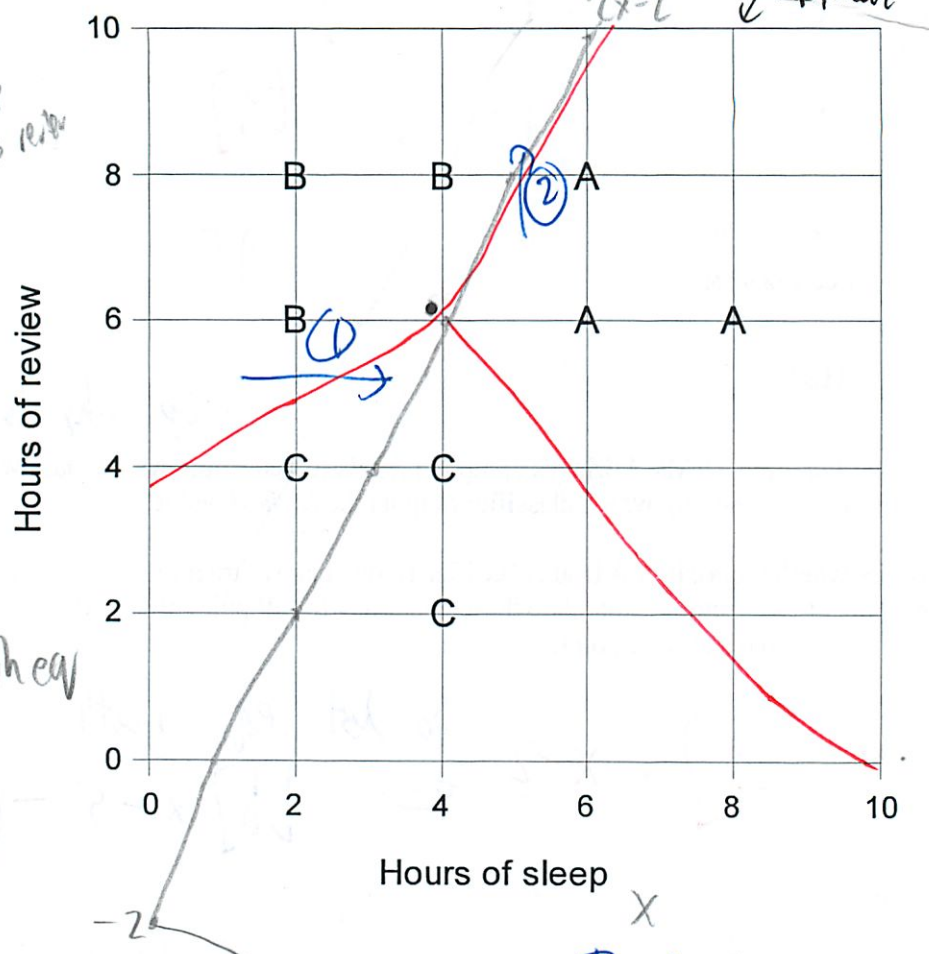
$h_B(\mathbf{x}) - h_C(\mathbf{x}) =$

$h_A(\mathbf{x}) - h_C(\mathbf{x}) =$

Part C (8 points)

Draw the boundary lines that separate classes A, B, and C in the final classifier on this graph.

The intent here is ***not*** for you to have to solve the equations down to the decimal point. We're looking for a graph that graphically shows the correct overall behavior. In fact, if you're trying to find the point where all three classifiers output the same value, we'll save you the effort by telling you that it should be at 3.8 hours of sleep and 6.2 hours of review.



Part D (6 points)

Describe a way in which the behavior of your classifier is counterintuitive, given the meaning of the input data?

↑ #s of sleep the moves from B → C in some areas

↑ # hrs of review A → B

Clarity

14/16

So why was $c=1$ there?

Recitation notes

width of road

$$m = \frac{2}{\|\vec{w}\|}$$

So ~~remain~~ remember w are from the previas one

which is what I did...

~~$\frac{2}{\|\vec{w}\|} = 1$~~

$$\frac{2}{\sqrt{c^2}} = 1$$

$$\frac{2}{c} = 1$$

$$2 = 1c$$

~~So need to multiply~~

Any multiple of c is still same decision boundry

$$(x - c_y - 2c \geq 0$$

in example problem

I had street width (all the way across - right?)

Yeah ⊖ gutter + ⊖ gutter

(2)

So should be able to test

$$\frac{2}{\sqrt{1^2+(0)^2}} = 1$$

$$\frac{2}{1} = 1 \quad ??$$

Now must solve for c

Try for other one

$$\frac{2}{\sqrt{(-1)^2+(1)^2}} = 1$$

$$\frac{2}{\sqrt{2}} = 1 \quad ??$$

Ohhhh one tick = 2 units on graph

So $\frac{2}{\sqrt{c^2}} = 2$

#1,3 $2 = 2c$

$c = 1$

#2 $\frac{2}{\sqrt{c^2+c^2}} = 2$

$2 = 2\sqrt{2c^2}$

$4 = 4 \cdot 2c^2$

$1 = 2c^2$

$\frac{1}{2} = c^2$

$\frac{1}{4} = c$

Still seems off

(3)

Yeah what is going on

road width is clearly 2

$$\frac{2}{\sqrt{c^2+c^2}} = 2$$

$$2 = 2\sqrt{2c^2}$$

$$4 = 4 \cdot 2c^2$$

$$1 = 2c^2$$

$$\frac{1}{2} = c^2$$

actually $\sqrt{}$ is

$$\frac{\sqrt{2}}{2}$$

So should be $\begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$

$$-3\frac{\sqrt{2}}{2}$$

Test

$$\frac{2}{\sqrt{\left(\frac{\sqrt{2}}{2}\right)^2 \cdot 2}} = 2$$

Works

(4)

Oh duh was not measuring road straight across

Diagonal of square?

$\sqrt{2^2+2^2} = \sqrt{8} = 2\sqrt{2}$

half of that

$\sqrt{2}$

$$\frac{2}{\sqrt{c^2+c^2}} = \sqrt{2}$$

$$2 = \sqrt{2}\sqrt{2c^2}$$

$$4 = 2 \cdot 2c^2$$

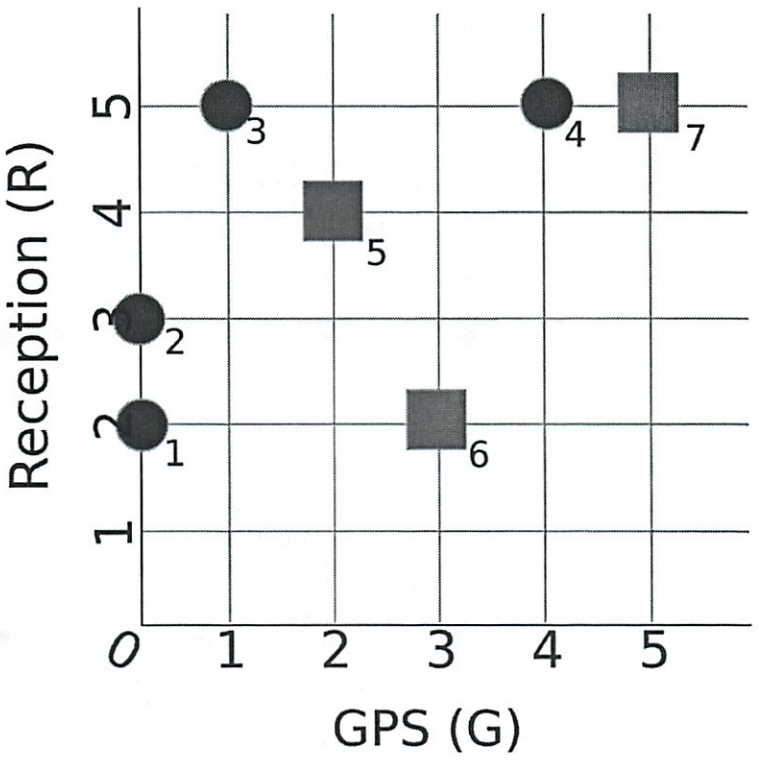
$$1 = c^2$$

$$c = 1$$

Grrr annoying - if ya forgot c thing, problem would still be right

Quiz 4, Question 2, Boosting(50 points)

You want to use your cell phone to detect whether you are indoors or outdoors. You use the number of bars of reception (R) and your GPS signal quality (G) as features. Here's the data you gather:



Part A (10 points)

You want to choose a learning algorithm for this data. Circle all the algorithms that can learn this data with zero training error:

AdaBoost ID-Tree Linear SVM RBF SVM

RBF = Radial basis function, the exponential kernel; you can choose the RBF's σ.

Part B (10 points)

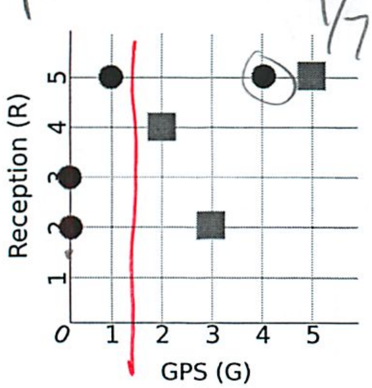
You decide to train AdaBoost using these stumps:

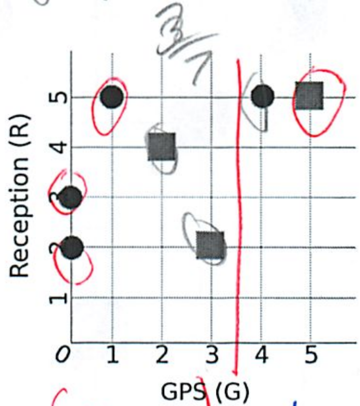
- $h_1(G,R) = \text{sign}(1.5 - G)$
- $h_2(G,R) = \text{sign}(G - 3.5)$
- $h_3(G,R) = \text{sign}(4.5 - G)$
- $h_4(G,R) = \text{sign}(R - 2.5)$
- $h_5(G,R) = \text{sign}(R - 4.5)$

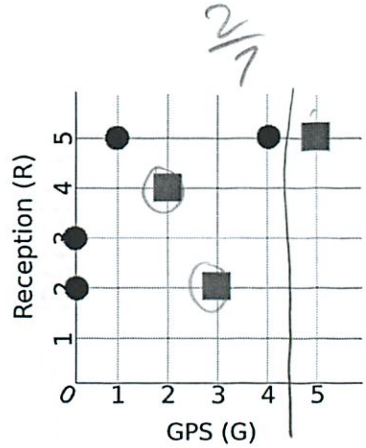
what? oh line eqns !!

For each stump, draw the line and circle the misclassified data points.

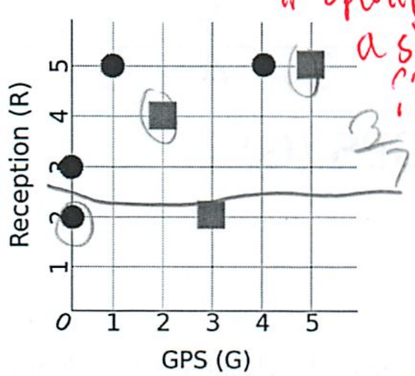
So pick one w/ lowest. oh opps

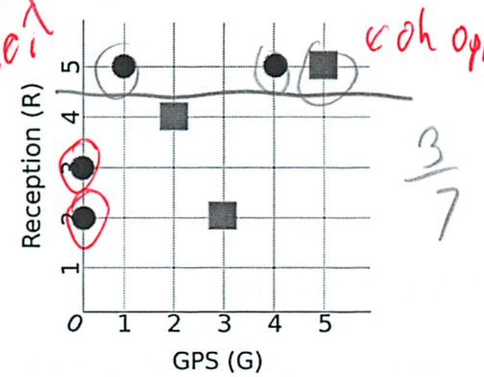






so it specifies a side!! wacky





Part C (30 points)

Complicated

Starting with all weights set to 1/7, run AdaBoost for three rounds. For each round, give the classifier you chose and its error. Also give the final AdaBoost classifier and its error rate. (You don't need to give weights or alphas, but they'll help us give you partial credit.) Break ties by preferring low-numbered stumps.

$h_a(G,R) =$ #1

$\varepsilon_a =$ $\frac{1}{7}$ ✓

Correct $\frac{7}{12}w \to \frac{1}{12}$ inc $\frac{1}{2}$

$h_b(G,R) =$ #3 ✓

$\varepsilon_b =$ $\frac{2}{12} = 1/6$ ✓

Correct $\frac{3}{5}w_i$ 1/20 3/10 ✓ inc $3w_i$ $\frac{1}{4}$ ✓

$\frac{1}{20}$ $\frac{1}{20}$ $\frac{1}{20}$	$\frac{1}{4}$ $\frac{1}{4}$	3/10 $\frac{1}{20}$

$h_c(G,R) =$ #5 ✓

$\varepsilon_c =$ $\frac{1}{20} + \frac{1}{20} + \frac{3}{10} = 4/10$ $\frac{3}{20}$ (had diff pts circled)

~~Correct $\frac{5}{7}w_i$~~ don't need

Signex

H(G,R)= $\frac{1}{2}\ln\left(\frac{6/7}{1/7}\right)$ #1 + $1/2 \ln\left(\frac{5/6}{1/6}\right)$ # 3 + $\frac{1}{2}\ln\left(\frac{7/10}{3/10}\right)$ #5

why no weights?

ε= 0

? how do we have an error rate? could add it doing it right but complicated...

how did they get that?

that was wacky

Quiz 5, Problem 1, (25 points)

Tired of searching for a new president every 10-15 years, the MIT corporation decides to elect a permanent, hereditary monarch. Times are tough, so they are overwhelmed with applicants and decide to hire you to create a kind of filter to reduce the candidate set to a manageable number.

Your first task is to pick some descriptors. You decide to look at intelligence, field, group memberships (if any) , and gender. Men an women form a set. Fields are grouped as at MIT: EE and MechE are part of engineering, Physics and Math are part of science, and XV is part of management. Engineering, science, management, architecture, and humanities are "MIT-type fields."

Using Arch learning, indicate in the tabel what is learned from each example and identify the heuristic involved by name, if known. If nothing is learn, put an x in the corresponding 2 columns.

Part A (18 points)

Candidate	Intelligent	Field	Member of	Gender	Heuristic	What is learned
Good	Yes	EE		Woman		
Bad	No	EE		Woman		
Bad	Yes	EE	Anarchists	Woman		
Good	Yes	MechE		Woman		
Good	Yes	MechE		Man		
Good	Yes	Physics		Man		
Good	Yes	Math		Man		
Bad	No	XV		Man		
Bad	Yes	XV	Anarchists	Woman		

Part B (7 points)

Exhibit an example, which if used as the second example, would teach two characteristics at once.

Candidate	Intelligent	Field	Member of	Gender	What is learned

Quiz 5, Problem 2, (25 points)

Circle the best answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

Experiments with the newly sighted children indicate that they have difficulty with

1. Naming colors
2. Recognizing objects that are moving
3. Recognizing objects while the child is moving
4. Recognizing overlapping objects
5. None of the above

A newly sighted child, presented with a picture of a ball and a picture of a cube will

1. Readily identify which is which
2. Think they are both the same
3. Think they are both balls
4. Not know which is which
5. None of the above

A newly sighted child will have trouble recognizing a triangle if

1. The triangle is seen against a background of short black lines
2. It is a filled with black, not just three black lines
3. Only three black lines are shown, with no fill
4. The triangle is seen between a circle and a square
5. None of the above

Experiments with newly sighted children suggest the key to visual bootstrapping is

1. Cooperation between vision and sound systems
2. Use of color in figure-ground separation
3. Restricted motion of Infants during the first few months of life
4. Clues provided by the visual motion detection system
5. None of the above

Human children...

1. Like cats, cannot have vision usefully restored if sightless beyond a critical period
2. Like cats, have trouble tracking moving objects if sightless beyond a criical period
3. Prefer to remain blind if sightless beyond a critical period
4. Cannot identfy geometric figures on a computer screen
5. None of the above

Quiz 5, Problem 3, (25 points)

Circle the best answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

Disorientation experiments with rats, children, and adults demonstrate

1. Rats will not move in the presence of loud rock music
2. Children that behave like rats do not spontaneously use words like left and right
3. Rats and infants do not have full-spectrum color vision
4. Neither small children nor adults cannot reorient themselves if severely sleep derprived
5. None of the above

Chomsky believes the great differentiator between modern humans and humans 100,000 years ago is

1. Improved mitochondria
2. Improved memory
3. Improved hand-eye coordination
4. Ability to plan
5. None of the above

Brooks subsumption architecture is best described as

1. A demonstration of the importance of reasoning and planning in robot activity
2. An approach demonstrated by a robot that collected beer cans on cape-cod beaches
3. An approach to building robots focused on tightly coupled loops between vision and language
4. An idea reminiscent of the notion of abstraction layers in programming methodology
5. None of the above

Winston believes that to develop an account of human intelligence we must

1. Find new ways to exploit the SOAR architectural paradigm
2. Build systems that acquire permanent knowledge from visual experiences
3. Collect commonsense knowledge using volunteers contributing via the web
4. Focus for the time being on what insects can do
5. None of the above

The General Problem Solver Architecture

1. Is a general purpose architecture that unifies reasoning and perception
2. Is a special purpose architecture limited to proving theorems in first order logic
3. Uses a means-ends approach to operator selection
4. Gets stuck if a selected operator cannot be applied in the current state
5. None of the above

Quiz 5, Problem 4, (25 points)

Coen's method

1. Uses boosting to make clusters
2. Uses a neural net to make clusters
3. Uses subsumption to make clusters
4. Uses cross modal coupling to make clusters
5. None of the above

Coen's method

1. Explains why Zebra Finches use a mating song
2. Enables a program to learn to sing a Zebra Finch's mating song
3. Enables a program to learn to to recognize a Zebra Finch's mating song
4. Enables a program to recognize the mating song of particular Zebra Finches
5. None of the above

In Coen's method

- Two clusters are close if they are close in a Euclidian sense
- Two clusters are not close if they are not close in a Euclidian sense
- Two clusters are close if they project proportionately to clusters in another space
- Two clusters are not close if they have different shapes
- None of the above

Coen's method

1. Works well with any distance metric
2. Works well only with the Wasserstein distance metric
3. Works well only when the number of clusters is known in advance
4. Works well only when there are two or three clusters
5. None of the above

Coen's method

1. Works well only if corresponding points in two spaces are provided
2. Works well only if all the points in at lest one cluster are close together
3. Works well only if there is no noise
4. Works well only with labeled data
5. None of the above

This page left almost blank intensionally

6.034 Final Examination
December 15, 2008

Name	
EMail	

Quiz number	Maximum	Score	Grader
1	100		
2	100		
3	100		
4	100		
5	100		

There are 33 pages in this final, including this one. Additional pages of tear-off sheets are provided at the end with duplicate drawings and data. As always, open book, open notes, open just about everything.

Quiz 1, Question 1, Rules (50 points)

Before swimming in an unknown river, you want to figure out which animals are dangerous. You have a set of rules and assertions, given below.

Rules:

```
P0:  IF( '(?x) is a mammal',
         THEN( '(?x) is not a crocodile' ) )
P1:  IF( AND( '(?x) is not a crocodile',
              '(?x) lives underwater' ),
         THEN( '(?x) is a manatee'
P2:  IF( AND( '(?x) is a mammal',
              '(?x) lives underwater' ),
         THEN( '(?x) is a hippo' ) )
P3:  IF( OR( '(?x) is a crocodile',
             '(?x) is a hippo' ),
         THEN( '(?x) is dangerous' ) )
P4:  IF( '(?x) is not a crocodile',
         THEN( '(?x) is safe' ) )
```

Assertions:

```
A0:  ('Spike is a mammal')
A1:  ('Fido is a mammal')
A2:  ('Fido lives underwater')
A3:  ('Rover is a crocodile')
```

A4: Spike is not a crocodile (P0)
A5: Fido is not a crocodile (P0)
A6: Fido is a manatee (P1)
A7. Fido is a hippo (P2)
A8. Rover is dangerous (P3)
A9. Fido is dangerous (P3)
A10. Spike is safe (P4)
A11 Fido is safe (P4)

Part A: Forward Chaining (30 points)

You may make the following assumptions about forward chaining:

- Assume rule-ordering conflict resolution
- New assertions are added to the bottom of the dataset
- If a particular rule matches assertions in the dataset in more than one way, the matches are considered in the order corresponding to the top-to-bottom order of the matched assertions. Thus, if a particular rule has an antecedent that matches both A1 and A2, the match with A1 is considered first.

Run forward chaining on the rules and assertions provided. For the first two iterations, fill out the table below, noting the rules matched, fired, and new assertions added to the data set.

	Matched	Fired	New Assertions Added to Data Set
1	P0, P2, P3,	P0	Spike is not a crocodile
2	P0, P2, P3, P4	P0	Fido is not a crocodile

Which animals (Fido, Spike, Rover) are determined to be dangerous?

Rover, Fido

Which animals (Fido, Spike, Rover) are determined to be safe?

Spike, Fido

Would changing the order of the rules affect the final decision of which animals are dangerous or safe?

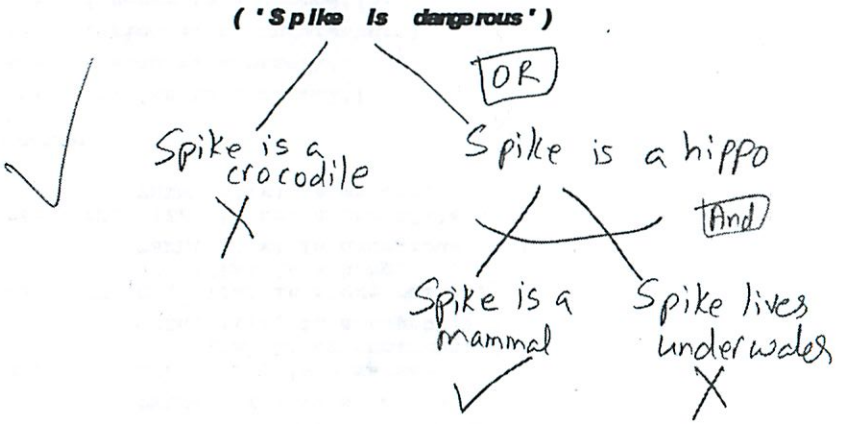
No

Part B: Backwards Chaining (20 points)

Make the following assumptions about backwards chaining:

- When working on a hypothesis, the backward chainer tries to find a matching assertion in the dataset. If no matching assertion is found, the backward chainer tries to find a rule with a matching consequent. In case none are found, then the backward chainer assumes the hypothesis is false.
- The backward chainer never alters the dataset, so it can derive the same result multiple times.
- Rules are tried in the order they appear.
- Antecedents are tried in the order they appear.

Evaluate the hypothesis 'Spike is dangerous' using backwards chaining. Draw a goal tree in the space below.



Is Spike dangerous?

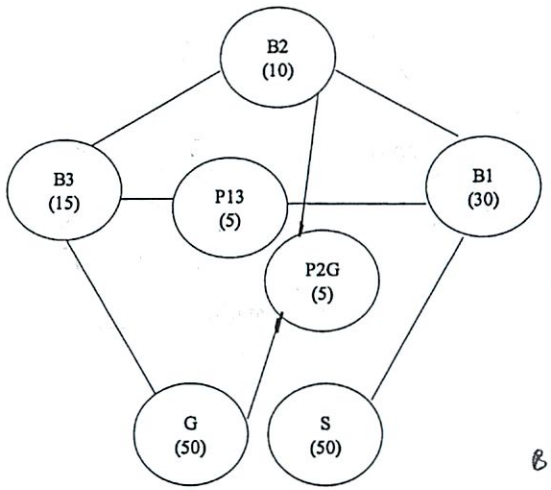
No

Quiz 1, Question 2, Search (50 points)

Part A

Your 6.034 Tas have invented a new game based on baseball called Blurnsball. As many people know, baseball is incredibly boring, so to jazz it up, they included several rules variants, including a variant for running the bases. In Blurnsball, it is legal to run across the pitcher's mound and to the opposite side (thus making it legal to run from 1st base to the pitcher's mound to 3rd base or from 2nd base to the pitcher's mound to home). They have hired you as a consultant to use 6.034 Search techniques in order to analyze the new rules variant. See the graph below for a diagram of the new set-up.

Break all ties in lexicographic order, treating the path from start to finish as a character string. Thus, S-B1-B2-B3 comes before S-B1-P13-B3.



Part A1 (10 points)

The TAs are impatient and demand that you perform a hill-climbing search using the heuristic distances to the goal, provided in parentheses in the diagram.

Hill climbing

S

B1 (30)

B2(10) P13 (5)

B3 (15)

B2(10) G(50)

P2G(5)

G(50)

What path do you find from the starting node S to the goal node G? Do not test a path to see if it reaches the goal until that path reaches the front of the search queue.

S-B1-P13-B3-B2-P2G-G ✓

How many paths do you extend? Be sure to count the path that contains just S.

✓ 6 (not including G)

DFS:

S

B1

B2

B3

G

Part A2 (10 points)

Something seems funky, so Sam suggests a depth-first search.

What path do you find? Do not test a path to see if it reaches the goal until that path reaches the front of the search queue.

S-B1-B2-B3-G

How many paths do you extend? Be sure to count the path that contains just S.

4 (not including G)

Part A3 (10 points)

Mark is bored by your answer and decides that you should use Beam Search instead to keep the options from getting too large. Perform beam search with a beam width of 2. Sort after each extension and keep the two paths with the best heuristic distance.

What path do you find this time? Be sure to use the tie breaker on the previous page in case ties need breaking.

S-B1-B2-P2G-G

How many paths do you extend? Be sure to count the path that contains just S.

6 (not including G)

S

B1

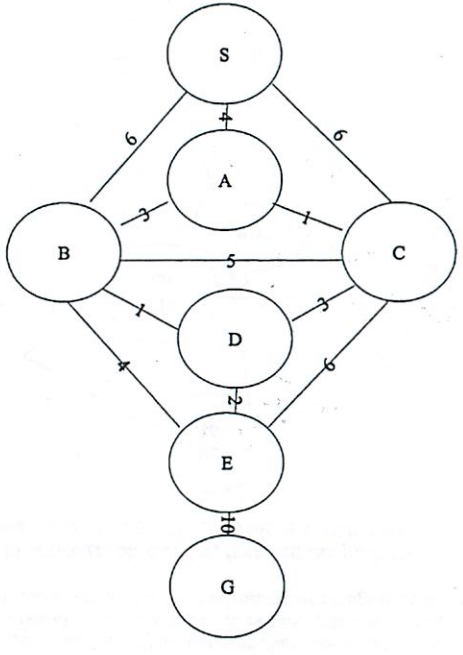
B2(10) P13(5)

B3(15) P2G(5) B3(15)

G(50) B2(10) G(50)

Part B

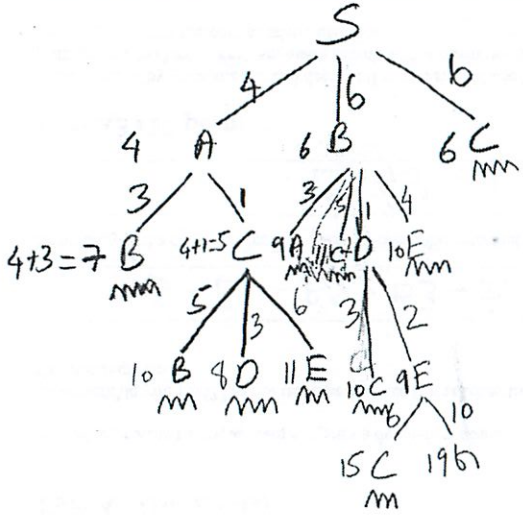
Alex has been chatting online for weeks with his new online girlfriend, Eliza. Stephanie and Maria are skeptical of this "Eliza" and insist that Alex meet with her in person, so when Alex asks Eliza out, Stephanie and Maria decide to come along to make sure the girlfriend is legit. They have to travel along the following streets, from S to G, and Alex wants to make sure he's on time to meet Eliza, so he insists that they take the shortest path.



Part B1 (10 points)

Using Branch and Bound with an Extended List, what is the final path from S to G? Be sure to show your goal tree for partial credit.

S B D E G ✓



What nodes are in your extended list?

S, A, C, B, D, E, G

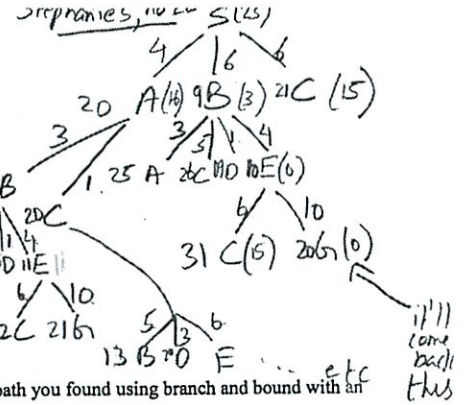
attempts to extend → but it is the goal node ✓

Part B2 (10 points)

Stephanie and Maria each suggest a heuristic:

Stephanie's heuristic:
S—25, A—16, B—3, C—15, D—12, E—0, G—0

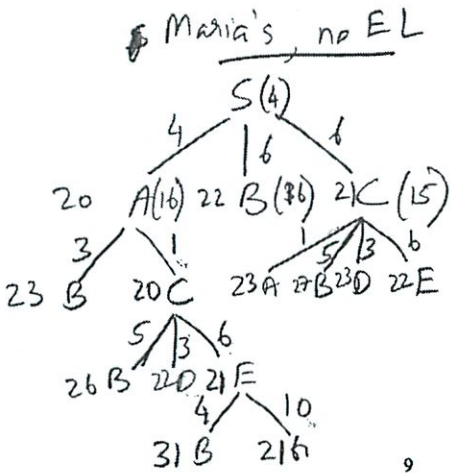
Maria's heuristic:
S—4, A—16, B—16, C—15, D—14, E—10, G—0



Which of the following searches will find the same path you found using branch and bound with an extended list?

Circle those that work; cross out those that do not.

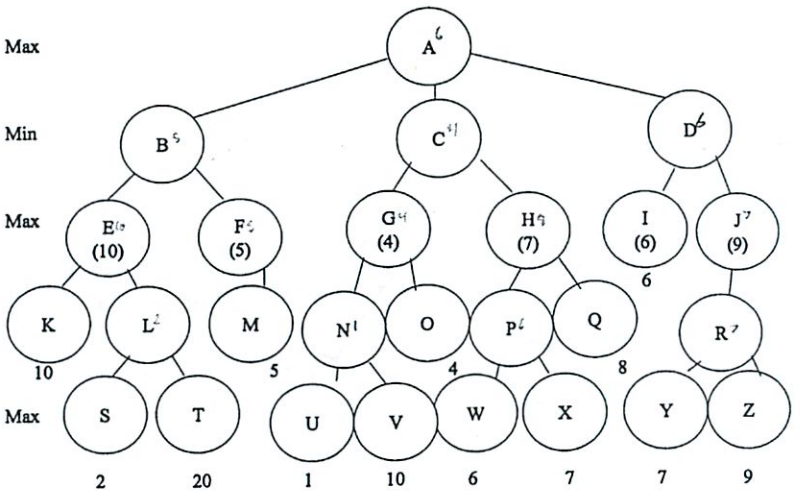
- ~~Branch and bound with Stephanie's heuristic and **NO** extended list.~~ — 5
- ~~Branch and bound with Maria's heuristic and **NO** extended list.~~
- ~~Branch and bound with Stephanie's heuristic and an extended list (aka A*)~~
- ~~Branch and bound with Maria's heuristic and an extended list (aka A*)~~



Quiz 2, Question 1, Games (50 points)

You are playing a new Sim game called Obamaquest, the Legend of the Lost International Credibility. In this game, you play a charismatic incoming president who must make a choice on various issues in order to save your country. After each of your turns, the outgoing president will attempt to perform the most meddlesome acts possible to make it less likely that you will succeed. You realize quickly that you can model this game using a simple Game Tree from 6.034, as shown below.

Static values are shown underneath leaf nodes. Ignore the numbers in parentheses for now.



Part A (15 points)

First, you decide to perform a simple minimax algorithm on the tree.

Which direction will the maximizer choose to go at node A?

D

What is the minimax value of node A?

6

Which static evaluations did you perform? (write the nodes you statically evaluated, in order).

K, S, T, M, U, V, O, W, X, Q, I, Y, Z

Part B (25 points)

Minimax was taking too many static evaluations, so you use alpha-beta instead.

This time what direction will the maximizer choose to go at node A?

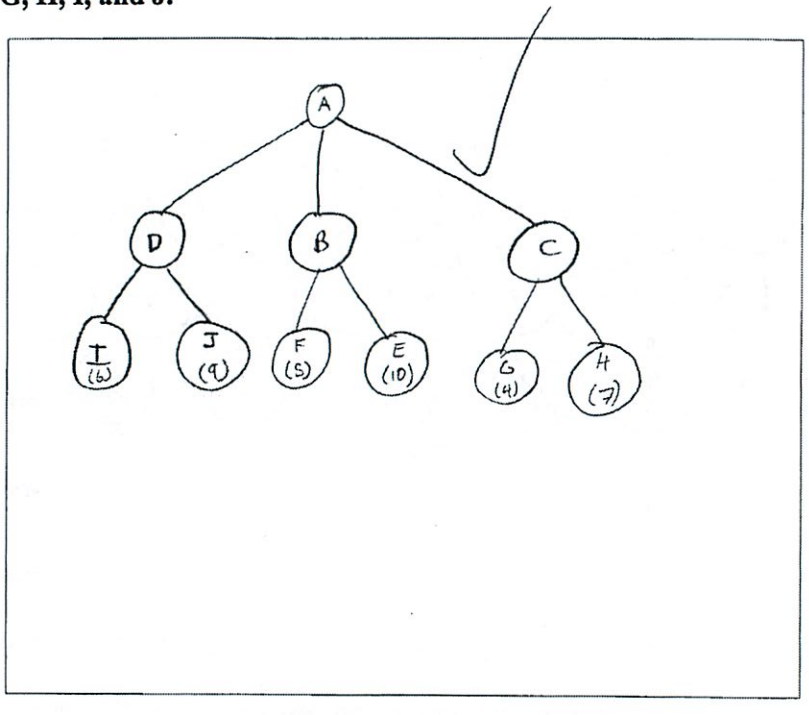
D

Which static evaluations did you perform? (write the nodes you statically evaluated, in order).

K, S, M, V, O, I, Y, Z

Part C (10 points)

The end result is still rather depressing in number of required static evaluations, so you decide to perform progressive deepening up to the second level and then reorder the tree to try for a more optimal pruning, using the static values for E,F,G,H,I, and J found in parentheses inside each circle. Draw your new ordering below. **You need not not draw any of the nodes below E, F, G, H, I, and J.**



Quiz 2, Question 2, Constraints (50 points)

Four 6.034 TAs (**Mark, Mike, Rob, Sam**) are trying to write eight questions for a final exam (somewhat like this one):

1. Constraints
2. Optimal Search
3. Games
4. Rules
5. ID-Trees
6. Neural Nets
7. SVMs
8. Boosting

Some questions are harder to write than others, so they should be distributed about equally. And some problems are so similar to others that for variety, we don't want those written by the same TA. This leads to some constraints:

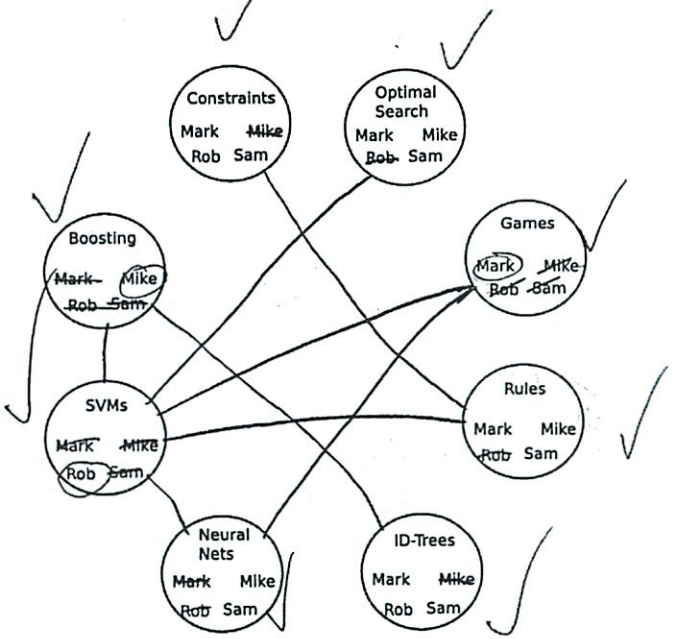
- TA(Constraints) != TA(Rules)
- TA(Boosting) != TA(ID-Trees)
- TA(Boosting) != TA(SVMs)
- TA(SVMs) != TA(Optimal Search)
- TA(SVMs) != TA(Games)
- TA(SVMs) != TA(Rules)
- TA(SVMs) != TA(Neural Nets)
- TA(Neural Nets) != TA(Games)

Also, there are some demands that have to be honored.

- Mike insists on doing Boosting
- Mike will not do Constraints
- Mark insists on doing Games
- Only Rob and Mike are willing to do SVMs.

Part A (10 points)

Draw lines between variables with a != constraint, and use the TAs' demands to reduce domains by crossing out names. Continue to reduce domains using the constraints while possible.



Part B (20 points)

Starting with your domains reduced in Part A, find a solution using depth-first search only, using no constraint propagation , checking constraints at assignments only. Consider TAs in alphabetical order: Mark, Mike, Rob, Sam. Please feel free to abbreviate unambiguously. **Draw your search tree on the next page.**

Constraints	Mark Rob Sam
Optimal Search	Mark Mike Sam
Games	Mark
Rules	Mark Mike Sam
	X
ID-Trees	Mark Rob Sam
Neural Nets	Mike Sam
SVMs	Rob
Boosting	Mike

Part C (15 points)

You consider two more advanced constraint propagation algorithms that

- propagate choices to neighbors only
- propagate choices to neighbors and continue through any domains reduced to size one

Do these algorithms find the same result as the DFS? Which domains do these algorithms reduce during the course of their runs? Circle the answers.

Neighbors only

Same result as DFS? (YES) NO

Domains reduced: C O G (R) I N S B

Neighbors and any domain reduced to size one

Same result as DFS? (YES) NO

Domains reduced: C O G (R) I N S B

Part D (5 points)

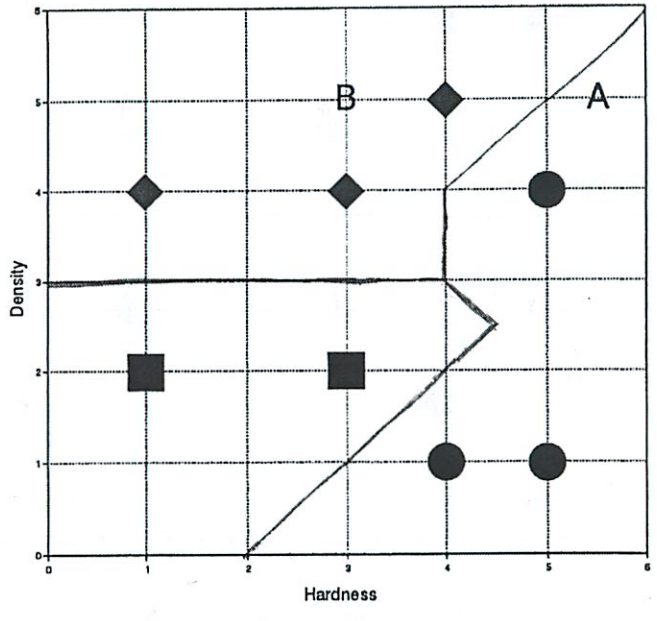
One TA thinks the assignments made in this problem are unfair. Suggest a way to ensure the test questions are more evenly distributed across the four TAs.

Instead of trying Mark first at each level, try the TAs in a random order.

Quiz 3, Question 1, NNand ID trees (50 points)

Part A: Nearest Neighbors

On the following graph, draw the decision boundaries produced by 1-Nearest Neighbor. Ignore the letters A and B.



◆ Sedimentary
■ Metamorphic
● Igneous

How is Sample A classified by 1-NN?

Igneous

By 3-NN?

Sedimentary

How is Sample B classified by 1-NN?

Sedimentary

By 3-NN?

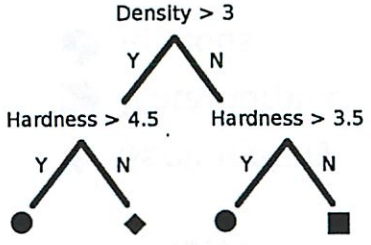
Sedimentary

Part B: ID Trees

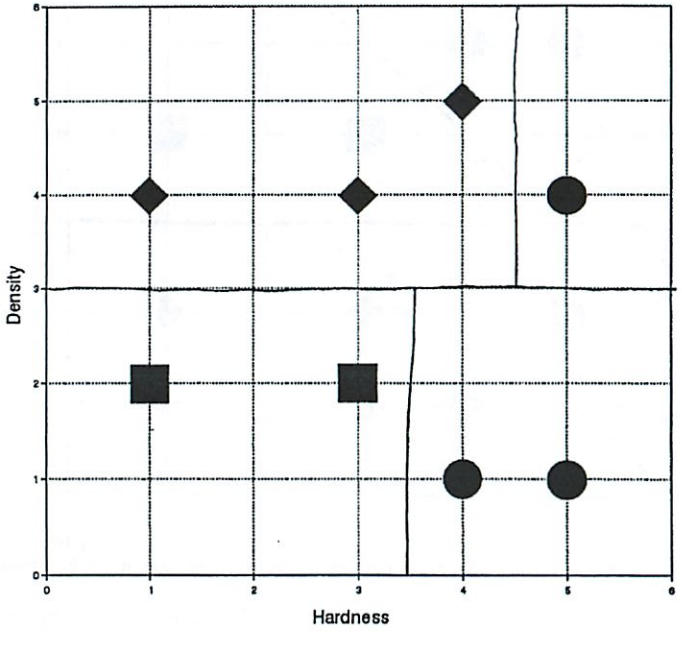
Using the same data as in Part A, calculate the disorder of the following ID Tree tests. Your answers may contain logarithms.

Density > 3 Y N	**Disorder:** $\frac{1}{2}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)+\frac{1}{2}\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)$
Hardness > 3.5 Y N	**Disorder:** $\frac{1}{2}\left(-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}\right)+\frac{1}{2}\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right)$

Suppose you created the following ID Tree.



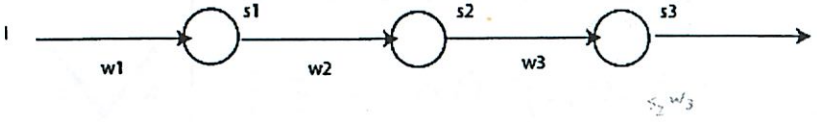
Draw the decision boundaries produced by that ID Tree on the graph below.



Quiz 3, Question 2, Neural Nets (50 points)

Part A (21 points)

Consider the following neural net. Note that all the neurons have sigmoid units, $s(z)=\frac{1}{1+e^{-z}}$ and the performance function is $P=-\frac{1}{2}(s_3-d)^2$



Note that the input to the net is i and the outputs of the sigmoid units are $s_1, s_2,$ and s_3.

In terms of $i, w_1, w_2, w_3, i, s_1, s_2, s_3,$ and d calculate the following partial derivatives:

$\frac{\partial P}{\partial w_3} = \frac{\partial P}{\partial s_3}\frac{\partial s_3}{\partial w_3} = -(s_3-d)s_3(1-s_3)$

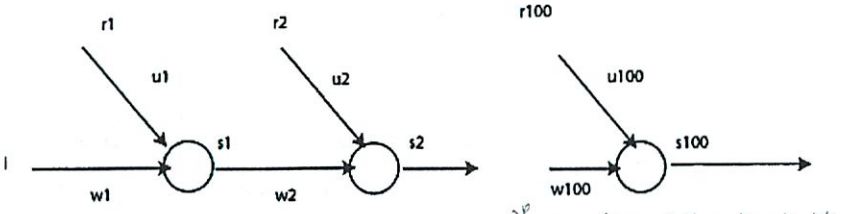
$\frac{\partial P}{\partial w_2} = s_2(1-s_2)w_3\frac{\partial P}{\partial w_3}$

$\frac{\partial P}{\partial w_1} = s_1(1-s_1)w_2\frac{\partial P}{\partial w_2}$

Part B (24 points)

Consider the following neural net. There are 100 neurons. All neurons have sigmoid units, $s(z)=\frac{1}{1+e^{-z}}$ and the performance function is $P=-\frac{1}{2}(s_{100}-d)^2$

The inputs to the network are i and r_i .The outputs of the sigmoid units in the network are s_i



$\frac{\partial P}{\partial w_{100}} = r_{100}(1-r_{100})(d-r_{100}) = \rho(1-\rho)(d-\rho)$

$\frac{\partial ?}{\partial w_{100}} = s_{100}(1-s_{100})(d-s_{100}) = \sigma(1-\sigma)(d-$

$\frac{\partial P}{\partial w_{99}} = s_{99}(1-s_{99})\left(w_{100}\frac{\partial P}{\partial w_{100}}\right)$

$= \sigma(1-\sigma)(\omega\sigma(1-\sigma)(d-\sigma))$

$= \omega\sigma^2(1-\sigma)^2(d-\sigma)$

At a certain time, t, $i=y$ and

$r_1=r_2=\cdots=r_{100}=\rho$

$s_1=s_2=\cdots=s_{100}=\sigma$

$u_1=u_2=\cdots=u_{100}=1$

$w_1=w_2=\cdots=w_{100}=\omega$

Calculate $\frac{\partial P}{\partial w_1}$ in terms of y ρ σ ω and d .

$\frac{\partial P}{\partial w_1} = \omega^{99}\sigma^{100}(1-\sigma)^{100}(d-\sigma)$

Part C (5 points)

Finally, exhibit an equation that relates σ , ρ and ω .

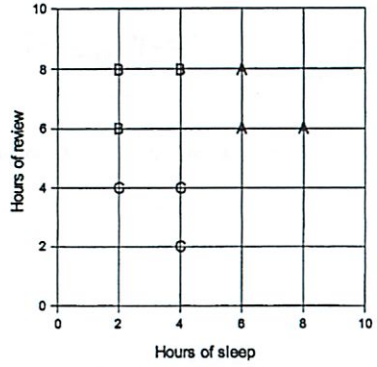
Do not attempt to solve the equation.

$\sigma = \frac{1}{1+e^{-(\rho+\omega\sigma)}}$

Quiz 4, Question 1, Support Vectors (50 points)

Miriam has a lot of finals. She wants to know how best to use her time, so naturally she collects some data to train a support vector machine.

She finds that nine of her friends have already taken the class that her next final is in, so she asks them how long they spent reviewing for the class's final exam, how many hours of sleep they got the night before the final, and what grade they got (A, B, or C). The results appear on this graph:



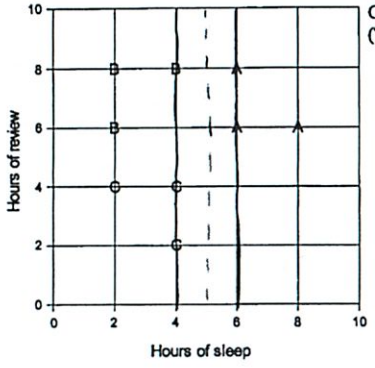
The problem she runs into is that SVMs can typically only distinguish two classes, and her data contains three classes of grades. However, she finds a way to make a three-class SVM:

- Divide the problem among three ordinary, two-class SVM classifiers, called $h_A(\mathbf{x})$, $h_B(\mathbf{x})$, and $h_C(\mathbf{x})$.
- Each SVM treats one of the classes (A, B, and C respectively) as +, and treats the other two classes as –.
- The final classifier outputs A, B, or C based on which of the three sub-classifiers outputs the highest value. For example, if h_A outputs -5, h_B outputs -3, and h_C outputs -2, the overall result should be C. If h_A outputs 2, h_B outputs -1, and h_C outputs -3, the overall result should be A.

Part A (30 points)

The three SVM classifiers are linear, so their output will be defined by the equation $h = \mathbf{w}\cdot\mathbf{x} + b$. On the next three graphs:

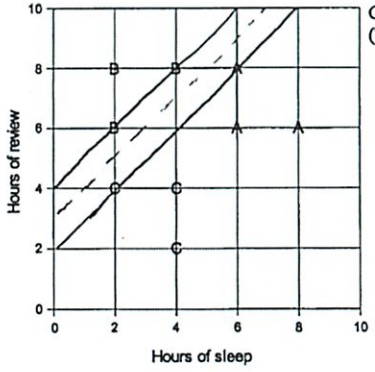
- Draw the "street" that separates the data by drawing a dotted line at the $h = 0$ boundary, and solid lines at $h = 1$ and $h = -1$.
- Write the values of w and b for that classifier.



Classifier A: $h_A = \mathbf{w}_A\cdot\mathbf{x} + b_A$
(This classifier separates A's from other grades.)

$\mathbf{w}_A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$

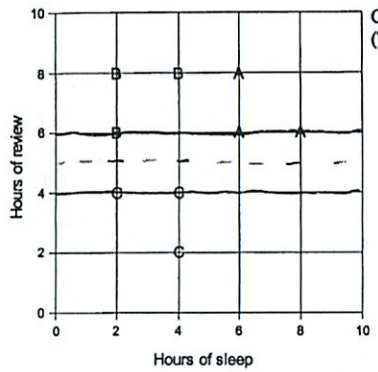
$b_A = -5$



Classifier B: $h_B = \mathbf{w}_B\cdot\mathbf{x} + b_B$
(This classifier separates B's from other grades.)

$\mathbf{w}_B = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

$b_B = -3$



Classifier C: $h_C = w_C \cdot x + b_C$
(This classifier separates C's from other grades.)

$w_C = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$

$b_C = 5$

Part B (6 points)

After running the three SVMs, Miriam's computer needs to determine which class wins overall. Remember that this is determined by which classifier outputs the *highest* value.

To determine whether classifier A beats classifier B, one can subtract their equations, giving a new equation in terms of the vector **x**. Calculate these equations for all pairs of classifiers – this will help to analyze which classifier wins at each point.

$h_A(x) - h_B(x) = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \cdot x - 2$

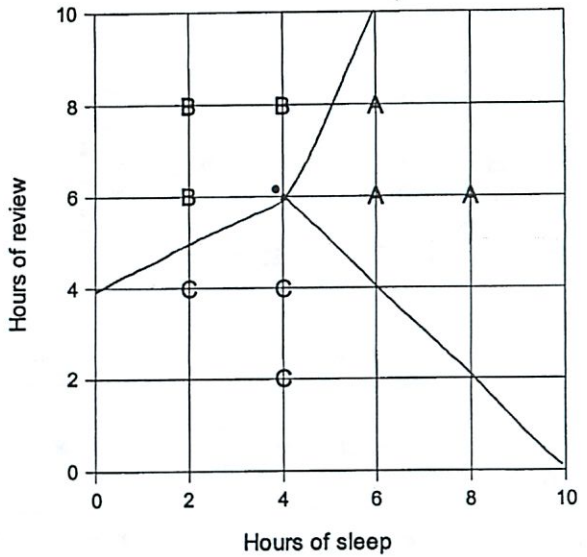
$h_B(x) - h_C(x) = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \cdot x - 8$

$h_A(x) - h_C(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot x - 10$

Part C (8 points)

Draw the boundary lines that separate classes A, B, and C in the final classifier on this graph.

The intent here is ***not*** for you to have to solve the equations down to the decimal point. We're looking for a graph that graphically shows the correct overall behavior. In fact, if you're trying to find the point where all three classifiers output the same value, we'll save you the effort by telling you that it should be at 3.8 hours of sleep and 6.2 hours of review.



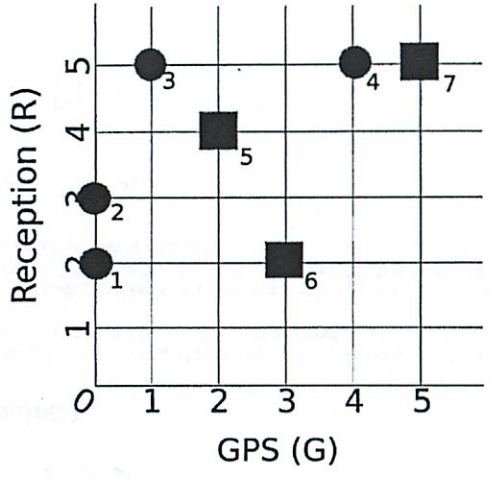
Part D (6 points)

Describe a way in which the behavior of your classifier is counterintuitive, given the meaning of the input data?

Increasing the hours of sleep moves from a B to a C in some areas, and increasing the hours of review moves from an A to a B in other areas (for example, given 5 hours of review, no sleep is better than 4 hours of sleep according to this classifier)

Quiz 4, Question 2, Boosting(50 points)

You want to use your cell phone to detect whether you are indoors or outdoors. You use the number of bars of reception (R) and your GPS signal quality (G) as features. Here's the data you gather:



O = +
□ = −

Part A (10 points)

You want to choose a learning algorithm for this data. Circle all the algorithms that can learn this data with zero training error:

AdaBoost ID-Tree Linear SVM RBF SVM

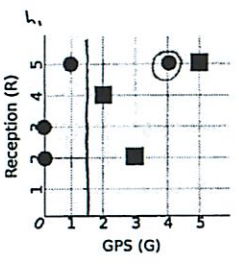
RBF = Radial basis function, the exponential kernel; you can choose the RBF's σ.

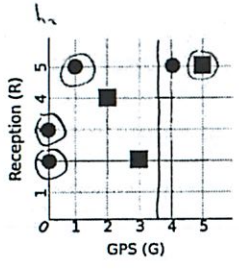
Part B (10 points)

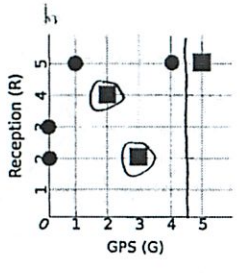
You decide to train AdaBoost using these stumps:

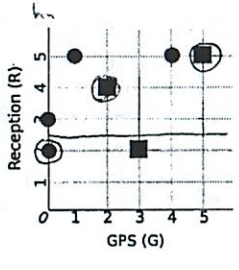
- $h_1(G,R) = \text{sign}(1.5 - G)$
- $h_2(G,R) = \text{sign}(G - 3.5)$
- $h_3(G,R) = \text{sign}(4.5 - G)$
- $h_4(G,R) = \text{sign}(R - 2.5)$
- $h_5(G,R) = \text{sign}(R - 4.5)$

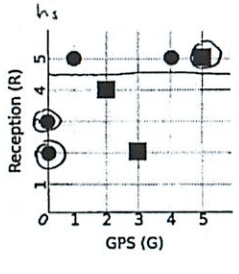
For each stump, draw the line and circle the misclassified data points.











Part C (30 points)

Starting with all weights set to 1/7, run AdaBoost for three rounds. For each round, give the classifier you chose and its error. Also give the final AdaBoost classifier and its error rate. (You don't need to give weights or alphas, but they'll help us give you partial credit.) Break ties by preferring low-numbered stumps.

$h_a(G,R) = h_1$

$\varepsilon_a = \frac{1}{7}$

$\alpha = \frac{1}{2} \ln 6$

new weights:

k	w_k
1	1/12
2	1/12
3	1/12
−4	1/2 = 6/12
5	1/12
6	1/12
7	1/12

$h_b(G,R) = h_3$

$\varepsilon_b = \frac{1}{6}$

$\alpha = \frac{1}{2} \ln 5$

k	w_k
1	1/20
2	1/20
3	1/20
4	6/20
−5	1/4 = 5/20
−6	1/4 = 5/20
7	1/20

$h_c(G,R) = h_5$

$\varepsilon_c = \frac{3}{20}$

$\alpha = \frac{1}{2} \ln \frac{17}{3}$

H(G,R)= $\text{sign}\left(h_1(G,R) + h_3(G,R) + h_5(G,R)\right)$

$\varepsilon = 0$

Quiz 5, Problem 1, (25 points)

Tired of searching for a new president every 10-15 years, the MIT corporation decides to elect a permanent, hereditary monarch. Times are tough, so they are overwhelmed with applicants and decide to hire you to create a kind of filter to reduce the candidate set to a manageable number.

Your first task is to pick some descriptors. You decide to look at intelligence, field, group memberships (if any) , and gender. Men an women form a set. Fields are grouped as at MIT: EE and MechE are part of engineering, Physics and Math are part of science, and XV is part of management. Engineering, science, management, architecture, and humanities are "MIT-type fields."

Using Arch learning, indicate in the tabel what is learned from each example and identify the heuristic involved by name, if known. If nothing is learn, put an x in the corresponding 2 columns.

Part A (18 points)

Candidate	Intelligent	Field	Member of	Gender	Heuristic	What is learned
Good	Yes	EE		Woman		(starting model)
Bad	No	EE		Woman	X	X
Bad	Yes	EE	Anarchists	Woman	forbid link	member of != anarchists
Good	Yes	MechE		Woman	climb tree	field = engineering
Good	Yes	MechE		Man	drop link	gender
Good	Yes	Physics		Man	climb tree	field = MIT-type
Good	Yes	Math		Man	X	X
Bad	No	XV		Man	X	X
Bad	Yes	XV	Anarchists	Woman	X	X

Part B (7 points)

Exhibit an example, which if used as the second example, would teach two characteristics at once.

Candidate	Intelligent	Field	Member of	Gender	What is learned
Good	no	EE		Man	drop link intelligent, drop link gender

Quiz 5, Problem 2, (25 points)

Circle the best answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

Experiments with the newly sighted children indicate that they have difficulty with

1. Naming colors
2. Recognizing objects that are moving
3. Recognizing objects while the child is moving
4. Recognizing overlapping objects
5. None of the above

A newly sighted child, presented with a picture of a ball and a picture of a cube will

1. Readily identify which is which
2. Think they are both the same
3. Think they are both balls
4. Not know which is which
5. None of the above

A newly sighted child will have trouble recognizing a triangle if

1. The triangle is seen against a background of short black lines
2. It is a filled with black, not just three black lines
3. Only three black lines are shown, with no fill
4. The triangle is seen between a circle and a square
5. None of the above

Experiments with newly sighted children suggest the key to visual bootstrapping is

1. Cooperation between vision and sound systems
2. Use of color in figure-ground separation
3. Restricted motion of Infants during the first few months of life
4. Clues provided by the visual motion detection system
5. None of the above

Human children...

1. Like cats, cannot have vision usefully restored if sightless beyond a critical period
2. Like cats, have trouble tracking moving objects if sightless beyond a criical period
3. Prefer to remain blind if sightless beyond a critical period
4. Cannot identfy geometric figures on a computer screen
5. None of the above

Quiz 5, Problem 3, (25 points)

Circle the best answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

Disorientation experiments with rats, children, and adults demonstrate

1. Rats will not move in the presence of loud rock music
2. Children that behave like rats do not spontaneously use words like left and right
3. Rats and infants do not have full-spectrum color vision
4. Neither small children nor adults cannot reorient themselves if severely sleep derprived
5. None of the above

Chomsky believes the great differentiator between modern humans and humans 100,000 years ago is

1. Improved mitochondria
2. Improved memory
3. Improved hand-eye coordination
4. Ability to plan
5. None of the above

Brooks subsumption architecture is best described as

1. A demonstration of the importance of reasoning and planning in robot activity
2. An approach demonstrated by a robot that collected beer cans on cape-cod beaches
3. An approach to building robots focused on tightly coupled loops between vision and language
4. An idea reminiscent of the notion of abstraction layers in programming methodology
5. None of the above

Winston believes that to develop an account of human intelligence we must

1. Find new ways to exploit the SOAR architectural paradigm
2. Build systems that acquire permanent knowledge from visual experiences
3. Collect commonsense knowledge using volunteers contributing via the web
4. Focus for the time being on what insects can do
5. None of the above

The General Problem Solver Architecture

1. Is a general purpose architecture that unifies reasoning and perception
2. Is a special purpose architecture limited to proving theorems in first order logic
3. Uses a means-ends approach to operator selection
4. Gets stuck if a selected operator cannot be applied in the current state
5. None of the above

Quiz 5, Problem 4, (25 points)

Coen's method

1. Uses boosting to make clusters
2. Uses a neural net to make clusters
3. Uses subsumption to make clusters
4. Uses cross modal coupling to make clusters
5. None of the above

Coen's method

1. Explains why Zebra Finches use a mating song
2. Enables a program to learn to sing a Zebra Finch's mating song
3. Enables a program to learn to to recognize a Zebra Finch's mating song
4. Enables a program to recognize the mating song of particular Zebra Finches
5. None of the above

In Coen's method

- Two clusters are close if they are close in a Euclidian sense
- Two clusters are not close if they are not close in a Euclidian sense
- Two clusters are close if they project proportionately to clusters in another space
- Two clusters are not close if they have different shapes
- None of the above

Coen's method

1. Works well with any distance metric
2. Works well only with the Wasserstein distance metric
3. Works well only when the number of clusters is known in advance
4. Works well only when there are two or three clusters
5. None of the above

Coen's method

1. Works well only if corresponding points in two spaces are provided
2. Works well only if all the points in at lest one cluster are close together
3. Works well only if there is no noise
4. Works well only with labeled data
5. None of the above

if the points in the two spaces aren't provided it has nothing to work with...

Name	
email	

6.034 Final Examination
December 16, 2009

Circle your TA and principle recitation instructor so that we can more easily identify with whom you have studied:

Erica Cooper	Matthew Peairs	Mark Seifter
Yuan Shen	Jeremy Smith	Olga Wichrowska

Robert Berwick	Randall Davis	Gregory Martin

Indicate the approximate percent of the lectures, mega recitations, recitations, and tutorials you have attended so that we can better gauge their correlation with quiz and final performance. Your answers have no effect on your grade.

	Lectures	Recitations	Megas	Tutorials
Percent attended				

Quiz	**Score**	**Grader**
Q1		
Q2		
Q3		
Q4		
Q5		

There are 38 pages in this final examination, including this one. In addition, tear-off sheets are provided at the end with duplicate drawings and data. As always, open book, open notes, open just about everything.

Quiz 1, Problem 1, Rules (50 points)

The administration, worried about the social habits of its students, agrees to finance cross-school-mixers. The 034 TA's decide to fly to England and mix with the students at Hogwarts School of Witchcraft and Wizardry. A merry old time ensues, but the morning after, due to an accidental confundo charm (and perhaps also a large consumption of butterbeer), no one can remember the events that transpired. The 034 staff, in an attempt to show off the power of Muggle logic, promise they can piece together the important events with a rule based system.

Using their keen sense of logic, Matt, Erica, and Mark are able to piece together the following rules:

RULES:
R0 : IF (?X) goes to MIT,
THEN (?X) is a muggle,
(?X) consumed butterbeer

R1: IF (?X) made math jokes AND
(?X) consumed butterbeer
THEN (?X) was transfigured into a porcupine

R2: IF (?Y) fancies (?X) AND
(?X) fancies (?Y) AND
(?Y) is a muggle
THEN (?X) snogged (?Y)

R3: IF (?X) fancies (?Y) AND
(?X) made math jokes,
THEN (?Y) fancies (?X)

R4: IF (?X) made math jokes
THEN (?X) goes to MIT

You start with the following list of assertions which **is all you have to go on.**

ASSERTIONS:
A0: Olga made math jokes
A1: Yuan goes to MIT
A2: Jeremy made math jokes
A3: Hermione consumed butterbeer
A4: Jeremy fancies Hermione

Part A: Forward Chaining (24 points)

Run forward chaining on the rules and assertions provided for the first 5 iterations. For the first two iterations, fill out the first two rows in the table below, noting the rules whose antecedents match the data, the rule that fires, and the new assertions that are added by the rule. For the remainder, supply only the fired rules and new assertions. As usual, break ties using the earliest rule on the list that matches. If the earliest rule matches more than once, break ties by assertion order.

	Matched	**Fired**	**New assertions added to database**
1			
2			
3	[redacted]		
4			
5			

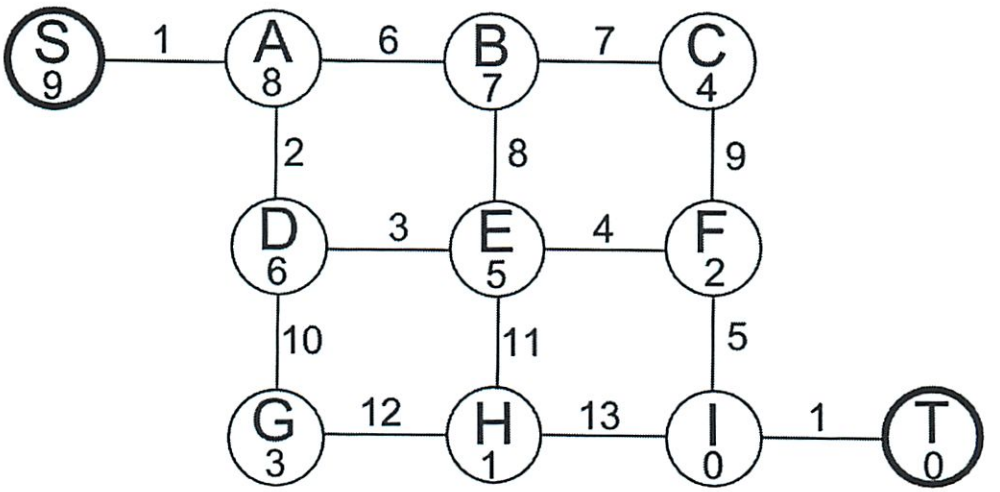
Part B: Backward Chaining (26 points)

Ron Weasley claims that Hermione snogged Jeremy. Use backward chaining to determine if this event occurred. **Draw the goal tree for this statement.** Partial credit will be given for partial completion of the goal tree.

Hermione snogged Jeremy

Is the claim that Hermione snogged Jeremy true?

Quiz 1, Problem 2: Search (50 points)



Aliens have invaded the MIT campus, thinking that the Stata Center was their mothership. Despite their realization that it was not, they went ahead with their invasion plan anyway. Advanced alien technology has disabled use of phones and internet, so you and your friends decide to go across the river to Boston to get help and to tell everyone what's going on. You decide to use the secret underground network of tunnels underneath the campus to get from your starting location at **S**, to the subway station at **T.**

You have the above graph of the underground tunnels. The edges are labeled with distances, and the nodes are labeled with a heuristic estimate to your destination at **T**. When performing search, ties are broken by choosing the node that is **alphabetically first.**

Note that node G is not the goal node in this problem.

Part A: Depth-First Search. (15 points)

A1: You first attempt to pick your route using **Depth-first search with an extended list.** Draw your search tree:

Also, show your extended list at the time your search terminates:

A2: What is the final path that is found using depth-first search?

Part B: Beam Search (15 Points)

B1: Next, you try to pick your route using **Beam Search with a beam width of 2, using an extended list.** In the event of a tie, use the alphabetical order of the final node on the contending partial paths. Draw your search tree:

Also, show your extended list at the time your search terminates:

B2: What is the route found using beam search?

Part C: Branch and Bound (20 Points)

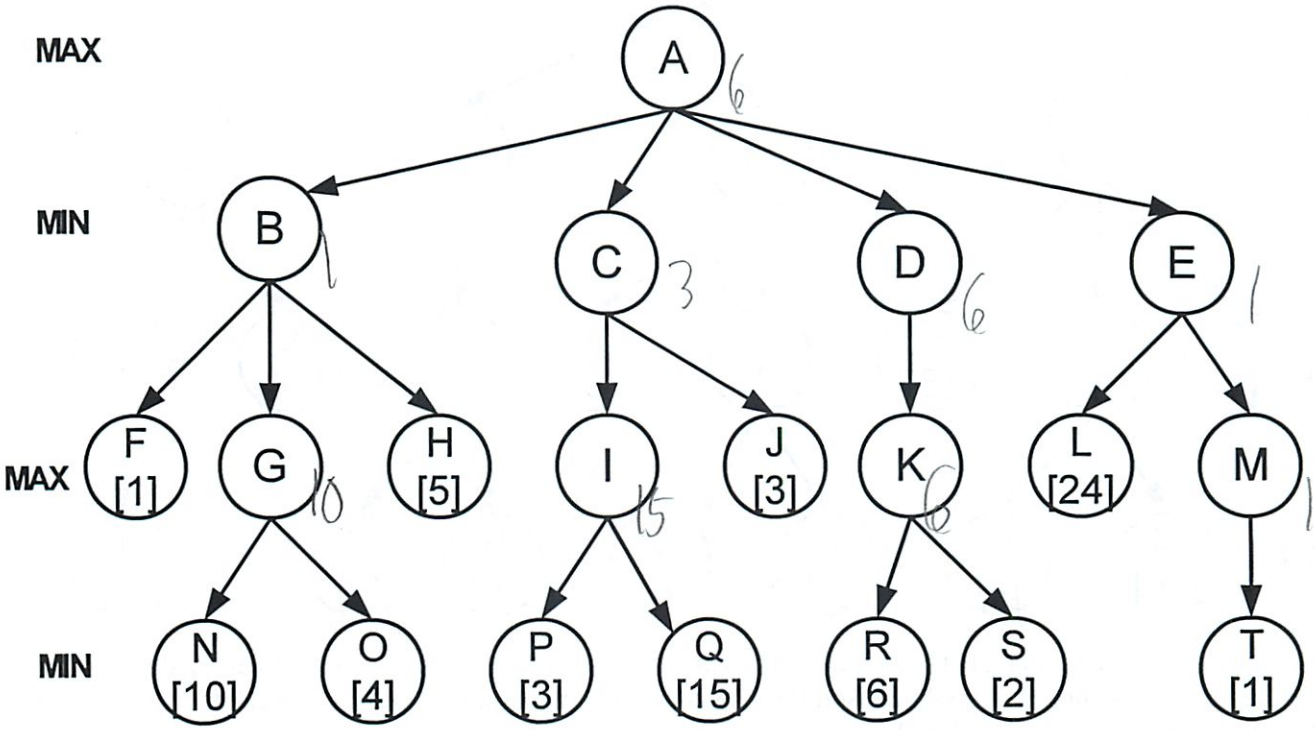
C1: Lastly, you try to plan your route using **Branch and Bound using the path lengths indicated on the graph and also using an extended list.** Draw your search tree and extended list below:

Also, show your extended list at the time your search terminates:

C2: What is the route found using branch and bound?

C3: If you repeat the search using the **A* algorithm with the heuristic values indicated on the graph,** will you find the same route? Include a brief explanation of why or why not.

Quiz 2 Problem 1: Games (50 points)



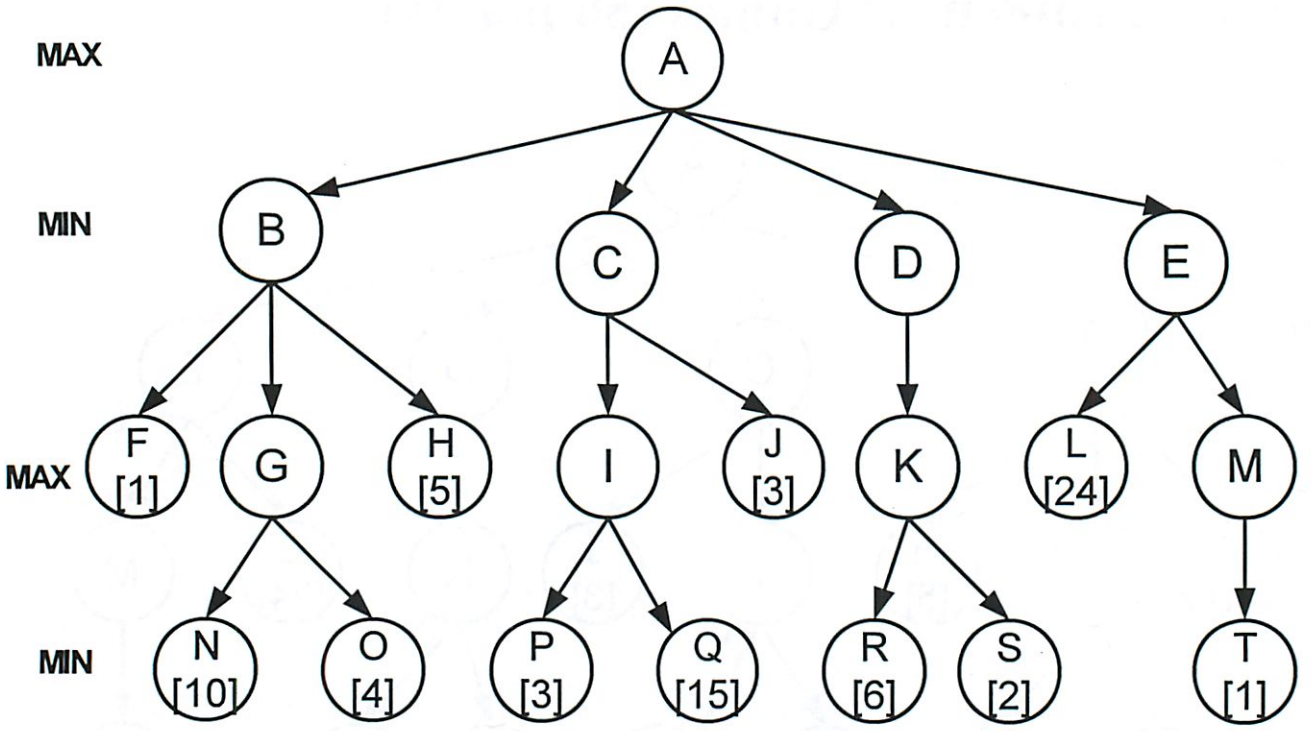
A: Using minimax only, no alpha-beta,, indicate the values of the following nodes. (10 pts)

A =	6	G =	10
B =	1	I =	15
C =	3	K =	6
D =	6	M =	1
E =	1		

✓

B: Using minimax only, what is the best next move from A? **(Indicate a letter)** (4 pts)

D ✓



C: Trace the steps of Alpha beta pruning on the same tree above. Note that alpha, betas are updated before pruning occurs, if in doubt consult the reference implementation given on the tear-off sheet. **List the leaf nodes in the order that they are statically evaluated.** (10 pts)

What are the final Alpha Beta values at node E (4pts)

Alpha =	Beta =

What are the final Alpha Beta values at node A (4pts)

Alpha =	Beta =

D: For a full **binary tree** (branching factor = 2) **of depth 3** (4 layers of nodes including the layer with the root node, 8 nodes in the bottom layer), with the root node being a MAX node, what is the fraction of leaf nodes that are statically evaluated under alpha beta pruning under conditions of maximum pruning? (11 pts)

Use the space below to show your work:

E: For a given depth, d, does the fraction requested in part D increase, decrease, or stay the same with increasing b. Circle your answer (7 pts)

Increase

Decrease

Stay the same

Explain:

Quiz 2, Problem 2:
Constraint Propagation (50 Points)

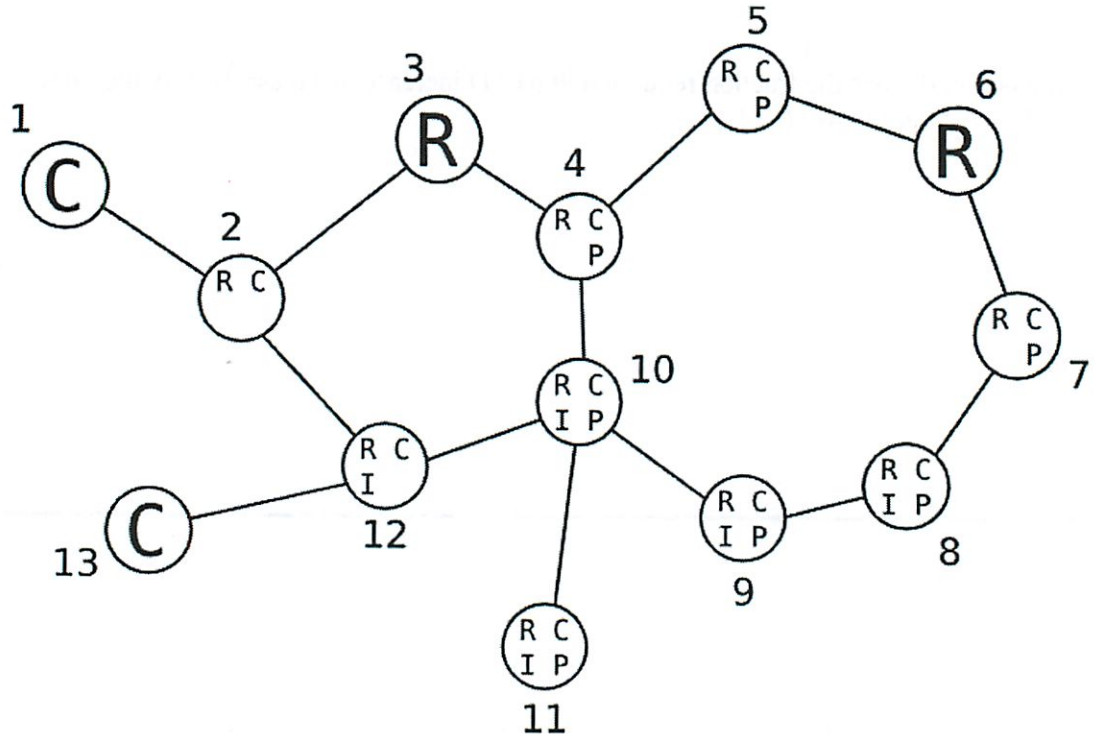
Important note: there is no search in this problem, only constraint propagation.

On the newly colonized planet Mars, cities consist of pressurized domes connected by a series of tubes. Each dome is designated for a specific type of use, and there are some restrictions on what sorts of domes can be connected to each other:

1. A **R**esidential dome can only be connected to another **R**esidential dome, a **C**ommercial dome, or a **P**ark.
2. A **C**ommercial dome can only be connected to a **R**esidential dome, another **C**ommercial dome, or an **I**ndustrial dome.
3. An **I**ndustrial dome can only be connected to a **C**ommercial dome or another **I**ndustrial dome.
4. A **P**ark can only be connected to a **R**esidential dome.

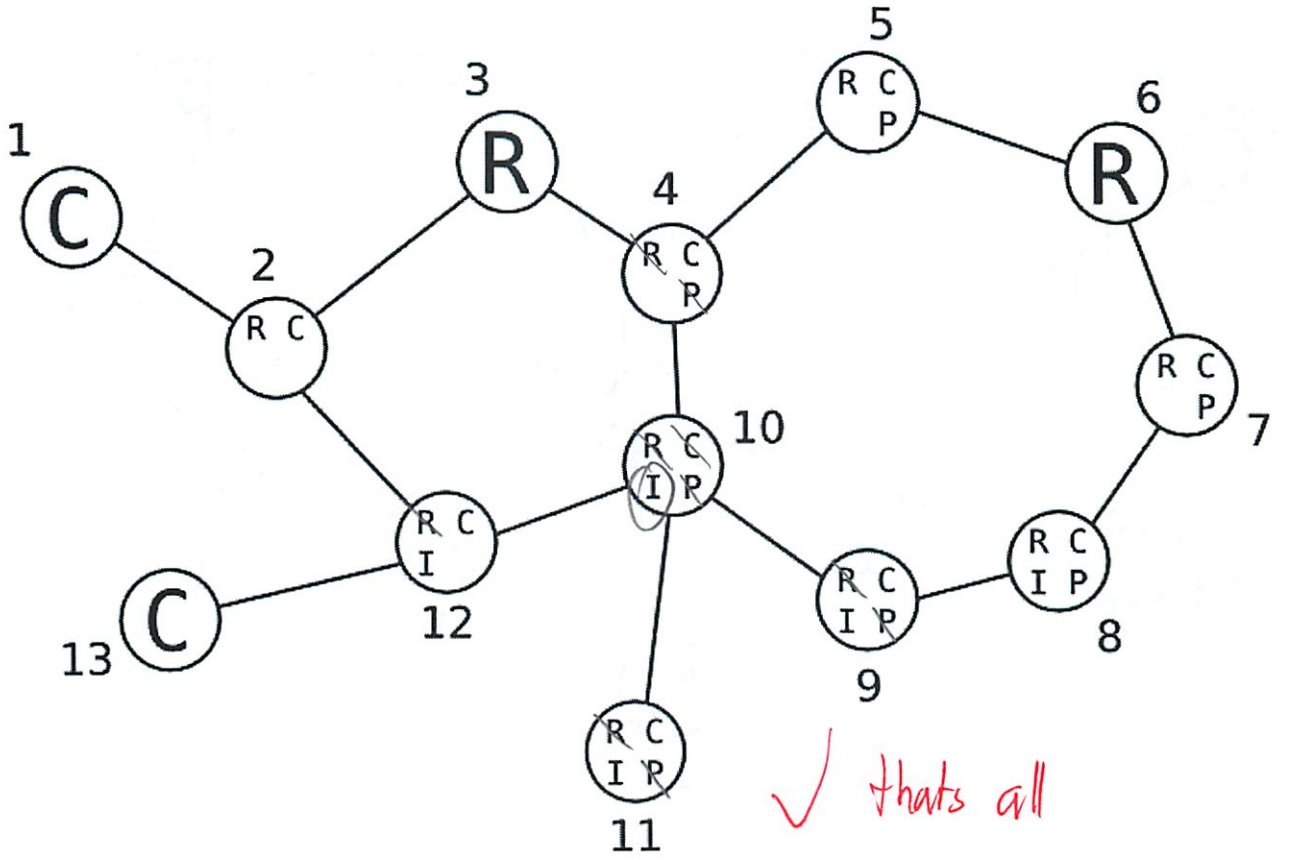
Silly

You decide to use constraint propagation to design a new Martian city. You begin with the following partially completed plan, with some domes already designated and their neighboring domains reduced accordingly:



Part A (12 points)

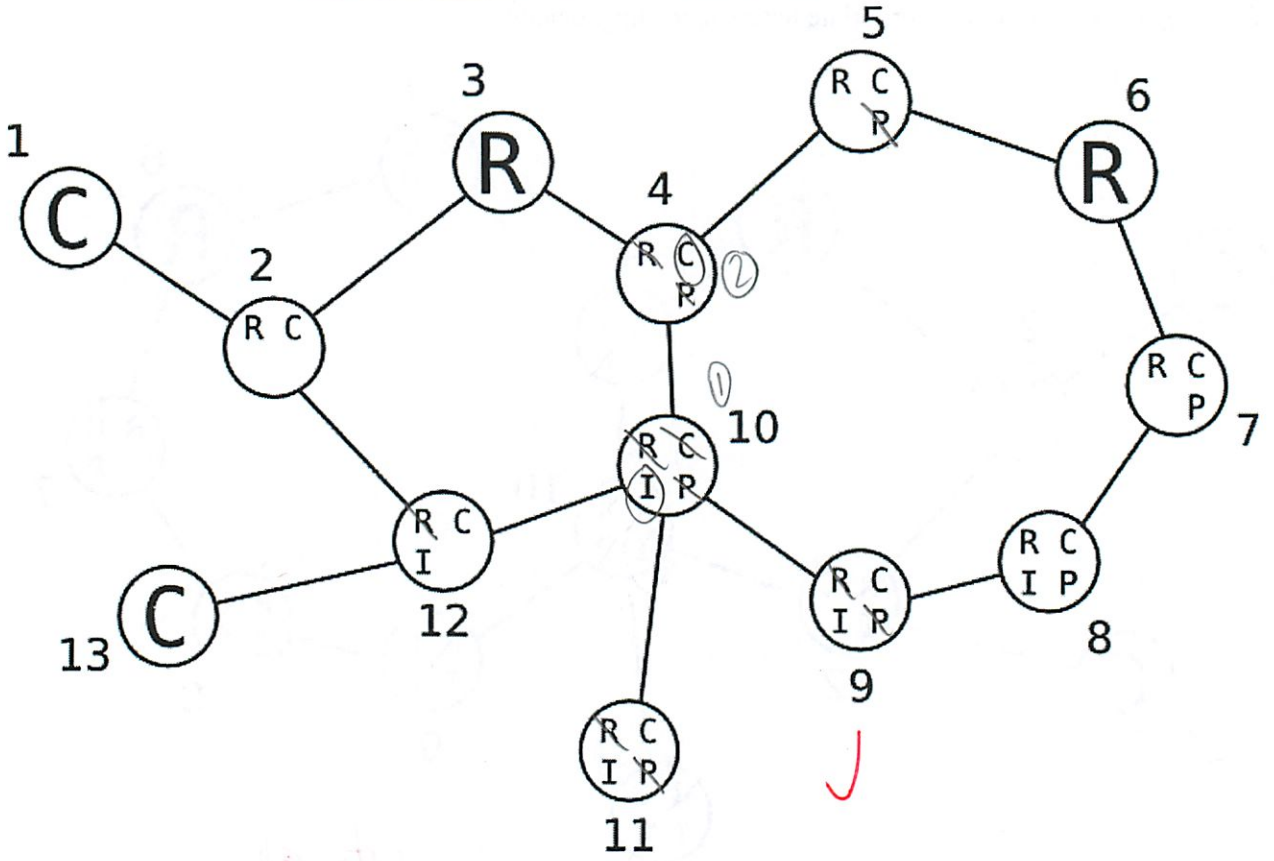
You decide to begin by designating Dome #10 as Industrial. Using forward checking (no propagation beyond the neighbors of the just-assigned variable), show how the domains of the city's domes are reduced by crossing out the appropriate letters in the map below:



✓ thats all

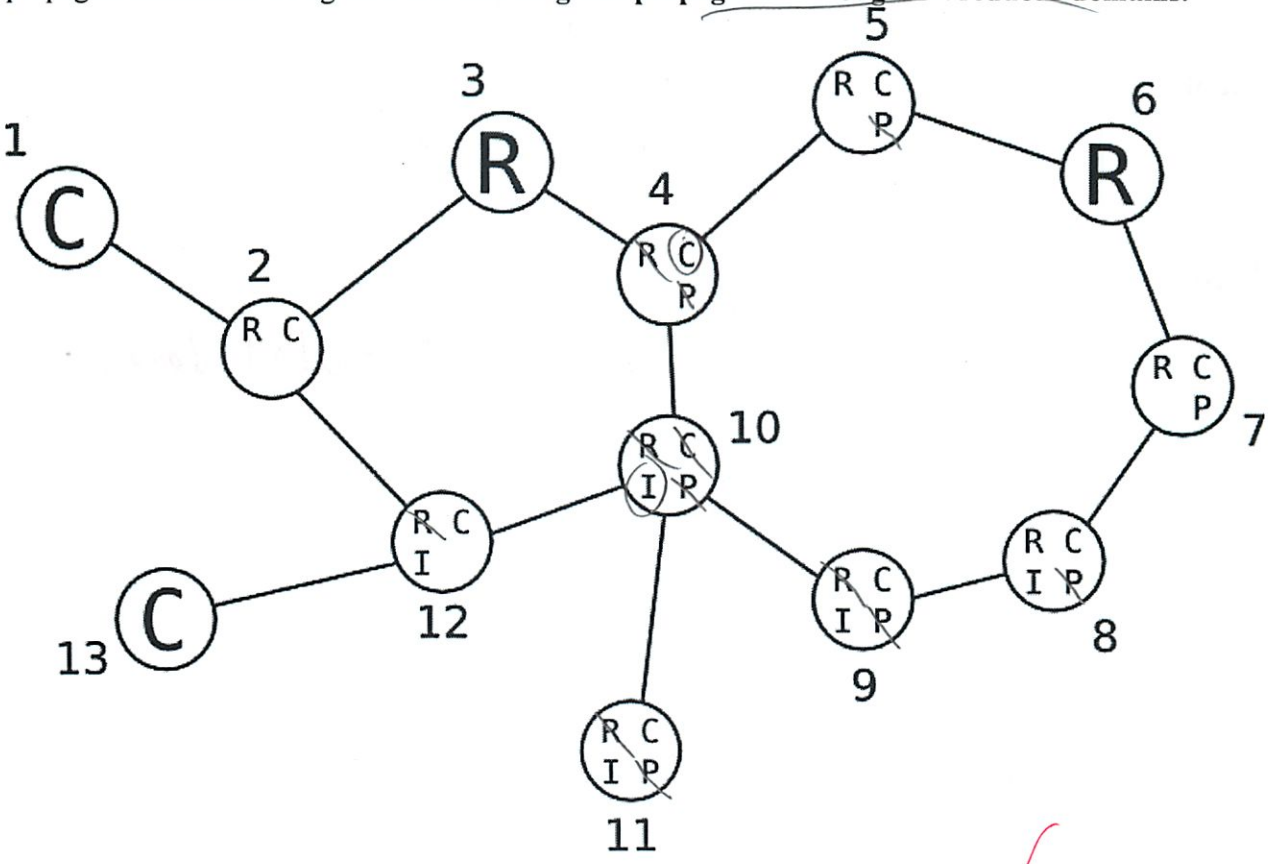
Part B (12 points)

Next, show how the domains are reduced if you designate Dome #10 as **Industrial**, and then propagate constraints using **forward checking and propagation through singleton domains.**



Part C (12 points)

Finally, show how the domains are reduced if you designate Dome #10 as **Industrial**, and then propagate constraints using **forward checking and propagation through all reduced domains.**



Part D (14 points)

Yuan says that you are right to pick a dome type for Dome #10 first, but Olga advises you to start with Dome #2 instead. What is the reasoning behind each of these suggestions?

1. start w/ most constrained

Yuan:

start w/ most constrained ✓ ← better?

Olga:

Start w/ least options ✓ Smallest domain.

✓ That was a very easy CP problem!

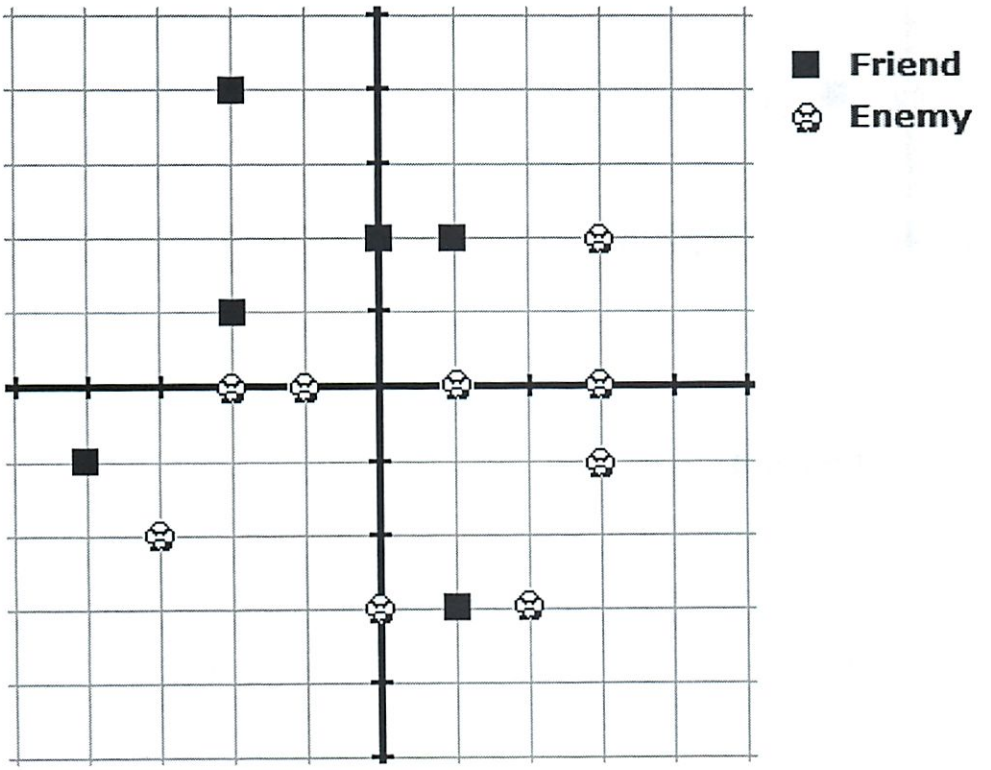
Quiz 3, Problem 1:
KNN & ID-Trees (50 points)

The Mario Bros. have hired you to help them identify potential dangers during their frequent expeditions to explore strange new worlds, battle evil monsters, and attempt to rescue some princesses.

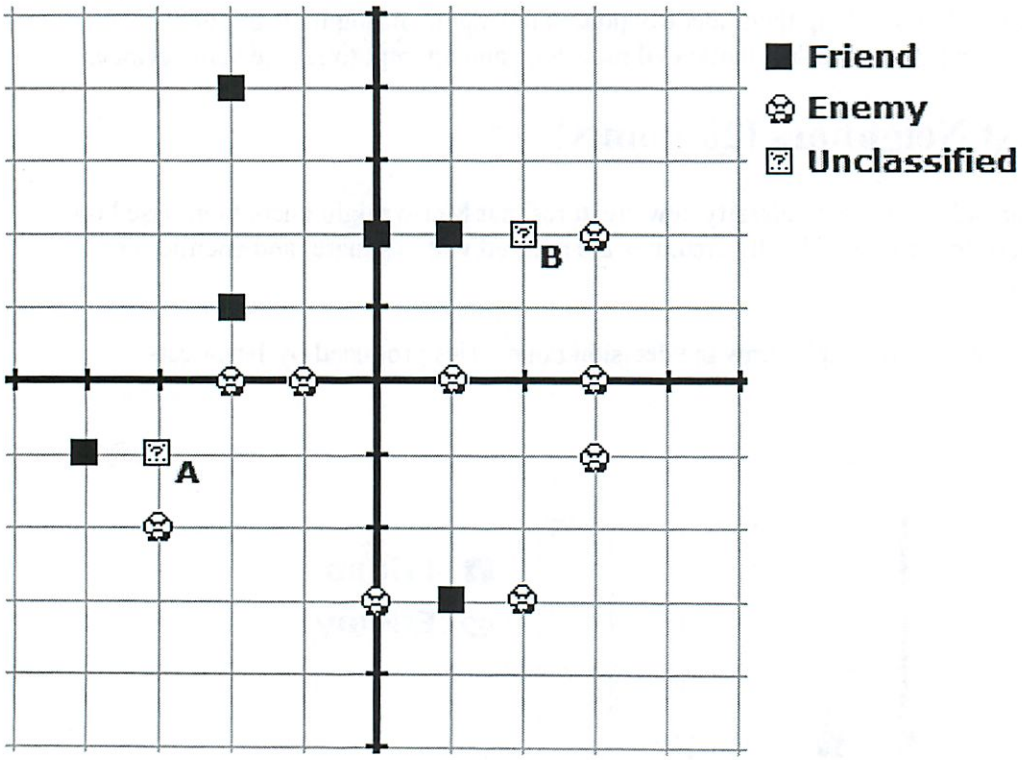
Part A: Nearest Neighbors (20 Points)

You decide to use Nearest Neighbors to classify new creatures that Mario might encounter, based on their positions on a map. In the map, friendly creatures are marked with a square, and enemies are marked with monster icons.

A1 (12 points): On the following graph, draw the decision boundaries produced by 1-nearest-neighbors.



A2 (8 points): The graph below shows two new creatures, marked with a question mark symbolsand labeled A and B. Show how these will be classified using 3-nearest-neighbors and 5-nearest-neighbors below.



	Creature A	Creature B
Using 3 NN:		
Using 5 NN:		

Part B: ID-Trees (30 Points)

It turns out that enemies move around the world, so a simple K-Nearest-Neighbors on their location won't do a very good job of protecting Mario. Instead, you decide to use an Identification Tree, based on some characteristics of the creatures in this world, to classify creatures as Enemies (Enemy=Y) and Friends (Enemy=N). The other 5 characteristics you note are in the table below.

	Character	Enemy?	Talks?	Boss?	Annoying?	Killable with Jump?
A	Yoshi	N	N	N	N	N
B	Tree	N	N	N	N	N
C	Luigi	N	N	N	N	N
D	Toad	N	Y	N	N	N
E	Peach	N	Y	N	Y	N
F	Bowser	Y	N	Y	N	N
G	Bob-omb	Y	N	N	N	Y
H	Goomba	Y	N	N	N	Y
I	Piranha	Y	N	N	Y	N
J	Dry Bones	Y	N	N	Y	N
K	Chain Chomp	Y	N	N	Y	N
L	Thwomp	Y	N	N	Y	N
M	Boo	Y	N	N	Y	N
N	Koopa Troopa	Y	N	N	Y	Y

B1 (8 points): What is the disorder of the test **"Boss = Y"**? Leave your answer in terms of fractions, real numbers, and logarithms.

B2 (7 points): What is the disorder of the test **"Annoying = Y"**? Leave your answer in terms of fractions, real numbers, and logarithms.

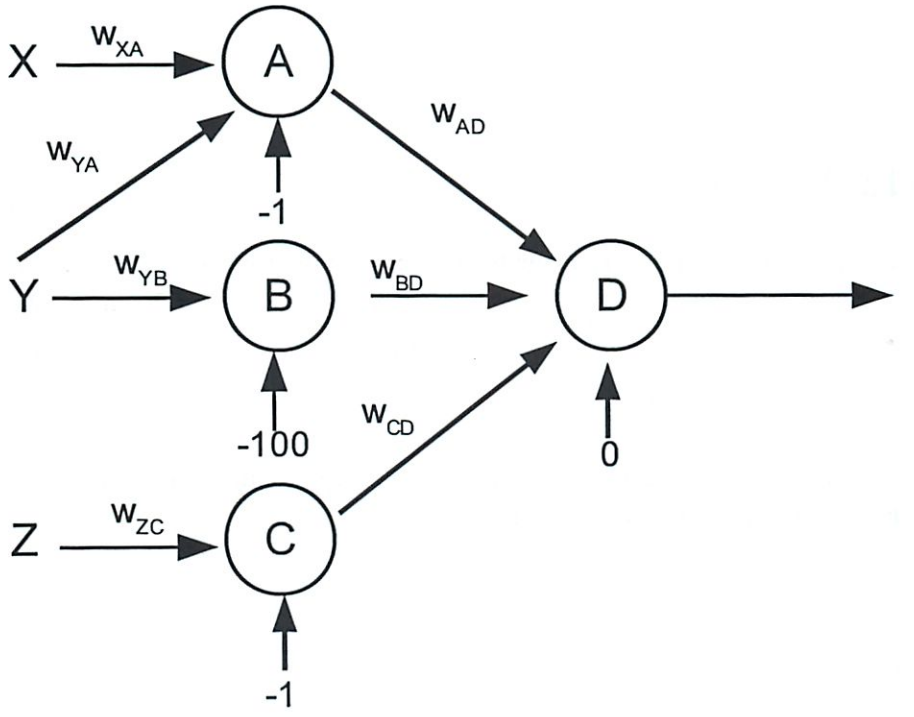
B3 (15 points): Draw the **disorder minimizing** identification tree Mario can use to correctly decide whether any of the above examples is an enemy (Enemy=Y) or friend (Enemy=N). Use the letters provided next to characters' names to show how the characters are separated by each decision.

Hint: neither Boss nor Annoying are the first test.

Quiz 3, Problem 2:
Neural Nets (50 points)

Hermione Granger decides to spend a semester abroad at MIT for her Muggle Studies class. Upon arriving, Hermione befriends the coolest people on campus. The 034 staff, over a festive round of Butter Beer, then teach Hermione about neural nets. She decides to practice.

Note that the tear off sheets include the neural net tear-off sheet from Quiz 3.



Hermione creates the above neural net. Each node uses a standard sigmoid function. All threshold units have fixed weights of 1. **They are never updated.**

Part A: (12 Points)

For each node (A, B, C, and D), using only the weight variables provided in the diagram, d (desired output), and the outputs of each node (O_A, O_B, O_C, and O_D), determine δ_A, δ_B, δ_C, and δ_D.

Part B: (12 Points)

Hermione, as an intellectual exercise, runs the neural net with the following parameters:

X=0.5	Y = 0	Z = 5	W_{XA} = 2
W_{YA} = 2	W_{YB} = 1	W_{ZC} = 0.2	W_{AD} = 1
W_{BD} = 1.5	W_{CD} = -1		

Determine the output for each node (O_A, O_B, O_C, and O_D). Approximate sigmoid(x) very simply by letting it evaluate to 0 for x < -50 and 1 for x > 50.

Part C: (18 Points)

Run back-propagation using these same parameter values for one iteration to determine new weight values for W_{AD} , W_{BD}, and W_{CD}. Use your results from Part A and Part B, and assume a learning rate of r = 2 and a desired output, d=1.

Part D: (8 Points)

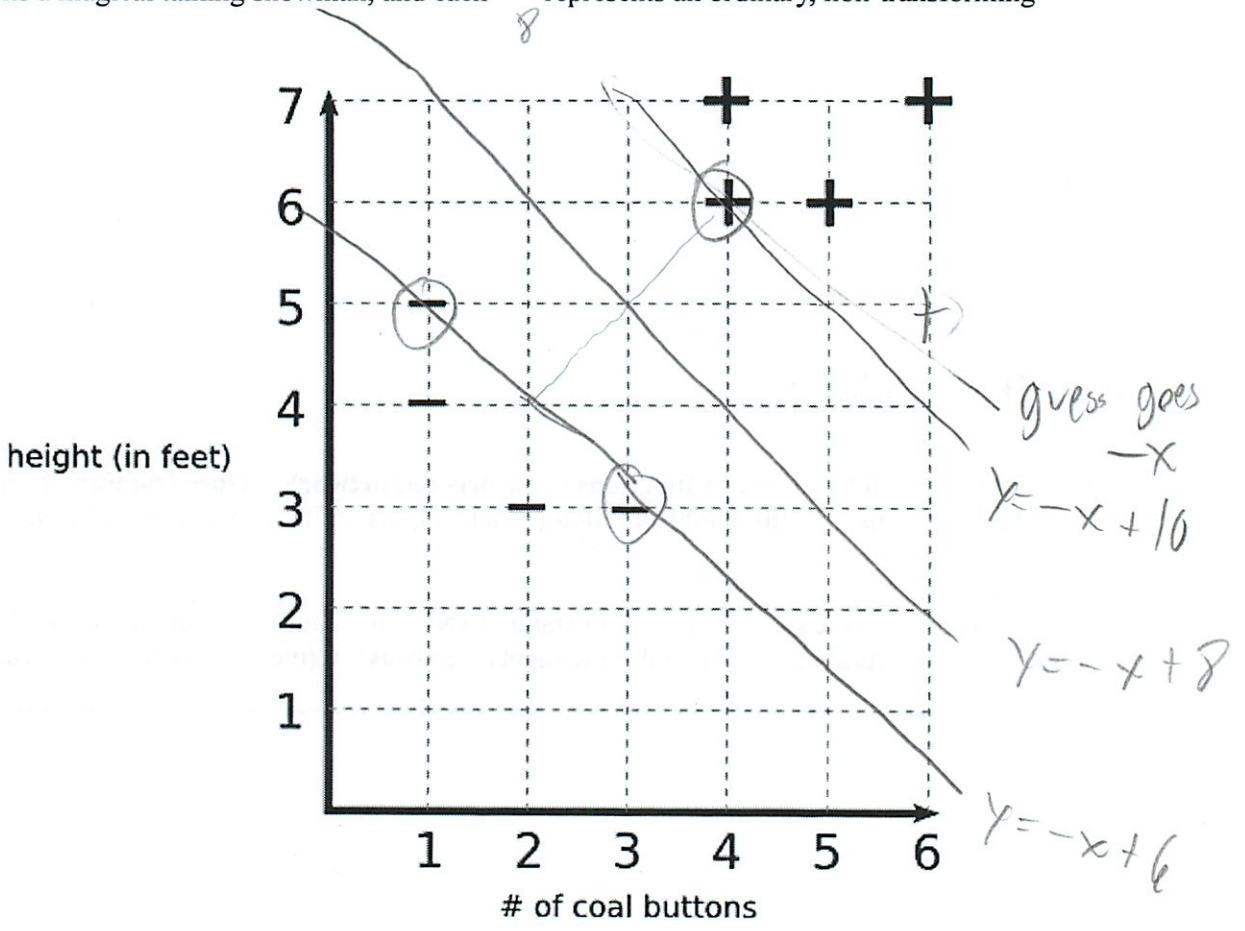
Hermione changes **all the decision functions** in the **original network, before training,** from using a sigmoid function to using a **threshold function** (which outputs 1 if the input is greater than 0, and outputs 0 otherwise).

Using X, Y, and Z and the standard logical operators (AND, NOT, and OR), determine a logic expression for the neural net. Assume that the inputs are always 0 (meaning false) or 1 (meaning true).

Quiz 4, Problem 1: SVMs (50 points)

Part A: F.R.O.S.T. and the Snowmen (25 Points)

You have been hired by the Foundation for Research into Occult Snow Transformations (F.R.O.S.T.) to investigate a strange new phenomenon. It seems that certain snowmen can be turned into living, talking creatures when an old silk hat is placed on their heads. So far, F.R.O.S.T. researchers have determined that the factors that best predict whether a particular snowman can be thus transformed are the snowman's height and the number of coal buttons on its chest. Their preliminary data is below: each "+" represents a magical talking snowman, and each "–" represents an ordinary, non-transforming snowman.



A1 (5 points): You decide to use a linear SVM to classify magical vs. non-magical snowmen. **Draw the resulting decision boundary in the graph above, and circle the support vectors.**

A2 (10 points): In this SVM, what is the vector **w** and the constant b?

$y \geq -x + 8$

$0 \geq -x - y + 8$

So $x + y - 8 \geq 0$ $\begin{bmatrix} c \\ c \end{bmatrix}$

So now $\frac{2}{\sqrt{c^2+c^2}} = 2\sqrt{2}$ $\quad \frac{1}{4} = c^2$ $\quad c = \frac{1}{2}$

$\frac{4}{2c^2} = 8$ $\quad 4 = 16c^2$ $\quad w = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$ $\quad b = -4$

Always $w_1 x + w_2 y + b \geq 0$

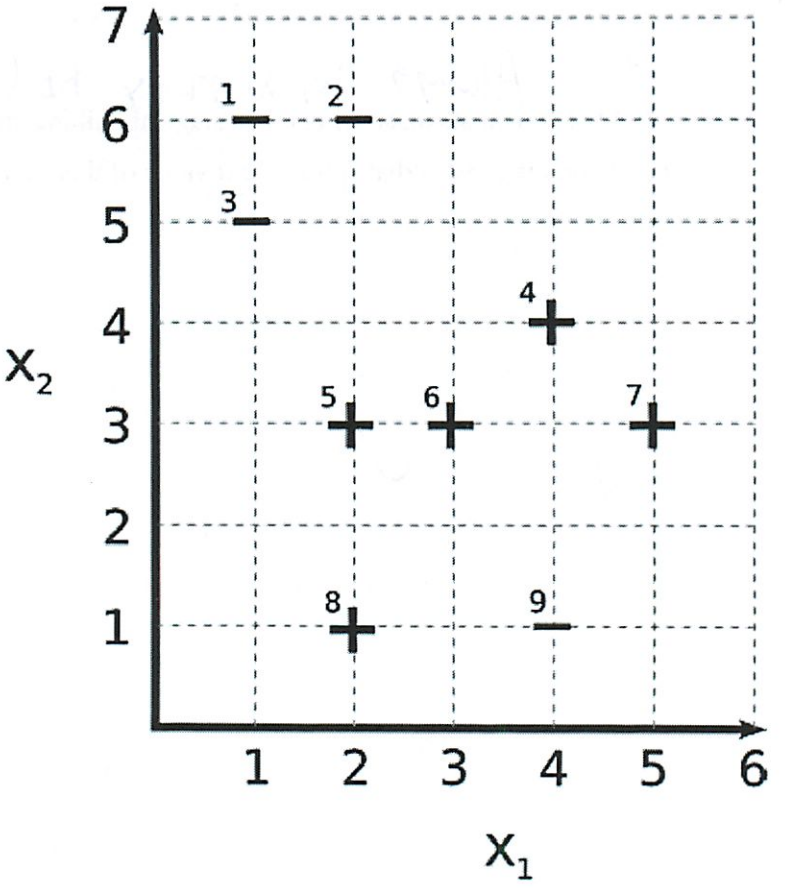
A3 (10 points): Suppose the F.R.O.S.T. scientists discover another magical, talking snowman that is five feet tall and has six coal buttons on his chest. What would the α value of this new data point be? Justify your answer.

0

Not a SV

Part B: Kernels (25 Points)

In this section, you will project the data below into a new space with $\phi(\mathbf{u}) = <| x_1 - x_2 |>$. That is, you project the two-dimensional vector **u** into a one-dimensional vector in a one-dimensional space.

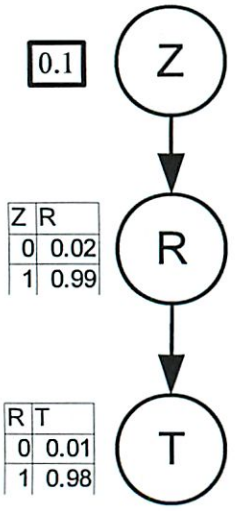


B1 (7 points): What is the kernel function K(u,v) for this transformation?

B3 (10 points): In the transformed space, what is the vector w and the constant b?

B2 (8 points): What is the final classifier produced. Express your answer in the original space, not the transformed space.

Quiz 4, Problem 2:
Bayesian Inference (50 points)



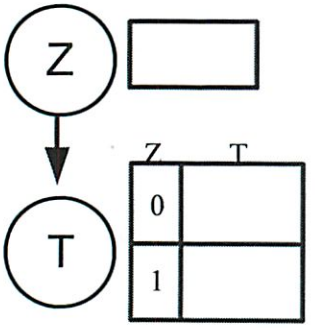
Zombies once again threaten the town, but this time the healthy folks are using science to help keep themselves on a brain-free diet!

When someone first becomes a zombie, their zombieness is latent for a while, as the transformation slowly progresses, but successful isolation at this stage is critical for avoiding new zombie infections.

New zombies quickly begin to produce a protein R that one can easily test for. One percent of infected individuals do not produce protein R (or perhaps just not yet), while protein R has been found in two percent of subjects who turned out not to have turned into zombies later,.

The test is 98% sensitive to the presence of the protein, and 99% specific to it, as shown in the Bayes Net on the left.

A (9 points) To make things a little simpler for the general population to understand, you want to give information in terms of how sensitive and specific the test is to zombiehood, not to the protein:



Show your work for partial credit:

B (6 points) Of 100 townfolk, how many will test positive:

C (6 points) ...and how many healthy ones will test positive:

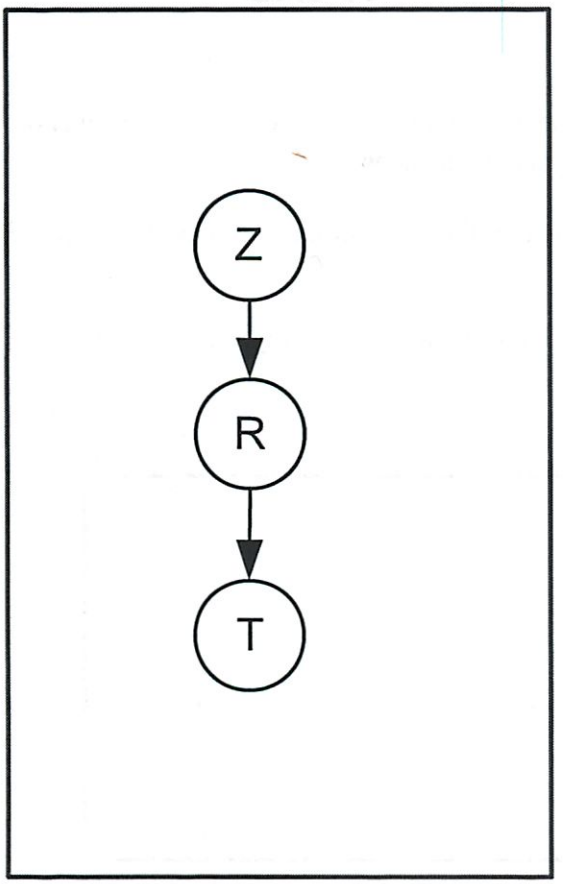
D (6 points) What is the overall accuracy of the test? (the probability that it predicts correctly)

E (5 points) A visitor, Ulric, was tested upon entering the town, and had a positive result. Ulric insists that he is healthy, and hasn't come anywhere near any zombies. To be sure, the guards decide to test him again. What assumption(s) are the guards making about the disease and the testing process?

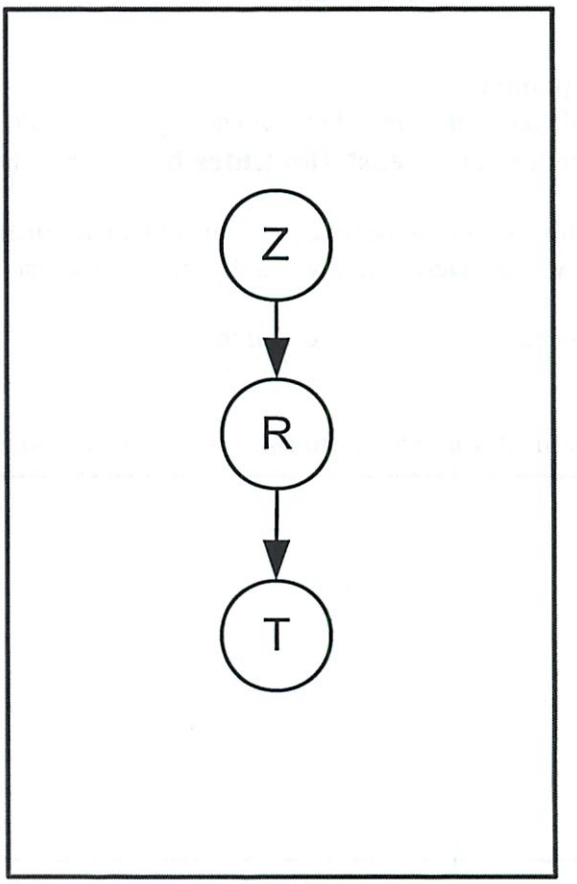
F (5 points) Victor and Wendy both arrive from Brainton, where Ulric came from too, and they both test positive, though they appear quite healthy. Ulric is doing fine in quarantine. The town scientists begin to consider two alternate hypotheses. Both of the alternate hypotheses allow being from Brainton (indicated with B) to explain away a positive test result. Both of the models extend the original three-node model with just the one extra node B. Draw the two possible models they should consider.

Draw only the new node B and its edges — you do not need conditional probability tables (yet):

Alternate Model A



Alternate Model B



Despite **two positive test results each**, Ulric, Victor, and Wendy all turn out to be healthy. **An expedition to (apparently zombie-free) Brainton revealed that 80% of its population test positive using your town's test.** One of the alternate models is very likely!

G (6 points)
Assume that you think a model whose Bayes Net representation has k+1 parameters is about **half** as likely *a priori* as a model with k parameters, unless k = 0, in which case assume that the model probability is some constant, c.

How likely is the original (three variable) model, *a priori*? [ ]

How likely are each of the models you drew, *a priori*?

Alternate Model A [ ]

Alternate Model B [ ]

H (7 points)
To rule out your original model entirely, you decide that an alternate model has to explain the data not just better, but **at least 100 times better** than the original model does.

Was the expedition necessary to rule out the original model, or did we know enough once we knew the three visitors were actually healthy, despite the tests?

Circle one: Needed the Expedition Ulric, Victor, and Wendy were enough

Explain, stating any assumptions you need to make:

Quiz 5, Problem 1, (40 points)

You have made it. You have graduated, started a thriving company, found a spouse, got a home in the suburbs, children. Now it is time to get a dog. Knowing nothing about dogs yourself, you ask a dog-loving, but inarticulate friend to characterize a series of dogs as Good or Bad. You hope to learn from your friend's characterizations.

Your first task is to pick some descriptors. You decide to look at intelligence, breed, exceptional characteristics, and gender. Male and female form a set. Yes and no form a set. Breeds are the leaves of a tree and form groups specified by the American Kennel Club: Bouvier and Collie are part of the Herding group, Beagle and Dachshund are part of the Hound group, and Chihuahua is part of the Toy group. Herding, Hound, Toy, Working, and Sporting all belong to the Recognized Dog category.

Using Arch learning, indicate in the table what is learned from each example and identify the heuristic involved by name, if known. If nothing is learned, put an x in the corresponding 2 columns.

Part A (32 points)

Candidate	Gender	Breed	Exceptional Quality	Intelligent	Heuristic	What is learned
Good	Male	Bouvier		Yes		
Bad	Female	Bouvier		Yes		
Bad	Male	Bouvier	Nasty	Yes		
Good	Male	Collie		Yes		
Good	Male	Collie		No		
Good	Male	Beagle		No		
Good	Male	Dachshund		No		
Bad	Female	Chihuahua		No		
Bad	Male	Chihuahua	Nasty	Yes		

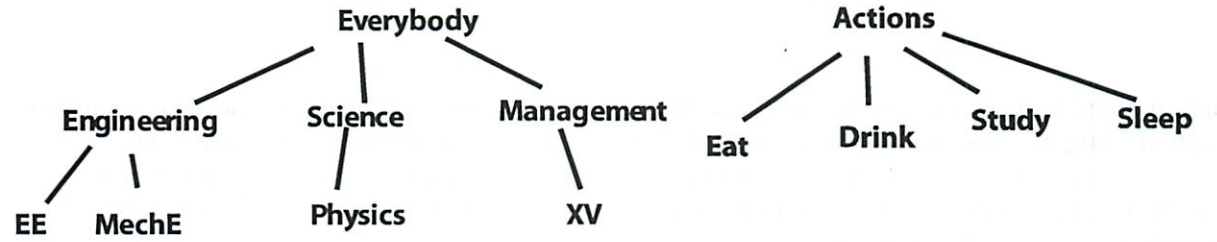
Part B (8 points)

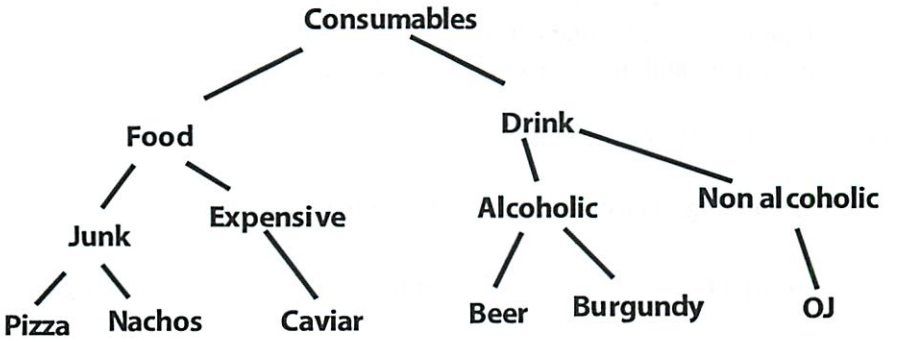
Exhibit an example, which if used as the second example, would teach two characteristics at once.

Candidate	Gender	Breed	Exceptional quality	Intelligent

Quiz 5, Problem 2, (21 points)

President Hockfield, eager to understand student dining preferences, decides to collect information in a self-organizing map. She starts by constructing some classification trees on a paper napkin. As you can see, she views students from the perspective of departments, such as EE, MechE, Physics, and XV:





Next, she develops an initial map with three relations:

Map element	Subject	Relation	Object
1	EE	Eats	Pizza
2	XV	Eats	Caviar
3	MechE	Drinks	Alcoholic beverage

Note that in this map, there is no concept of neighboring cell. She has decided to put that off for a while. Accordingly, each adjustment to the map alters only that cell judged closest to the new relation flowing into the map.

Part A

Now, President Hockfield observes the following:

Subject	Relation	Object
MechE	Eats	Nachos

Not fully remembering Professor Winston's lecture, she asks you to adjust the map to accommodate this information using the classification trees and edit distance (number of up and down steps in the classification tree needed to go from one combination to the other) to determine which element to change. You don't quite remember how to change the selected element, but fortunately, you took a photograph of Professor Winston's practice board:

If sample object classification extends map object classification
Then extend map classification object by one class
Else if sample object classification and map object classification are the same
Then do nothing
Else trim map object classification by one class

To save time, do not bother copying in any information that remains unchanged.

Map element	Subject	Relation	Object
1			
2			
3			

Part B

Repeat with the following observation. Note that any change you made in Part A is to be carried into this part:

Subject	Relation	Object
EE	Drinks	Beer

Again, you need only enter the changed row. You need not write in any row that remains the same after the work you did in Part A.

Map element	Subject	Relation	Object
1			
2			
3			

Part C

Repeat with the following observation. Note that any changes you made in previous parts are to be carried into this part:

Subject	Relation	Object
Management	Drinks	Burgundy

Again, you need only enter the changed row. You need not write in any row that remains the same after the work you did in previous parts.

Map element	Subject	Relation	Object
1			
2			
3			

Quiz 5, Problem 3, (15 points)

A: A Japanese Haiku is a poem that consists of 17 syllables arranged in 3 lines of 5, 7, and 5 syllables. These are very popular, so a publishing house has asked you to develop a system that can generate good ones automatically.

The syllables are drawn from the approximately 50 syllables that are found in Japanese.

James, who dropped 6.034 early in the term suggests that you just generate all possible Haikus and hand them to a panel of judges who will pick the best.

Part A (12 points)

Determine how long it would take to calculate all possible Haikus given a computer that produces one Haiku per nanosecond. So you won't think you need a calculator, you may make the following approximations:

$50 = 32$, $17 = 16$, and $1024 = 1000$, and seconds/year $= 10^7$

Part B (3 points)

Is James suggestion practical? Circle your answer

Yes No

Quiz 5, Problem 4, (24 points)

Circle the best answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

Genetic algorithms, without crossover, is best described as a kind of

1. Instance of General Problem Solver architecture
2. Instance of subsumption architecture
3. Instance of SOAR architecture
4. Hill climbing search
5. None of the above

Crossover is best described as

1. A product of means-ends analysis
2. A product of an abstraction barrier
3. A label for the strange mating behavior of Zebra Finches
4. A means to escape local maxima in a search space
5. None of the above

The SOAR architecture is best described as

1. A programming language
2. A descendant of the General Problem Solver architecture
3. An amalgam of several ideas
4. The design philosophy that led to the Stata center
5. None of the above

The Genesis architecture (Winston's research focus) is best described as

1. A search for a universal representation
2. A commitment to reasoning as the distinguishing feature of human intelligence
3. A search for principles that explain the intelligence of non human primates
4. A demonstration that natural language has little or no role in explaining human intelligence
5. None of the above

Minsky's Emotion Machine/Society of Mind architecture is best described as

1. Focused on explaining visual problem solving
2. Focused on reasoning on multiple levels
3. A commitment to the idea that language is the differentiating feature of human intelligence
4. A commitment to the idea that culture is reflected in the myths associated with the culture
5. None of the above

Intermediate features and the Goldilocks principle is best explained as

1. The observation that eyes are too small and faces are too large
2. The observation that trajectories are too small and transitions are too large
3. The observation that cultures are defined by their fairy tales
4. The observation that the intelligence of a songbird lies half way between insects and humans
5. None of the above

Tear off sheets—you need not hand in the tear off sheets

Q1P1

RULES:

R0 : IF (?X) goes to MIT,
THEN (?X) is a muggle,
(?X) consumed butterbeer

R1: IF (?X) made math jokes AND
(?X) consumed butterbeer
THEN (?X) was transfigured into a porcupine

R2: IF (?Y) fancies (?X) AND
(?X) fancies (?Y) AND
(?Y) is a muggle
THEN (?X) snogged (?Y)

R3: IF (?X) fancies (?Y) AND
(?X) made math jokes,
THEN (?Y) fancies (?X)

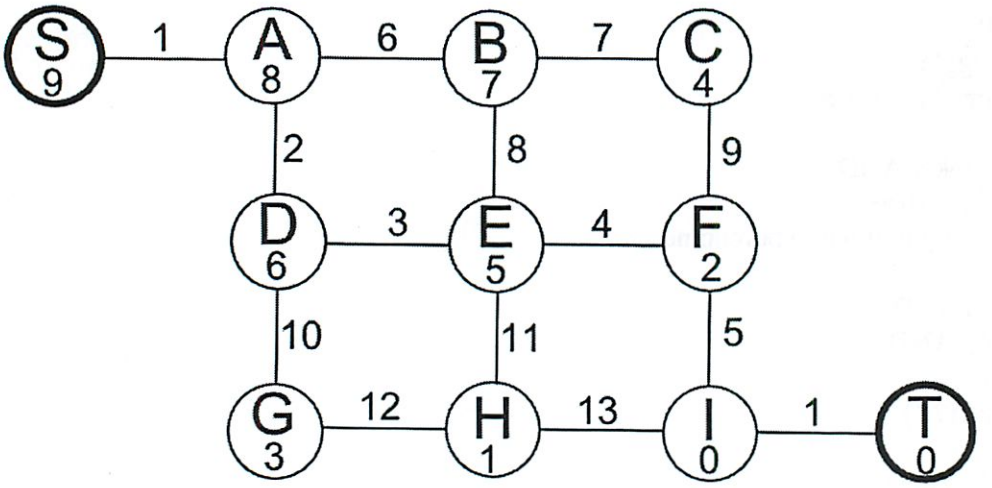
R4: IF (?X) made math jokes
THEN (?X) goes to MIT

You start with the following list of assertions which **is all you have to go on.**

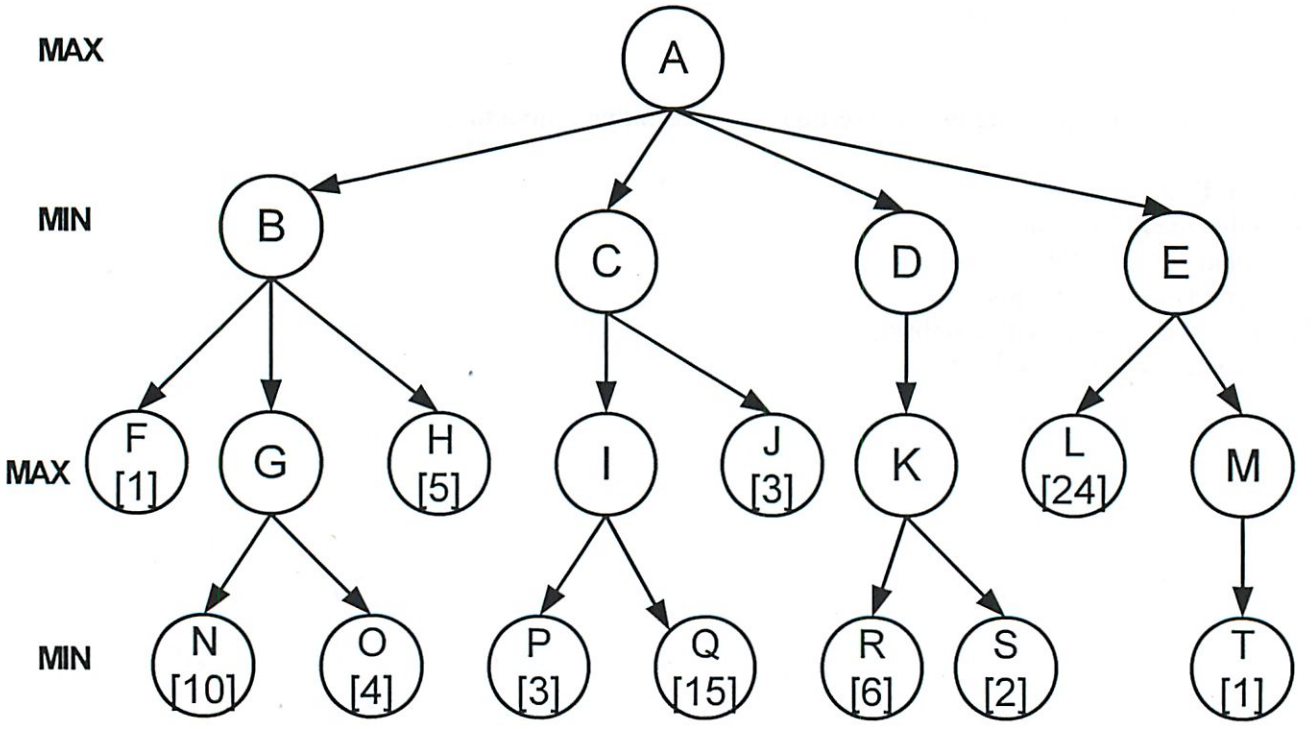
ASSERTIONS:

A0: Olga made math jokes
A1: Yuan goes to MIT
A2: Jeremy made math jokes
A3: Hermione consumed butterbeer
A4: Jeremy fancies Hermione

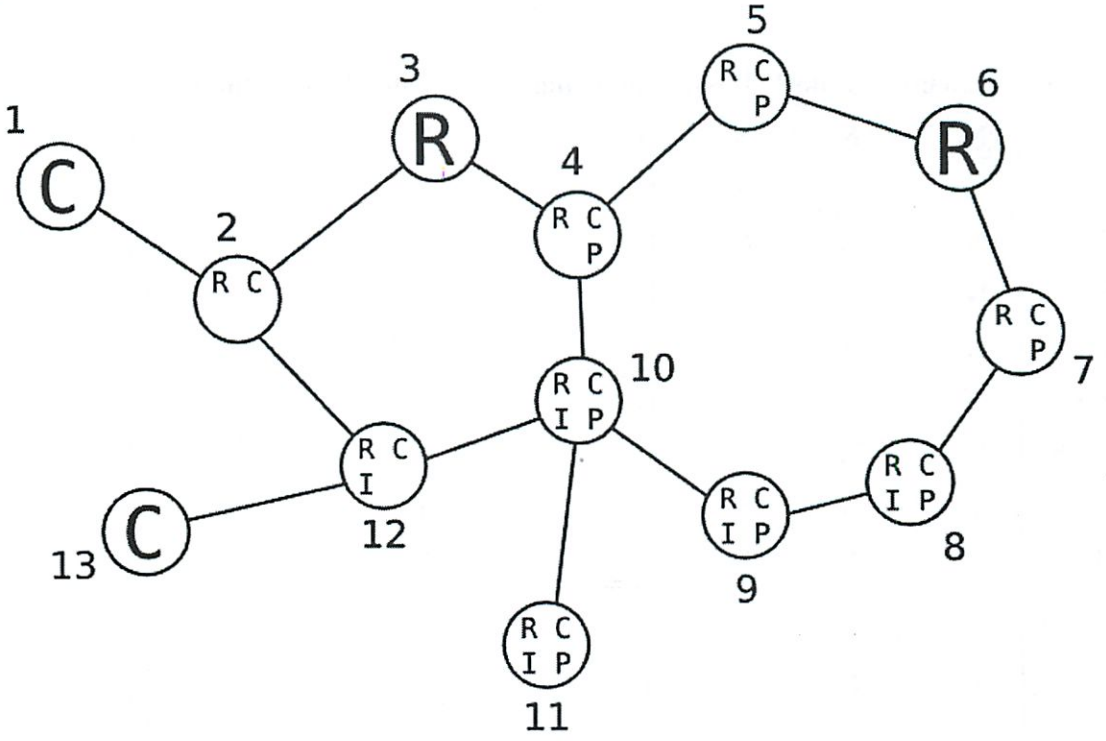
Q1P2


S
9
1
A
8
6
B
7
7
C
4
2
8
9
D
6
3
E
5
4
F
2
10
11
5
G
3
12
H
1
13
I
0
1
T
0

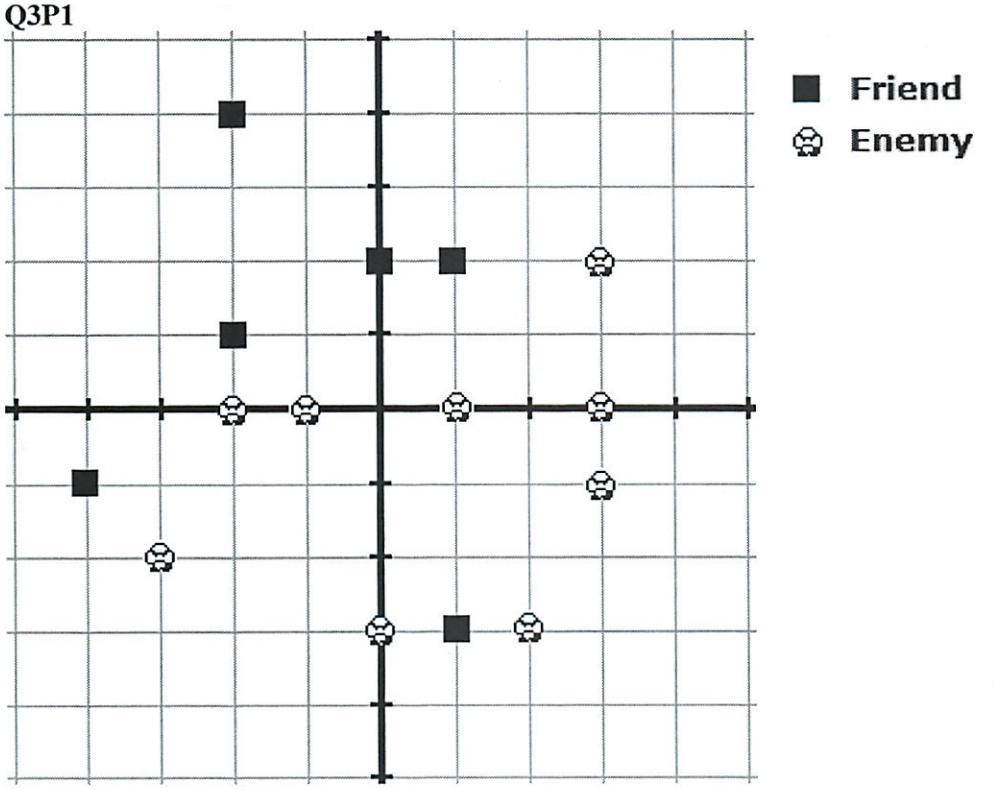
Q2P1


MAX
A
MIN
B
C
D
E
MAX
F
[1]
G
H
[5]
I
J
[3]
K
L
[24]
M
MIN
N
[10]
O
[4]
P
[3]
Q
[15]
R
[6]
S
[2]
T
[1]

Q2P2


1
C
2
R C
3
R
4
R C P
5
R C P
6
R
7
R C P
8
R C I P
9
R C I P
10
R C I P
11
R C I P
12
R C I
13
C

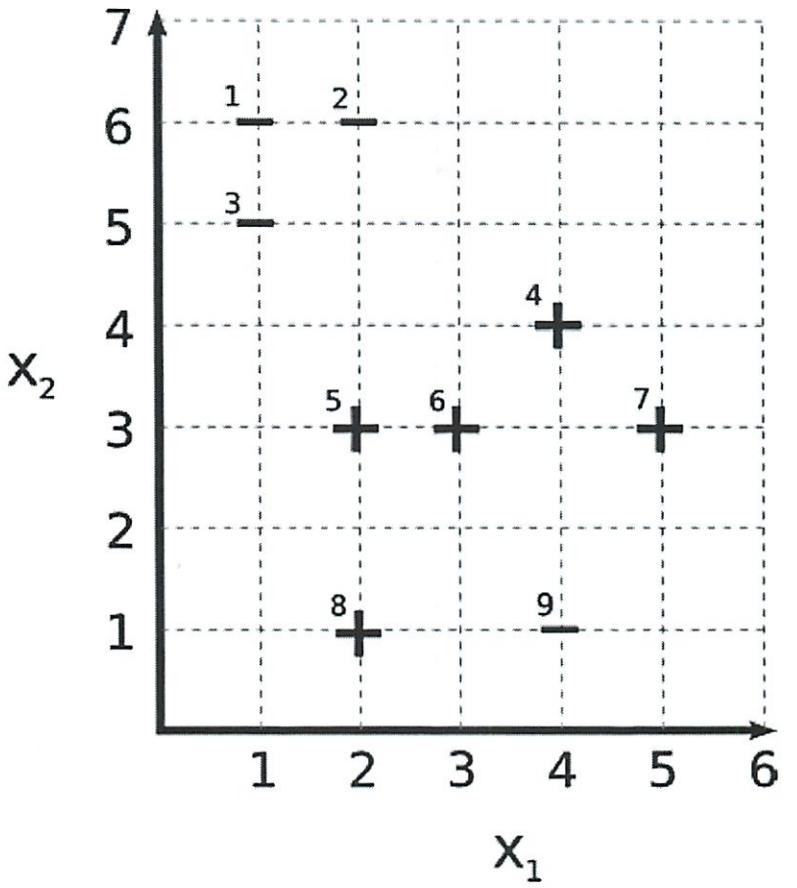
Q3P1


Friend
Enemy

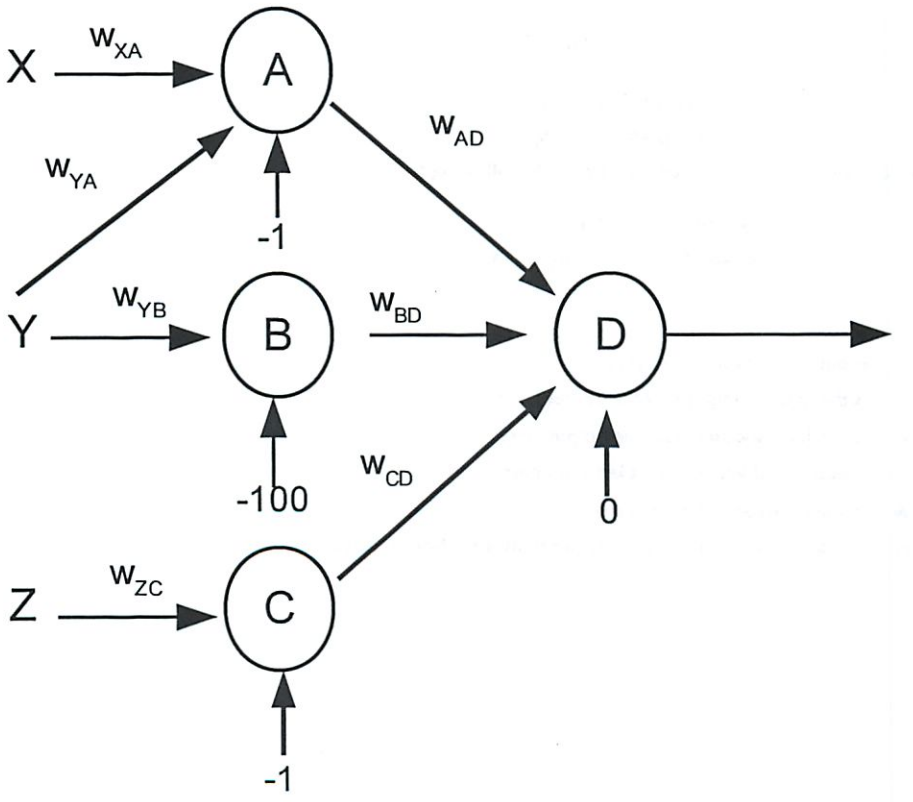
Q3P1

	Character	Enemy?	Talks?	Boss?	Annoying?	Killable with Jump?
A	Yoshi	N	N	N	N	N
B	Tree	N	N	N	N	N
C	Luigi	N	N	N	N	N
D	Toad	N	Y	N	N	N
E	Peach	N	Y	N	Y	N
F	Bowser	Y	N	Y	N	N
G	Bob-omb	Y	N	N	N	Y
H	Goomba	Y	N	N	N	Y
I	Piranha	Y	N	N	Y	N
J	Dry Bones	Y	N	N	Y	N
K	Chain Chomp	Y	N	N	Y	N
L	Thwomp	Y	N	N	Y	N
M	Boo	Y	N	N	Y	N
N	Koopa Troopa	Y	N	N	Y	Y

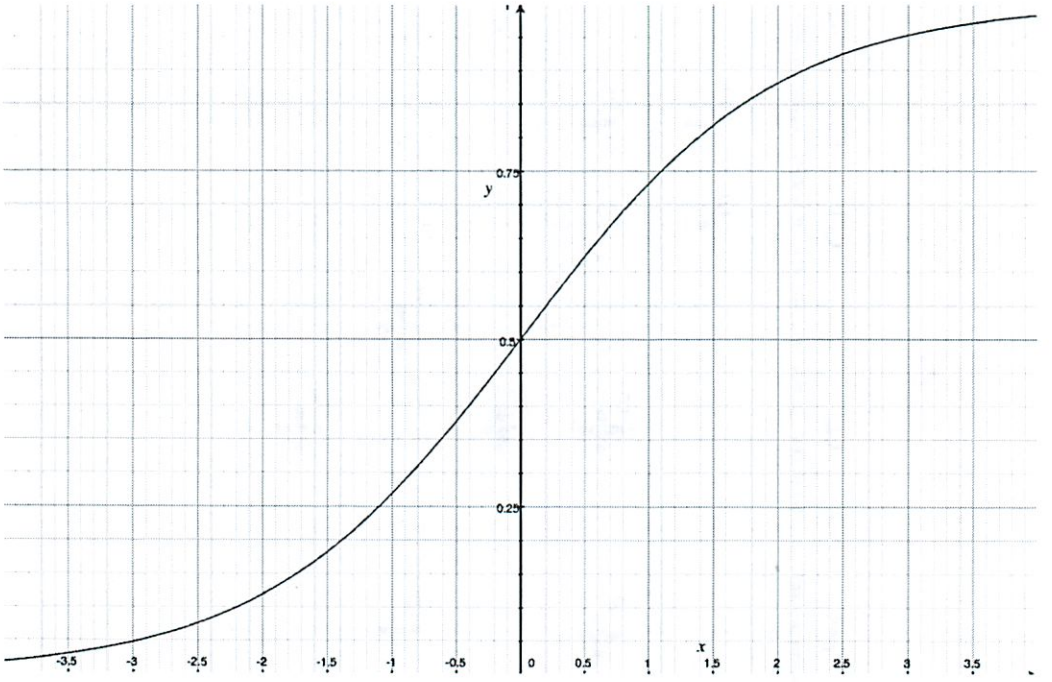
Q4P1


7
6
5
4
3
2
1
X_2
1
2
3
4
5
6
7
8
9
1 2 3 4 5 6
X_1

Q3P2


X
Y
Z
W_{XA}
W_{YA}
W_{YB}
W_{ZC}
W_{AD}
W_{BD}
W_{CD}
A
B
C
D
-1
-100
-1
0

Q3P2



Neural net update:

$$E = \frac{1}{2}\sum_k (o_k - d_k)^2$$

$$w_{i \to j} = w_{i \to j} - \Delta w_{i \to j}$$

$$\Delta w_{i \to j} = R \times o_{l_i} \times \delta_{r_j}$$

where R is a rate constant and the δs are computed with the following formulas:

$$\delta_k = o_k(1 - o_k) \times (o_k - d_k)$$

$$\delta_{l_i} = o_{l_i}(1 - o_{l_i}) \times \sum_j w_{i \to j} \times \delta_{r_j}$$

where

- o_k is output k of the output layer
- d_k is the desired output k of the output layer
- δ_k is a delta associated with the output layer
- o_{l_i} is output i of left layer in a left-right pair
- δ_{l_i} is a delta associated with the layer l
- δ_{r_j} is a delta associated with the adjacent layer to the right, layer r

Name	Albert Einstein
email	einstein@heaven.org

6.034 Final Examination
December 16, 2009

Circle your TA and principle recitation instructor so that we can more easily identify with whom you have studied:

Erica Cooper	Matthew Peairs	Mark Seifter
Yuan Shen	Jeremy Smith	Olga Wichrowska

Robert Berwick	Randall Davis	Gregory Marton

Indicate the approximate percent of the lectures, mega recitations, recitations, and tutorials you have attended so that we can better gauge their correlation with quiz and final performance. Your answers have no effect on your grade.

	Lectures	Recitations	Megas	Tutorials
Percent attended				

Quiz	Score	Grader
Q1		
Q2		
Q3		
Q4		
Q5		

There are 38 pages in this final examination, including this one. In addition, tear-off sheets are provided at the end with duplicate drawings and data. As always, open book, open notes, open just about everything.

Quiz 1, Problem 1, Rules (50 points)

The administration, worried about the social habits of its students, agrees to finance cross-school-mixers. The 034 TA's decide to fly to England and mix with the students at Hogwarts School of Witchcraft and Wizardry. A merry old time ensues, but the morning after, due to an accidental confundo charm (and perhaps also a large consumption of butterbeer), no one can remember the events that transpired. The 034 staff, in an attempt to show off the power of Muggle logic, promise they can piece together the important events with a rule based system.

Using their keen sense of logic, Matt, Erica, and Mark are able to piece together the following rules:

RULES:
R0 : IF (?X) goes to MIT,
THEN (?X) is a muggle,
(?X) consumed butterbeer

R1: IF (?X) made math jokes AND
(?X) consumed butterbeer
THEN (?X) was transfigured into a porcupine

R2: IF (?Y) fancies (?X) AND
(?X) fancies (?Y) AND
(?Y) is a muggle
THEN (?X) snogged (?Y)

R3: IF (?X) fancies (?Y) AND
(?X) made math jokes,
THEN (?Y) fancies (?X)

R4: IF (?X) made math jokes
THEN (?X) goes to MIT

You start with the following list of assertions which is all you have to go on.

ASSERTIONS:
A0: Olga made math jokes
A1: Yuan goes to MIT
A2: Jeremy made math jokes
A3: Hermione consumed butterbeer
A4: Jeremy fancies Hermione

ALBUS DUMBLEDORE

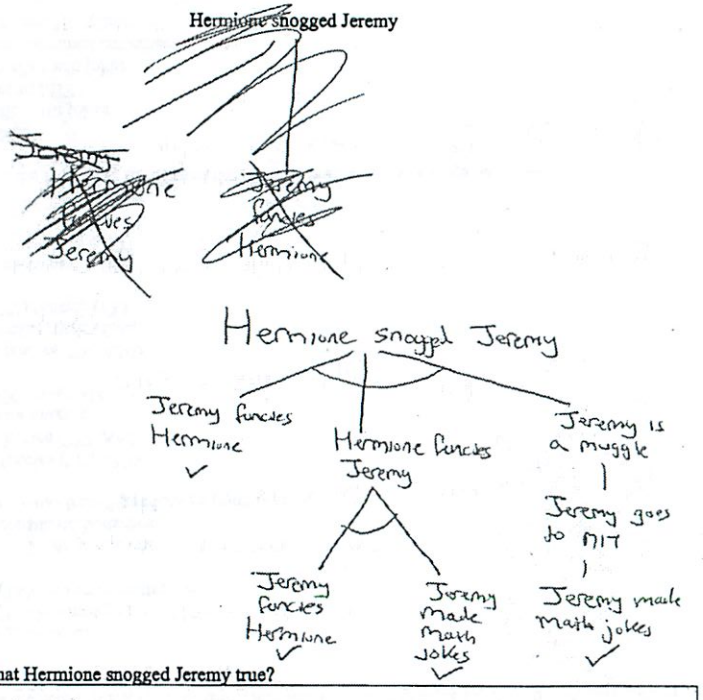
Part A: Forward Chaining (24 points)

Run forward chaining on the rules and assertions provided for the first 5 iterations. For the first two iterations, fill out the first two rows in the table below, noting the rules whose antecedents match the data, the rule that fires, and the new assertions that are added by the rule. For the remainder, supply only the fired rules and new assertions. As usual, break ties using the earliest rule on the list that matches. If the earliest rule matches more than once, break ties by assertion order.

	Matched	Fired	New assertions added to database
1	R0, R3, R4	R0	-Yuan is a muggle -Yuan consumed butterbeer
2	R0, R3, R4	R3	-Hermione fancies Jeremy (DUH)
3	[redacted]	R4	-Olga goes to MIT
4	[redacted]	R0	-Olga is a muggle -Olga consumed butterbeer
5	[redacted]	R1	-Olga was transfigured into a Porcupine (lol)

Part B: Backward Chaining (26 points)

Ron Weasley claims that Hermione snogged Jeremy. Use backward chaining to determine if this event occurred. **Draw the goal tree for this statement.** Partial credit will be given for partial completion of the goal tree.

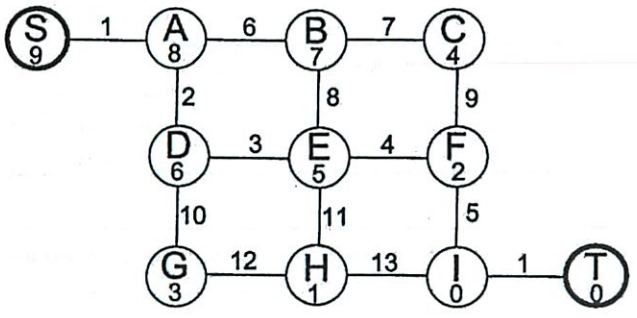


Is the claim that Hermione snogged Jeremy true?

HELL YES!

WOOHOO!

Quiz 1, Problem 2: Search (50 points)



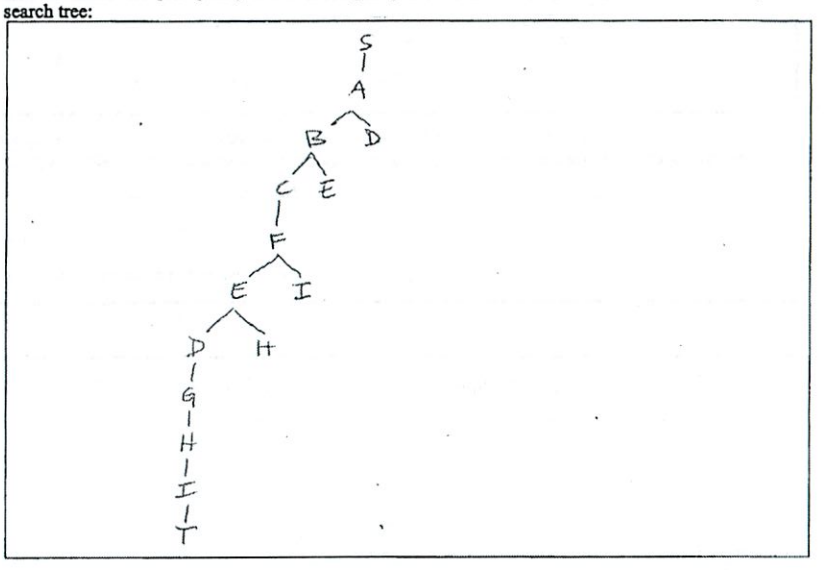
Aliens have invaded the MIT campus, thinking that the Stata Center was their mothership. Despite their realization that it was not, they went ahead with their invasion plan anyway. Advanced alien technology has disabled use of phones and internet, so you and your friends decide to go across the river to Boston to get help and to tell everyone what's going on. You decide to use the secret underground network of tunnels underneath the campus to get from your starting location at **S**, to the subway station at **T**.

You have the above graph of the underground tunnels. The edges are labeled with distances, and the nodes are labeled with a heuristic estimate to your destination at **T**. When performing search, ties are broken by choosing the node that is **alphabetically first.**

Note that node G is not the goal node in this problem.

Part A: Depth-First Search. (15 points)

A1: You first attempt to pick your route using **Depth-first search with an extended list.** Draw your search tree:



Also, show your extended list at the time your search terminates:

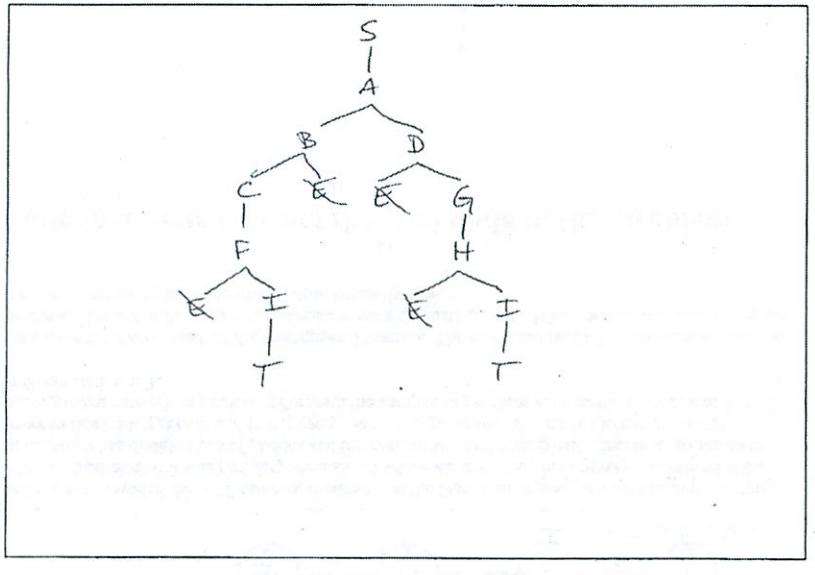
SABCFEDGHIT

A2: What is the final path that is found using depth-first search?

SABCFEDGHIT

Part B: Beam Search (15 Points)

B1: Next, you try to pick your route using **Beam Search with a beam width of 2, using an extended list.** In the event of a tie, use the alphabetical order of the final node on the contending partial paths. Draw your search tree:



Also, show your extended list at the time your search terminates:

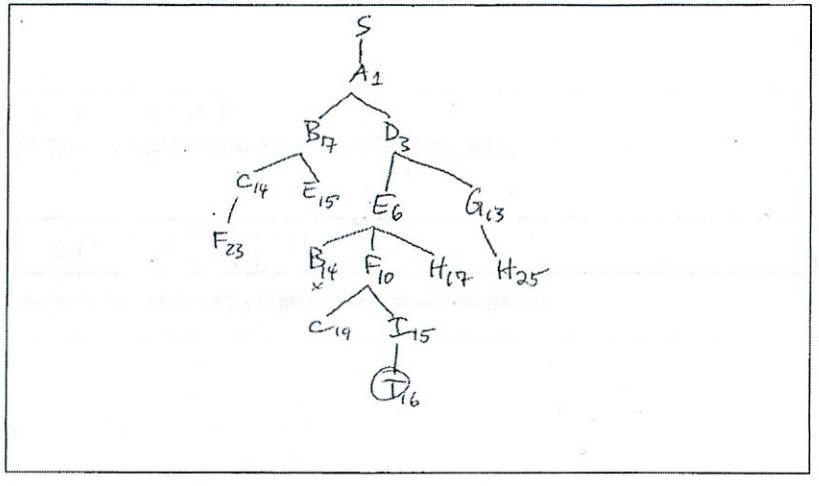
SABDCGFHIT

B2: What is the route found using beam search?

SABCFIT —OR— SADGHIT

Part C: Branch and Bound (20 Points)

C1: Lastly, you try to plan your route using **Branch and Bound using the path lengths indicated on the graph and also using an extended list.** Draw your search tree and extended list below:



Also, show your extended list at the time your search terminates:

SADEBFGCIT

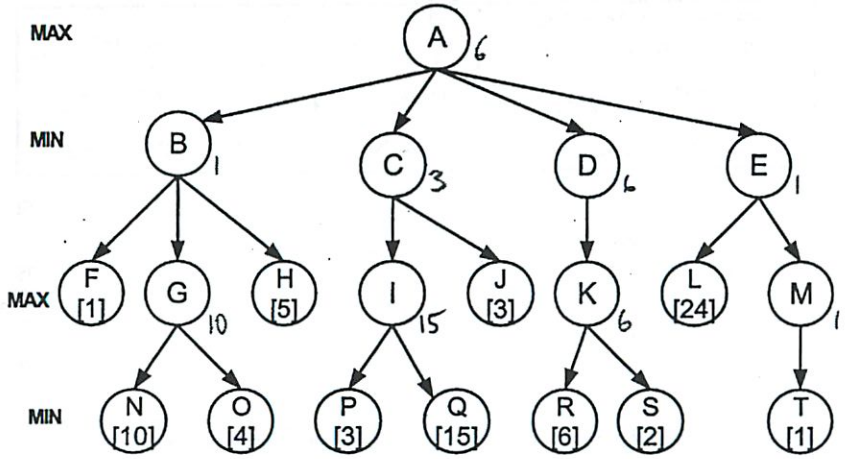
C2: What is the route found using branch and bound?

SADEFIT

C3: If you repeat the search using the **A* algorithm with the heuristic values indicated on the graph,** will you find the same route? Include a brief explanation of why or why not.

yes — the heuristic is consistent.

Quiz 2 Problem 1: Games (50 points)

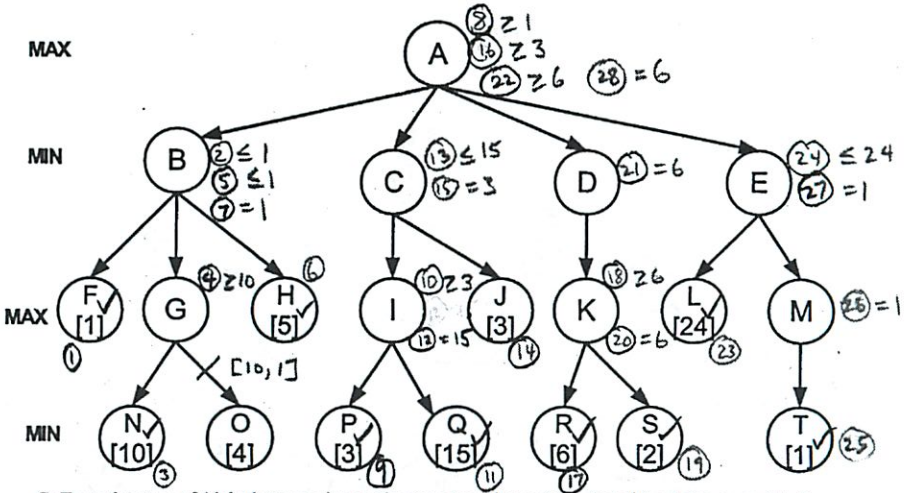


A: Using minimax only, no alpha-beta,, indicate the values of the following nodes. (10 pts)

A = 6	G = 10
B = 1	I = 15
C = 3	K = 6
D = 6	M = 1
E = 1	

B: Using minimax only, what is the best next move from A? **(Indicate a letter)** (4 pts)

D



C: Trace the steps of Alpha beta pruning on the same tree above. Note that alpha, betas are updated before pruning occurs, if in doubt consult the reference implementation given on the tear-off sheet.
List the leaf nodes in the order that they are statically evaluated. (10 pts)

F N H P Q J R S L T

(only O is pruned!)

What are the final Alpha Beta values at node E (4pts)

Alpha = 6	Beta = 1

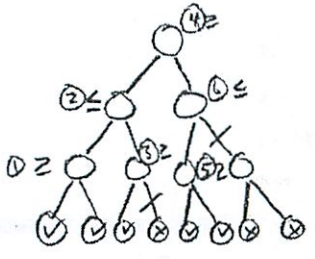
What are the final Alpha Beta values at node A (4pts)

Alpha = 6	Beta = ∞

D: For a full **binary tree** (branching factor = 2) **of depth 3** (4 layers of nodes including the layer with the root node, 8 nodes in the bottom layer), with the root node being a MAX node, what is the fraction of leaf nodes that are statically evaluated under alpha beta pruning under conditions of maximum pruning? (11 pts)

$\frac{5}{8}$

Use the space below to show your work:



E: For a given depth, d, does the fraction requested in part D increase, decrease, or stay the same with increasing b. Circle your answer (7 pts)

Increase

(Decrease)

Stay the same

Explain:

# of Nodes visited for $\alpha\beta$ pruning is $O(b^{d/2})$

# of Leaves for a tree of branching factor b, depth d = b^d

$\therefore$ Fraction is $O\left(\frac{b^{d/2}}{b^d}\right) = O\left(\frac{1}{b^{d/2}}\right)$ So as b increases b^d the fraction decreases!

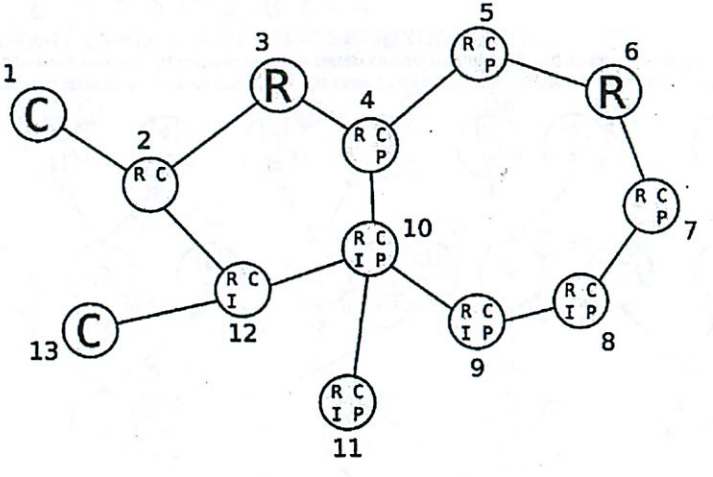
Quiz 2, Problem 2:
Constraint Propagation (50 Points)

Important note: there is no search in this problem, only constraint propagation.

On the newly colonized planet Mars, cities consist of pressurized domes connected by a series of tubes. Each dome is designated for a specific type of use, and there are some restrictions on what sorts of domes can be connected to each other:

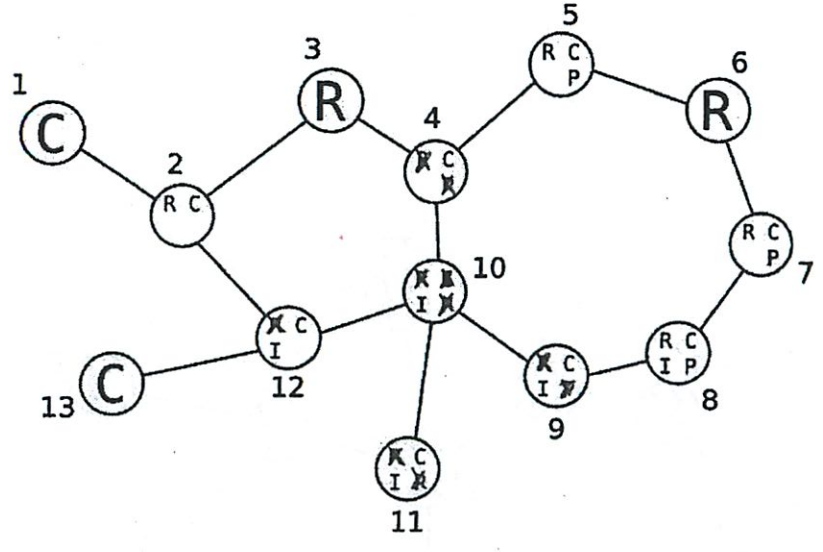
1. A Residential dome can only be connected to another Residential dome, a Commercial dome, or a Park.
2. A Commercial dome can only be connected to a Residential dome, another Commercial dome, or an Industrial dome.
3. An Industrial dome can only be connected to a Commercial dome or another Industrial dome.
4. A Park can only be connected to a Residential dome.

You decide to use constraint propagation to design a new Martian city. You begin with the following partially completed plan, with some domes already designated and their neighboring domains reduced accordingly:



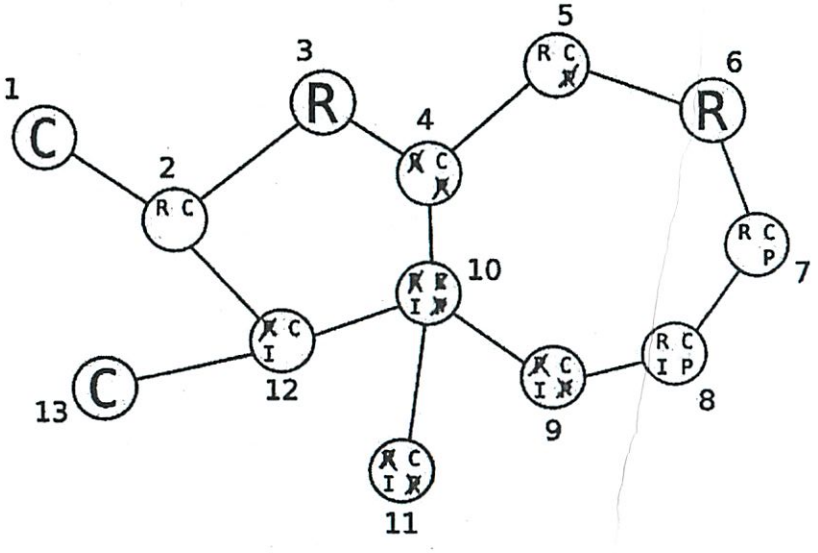
Part A (12 points)

You decide to begin by designating Dome #10 as Industrial. Using forward checking (no propagation beyond the neighbors of the just-assigned variable), show how the domains of the city's domes are reduced by crossing out the appropriate letters in the map below:



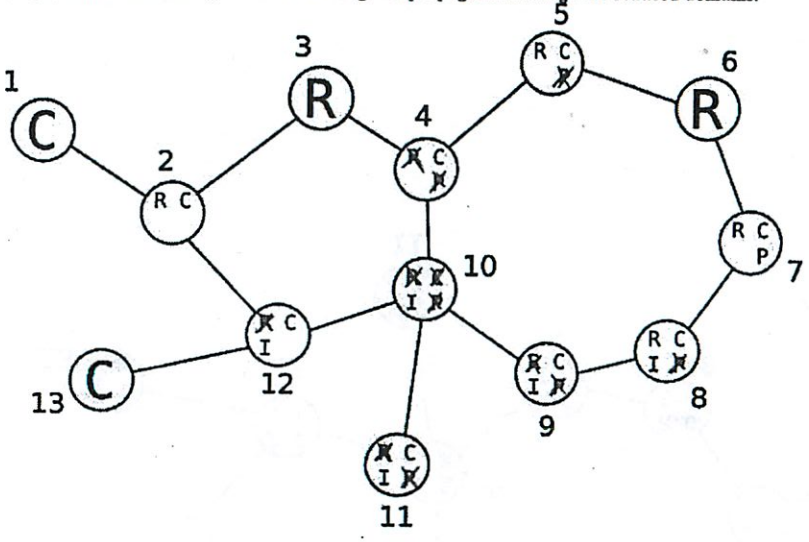
Part B (12 points)

Next, show how the domains are reduced if you designate Dome #10 as Industrial, and then propagate constraints using **forward checking and propagation through singleton domains.**



Part C (12 points)

Finally, show how the domains are reduced if you designate Dome #10 as Industrial, and then propagate constraints using **forward checking and propagation through all reduced domains.**



Part D (14 points)

Yuan says that you are right to pick a dome type for Dome #10 first, but Olga advises you to start with Dome #2 instead. What is the reasoning behind each of these suggestions?

Yuan: Dome #10 has the most constraints.

Olga: Dome #2 has the smallest domain.

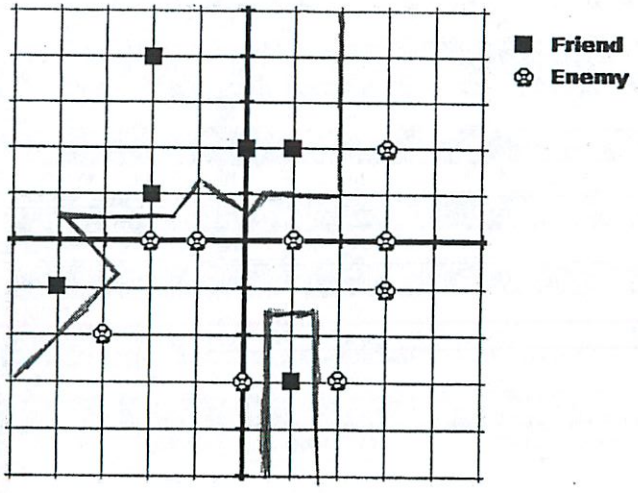
Quiz 3, Problem 1:
KNN & ID-Trees (50 points)

The Mario Bros. have hired you to help them identify potential dangers during their frequent expeditions to explore strange new worlds, battle evil monsters, and attempt to rescue some princesses.

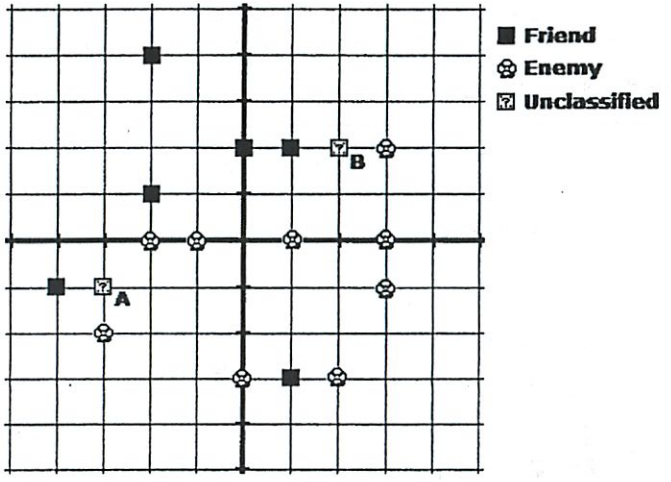
Part A: Nearest Neighbors (20 Points)

You decide to use Nearest Neighbors to classify new creatures that Mario might encounter, based on their positions on a map. In the map, friendly creatures are marked with a square, and enemies are marked with monster icons.

A1 (12 points): On the following graph, draw the decision boundaries produced by 1-nearest-neighbors.



A2 (8 points): The graph below shows two new creatures, marked with a question mark symbolsand labeled A and B. Show how these will be classified using 3-nearest-neighbors and 5-nearest-neighbors below.



	Creature A	Creature B
Using 3 NN:	ENEMY	FRIEND
Using 5 NN:	ENEMY	ENEMY

Part B: ID-Trees (30 Points)

It turns out that enemies move around the world, so a simple K-Nearest-Neighbors on their location won't do a very good job of protecting Mario. Instead, you decide to use an Identification Tree, based on some characteristics of the creatures in this world, to classify creatures as Enemies (Enemy=Y) and Friends (Enemy=N). The other 5 characteristics you note are in the table below.

	Character	Enemy?	Talks?	Boss?	Annoying?	Killable with Jump?
A	Yoshi	N	N	N	N	N
B	Tree	N	N	N	N	N
C	Luigi	N	N	N	N	N
D	Toad	N	Y	N	N	N
E	Peach	N	Y	N	Y	N
F	Bowser	Y	N	Y	N	N
G	Bob-omb	Y	N	N	N	Y
H	Goomba	Y	N	N	N	Y
I	Piranha	Y	N	N	Y	N
J	Dry Bones	Y	N	N	Y	N
K	Chain Chomp	Y	N	N	Y	N
L	Thwomp	Y	N	N	Y	N
M	Boo	Y	N	N	Y	N
N	Koopa Troopa	Y	N	N	Y	Y

B1 (8 points): What is the disorder of the test **"Boss = Y"**? Leave your answer in terms of fractions, real numbers, and logarithms.

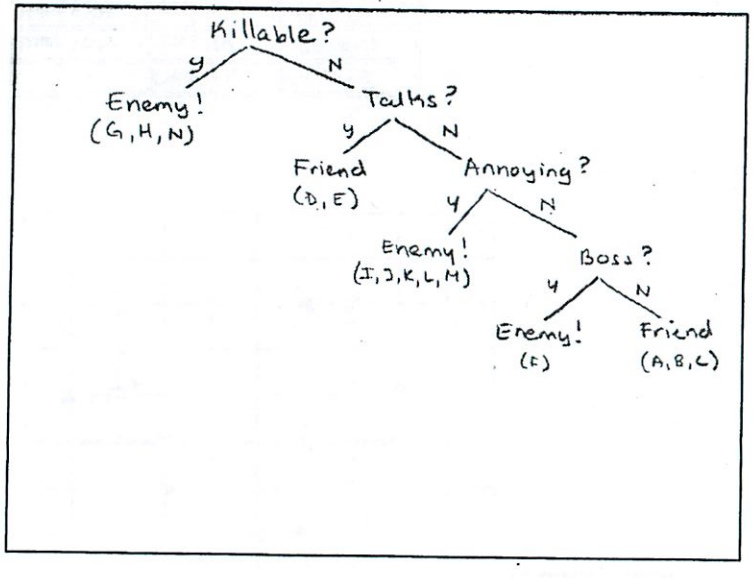
$$\frac{13}{14}\left(\frac{-8}{13}\lg\frac{8}{13}-\frac{5}{13}\lg\frac{5}{13}\right)+0=-\frac{4}{7}\lg\frac{8}{13}-\frac{5}{14}\lg\frac{5}{13}$$

B2 (7 points): What is the disorder of the test **"Annoying = Y"**? Leave your answer in terms of fractions, real numbers, and logarithms.

$$\frac{7}{14}\left(\frac{-1}{7}\lg\frac{1}{7}-\frac{6}{7}\lg\frac{6}{7}\right)+\frac{7}{14}\left(-\frac{3}{7}\lg\frac{3}{7}-\frac{4}{7}\lg\frac{4}{7}\right)=\frac{-1}{14}\lg\frac{1}{7}-\frac{3}{7}\lg\frac{6}{7}-\frac{3}{14}\lg\frac{3}{7}-\frac{2}{7}\lg\frac{4}{7}$$

B3 (15 points): Draw the **disorder minimizing** identification tree Mario can use to correctly decide whether any of the above examples is an enemy (Enemy=Y) or friend (Enemy=N). Use the letters provided next to characters' names to show how the characters are separated by each decision.

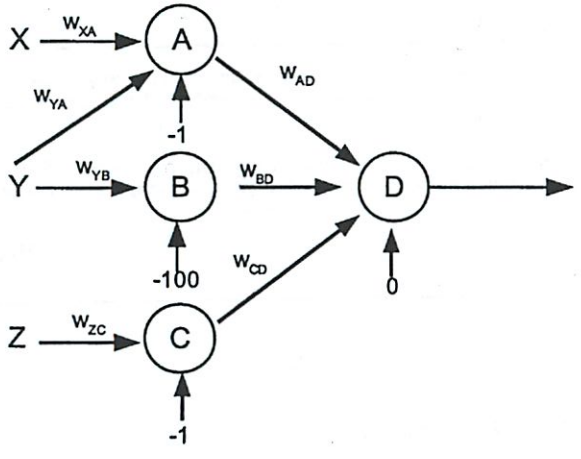
Hint: neither Boss nor Annoying are the first test.



Quiz 3, Problem 2:
Neural Nets (50 points)

Hermione Granger decides to spend a semester abroad at MIT for her Muggle Studies class. Upon arriving, Hermione befriends the coolest people on campus. The 034 staff, over a festive round of Butter Beer, then teach Hermione about neural nets. She decides to practice.

Note that the tear off sheets include the neural net tear-off sheet from Quiz 3.



Hermione creates the above neural net. Each node uses a standard sigmoid function. All threshold units have fixed weights of 1. **They are never updated.**

Part A: (12 Points)

For each node (A, B, C, and D), using only the weight variables provided in the diagram, d (desired output), and the outputs of each node (O_A, O_B, O_C, and O_D), determine δ_A, δ_B, δ_C, and δ_D.

$\delta_D = O_D(1-O_D)(d-O_D)$

$\delta_A = O_A(1-O_A)w_{AD}\delta_D$

$\delta_B = O_B(1-O_B)w_{BD}\delta_D$

$\delta_C = O_C(1-O_C)w_{CD}\delta_D$

Part B: (12 Points)

Hermione, as an intellectual exercise, runs the neural net with the following parameters:

X=0.5	Y = 0	Z = 5	W_{XA} = 2
W_{YA} = 2	W_{YB} = 1	W_{ZC} = 0.2	W_{AD} = 1
W_{BD} = 1.5	W_{CD} = -1		

Determine the output for each node (O_A, O_B, O_C, and O_D). Approximate sigmoid(x) very simply by letting it evaluate to 0 for x < -50 and 1 for x > 50.

$O_A = S(0.5(2) + 0(1) - 1) = S(0) = 1/2$

$O_B = S((0)(1) - 100) = S(-100) = 0$

$O_C = S(5(0.2) - 1) = S(0) = 1/2$

$O_D = S(1/2(1) + 0(1.5) - (1)(1/2)) = S(0) = 1/2$

Part C: (18 Points)

Run back-propagation using these same parameter values for one iteration to determine new weight values for W_{AD}, W_{BD}, and W_{CD}. Use your results from Part A and Part B, and assume a learning rate of r = 2 and a desired output, d=1.

$w'_{AD} = w_{AD} + r\,\delta_D\,i_{AD}$

$= 1 + 2(1/2 \cdot 1/2 \cdot 1/2) \cdot 1/2 = 9/8$

$w'_{BD} = w_{BD} + r\,\delta_D\,i_{BD} = 3/2$

$w'_{CD} = w_{CD} + r\,\delta_D\,i_{CD} = -7/8$

Part D: (8 Points)

Hermione changes **all the decision functions in the original network, before training,** from using a sigmoid function to using a **threshold function** (which outputs 1 if the input is greater than 0, and outputs 0 otherwise).

Using X, Y, and Z and the standard logical operators (AND, NOT, and OR), determine a logic expression for the neural net. Assume that the inputs are always 0 (meaning false) or 1 (meaning true).

C/Z is irrelevant b/c $O_C = 0$ for both $Z=1$ & $Z=0$

$O_C = \begin{cases} 1 & \text{if } 0.7Z - 1 \geq 0 \rightarrow Z \geq 5 \\ 0 & \text{o/w} \end{cases}$

B is irrelevant for same reason

$O_B = \begin{cases} 1 & \text{if } Y - 100 \geq 0 \rightarrow Y \geq 100 \\ 0 & \text{o/w} \end{cases}$

A has X & Y

$O_C = \begin{cases} 1 & \text{if } 2X + 2Y - 1 \geq 0 \\ 0 & \text{o/w} \end{cases}$ $\Big\}$ $O_C = X$ or Y

$O_D = \begin{cases} 1 & \text{if } O_A + 1.5O_B - O_C \geq 0 \\ 0 & \text{o/w} \end{cases}$

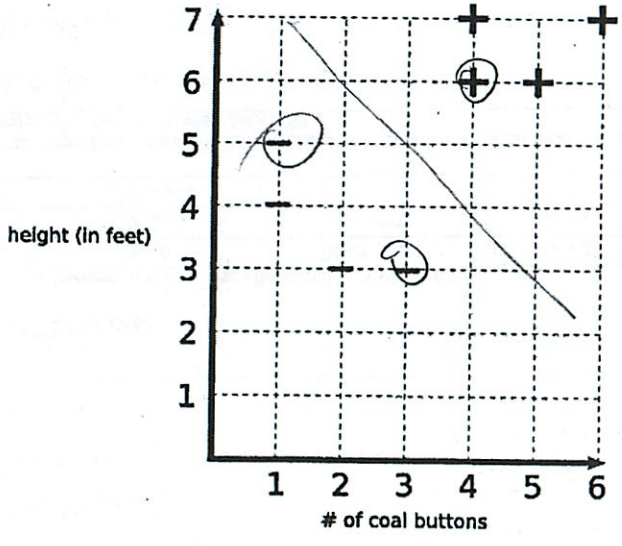
$= \begin{cases} 1 & \text{if } 1.5 \geq 0 \\ 0 & \text{o/w} \end{cases}$

So $O_D = X$ or Y

Quiz 4, Problem 1: SVMs (50 points)

Part A: F.R.O.S.T. and the Snowmen (25 Points)

You have been hired by the Foundation for Research into Occult Snow Transformations (F.R.O.S.T.) to investigate a strange new phenomenon. It seems that certain snowmen can be turned into living, talking creatures when an old silk hat is placed on their heads. So far, F.R.O.S.T. researchers have determined that the factors that best predict whether a particular snowman can be thus transformed are the snowman's height and the number of coal buttons on its chest. Their preliminary data is below: each "+" represents a magical talking snowman, and each "–" represents an ordinary, non-transforming snowman.



A1 (5 points): You decide to use a linear SVM to classify magical vs. non-magical snowmen. **Draw the resulting decision boundary in the graph above, and circle the support vectors.**

A2 (10 points): In this SVM, what is the vector w and the constant b?

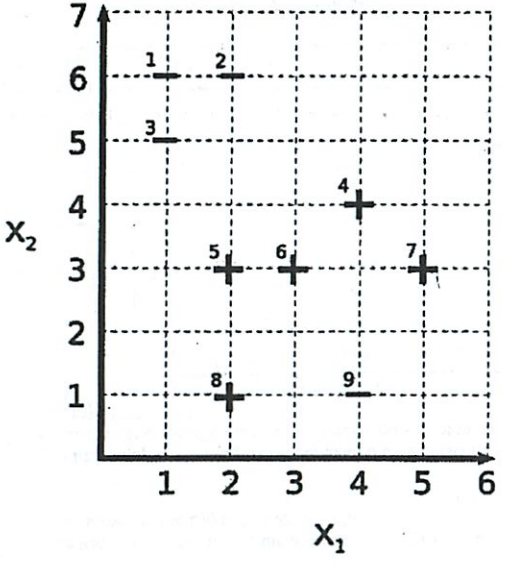
$w = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$ $b = -4$

A3 (10 points): Suppose the F.R.O.S.T. scientists discover another magical, talking snowman that is five feet tall and has six coal buttons on his chest. What would the α value of this new data point be? Justify your answer.

$\alpha = 0$, point is not a support vector

Part B: Kernels (25 Points)

In this section, you will project the data below into a new space with $\phi(u) = <|x_1 - x_2|>$. That is, you project the two-dimensional vector u into a one-dimensional vector in a one-dimensional space.



B1 (7 points): What is the kernel function K(u,v) for this transformation?

$K(\vec{u},\vec{v}) = \phi(\vec{u}) \cdot \phi(\vec{v}) = <|x_1 - x_2|><|y_1 - y_2|>$

B3 (10 points): In the transformed space, what is the vector w and the constant b?

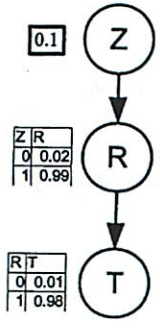
$\vec{w} = \langle -2 \rangle$

$b = 5$

B2 (8 points): What is the final classifier produced. Express your answer in the original space, not the transformed space.

$\text{sign}(5 - 2|x_1 - x_2|)$

Quiz 4, Problem 2:
Bayesian Inference (50 points)



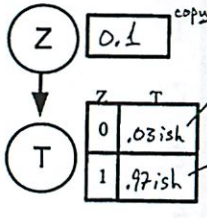
Zombies once again threaten the town, but this time the healthy folks are using science to help keep themselves on a brain-free diet!

When someone first becomes a zombie, their zombieness is latent for a while, as the transformation slowly progresses, but successful isolation at this stage is critical for avoiding new zombie infections.

New zombies quickly begin to produce a protein R that one can easily test for. One percent of infected individuals do not produce protein R (or perhaps just not yet), while protein R has been found in two percent of subjects who turned out not to have turned into zombies later,.

The test is 98% sensitive to the presence of the protein, and 99% specific to it, as shown in the Bayes Net on the left.

A (9 points) To make things a little simpler for the general population to understand, you want to give information in terms of how sensitive and specific the test is to zombiehood, not to the protein:



copy from above

Show your work for partial credit:

$P(T|\bar{z}) = P(T|R)\,P(R|\bar{z}) + P(T|\bar{z})\,P(\bar{R}|\bar{z})$
$.98 \quad .02 \quad .01 \quad .98$

$P(T|z) = P(T|R)\,P(R|z) + P(T|\bar{R})\,P(\bar{R}|z)$
$.98 \quad .99 \quad .01 \quad .01$

B (6 points) Of 100 townfolk, how many will test positive:

$P(T) = P(T|z)\,P(z) + P(T|\bar{z})\,P(\bar{z}) = .97 \cdot .1 + .03 \cdot .9 = .097 + .027 = .124 \rightarrow 12$ people

C (6 points) ...and how many healthy ones will test positive:

$.03 \cdot .9 = .027 \rightarrow 3$ people

D (6 points) What is the overall accuracy of the test? (the probability that it predicts correctly)

$P(T|z)\,P(z) + P(\bar{T}|\bar{z})\,P(\bar{z}) = .97 \cdot .1 + .97 \cdot .9 = .97$

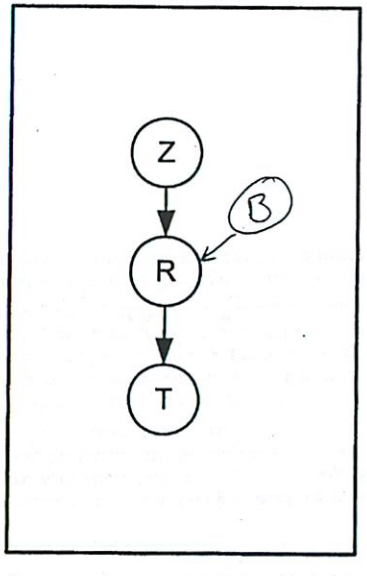
E (5 points) A visitor, Ulric, was tested upon entering the town, and had a positive result. Ulric insists that he is healthy, and hasn't come anywhere near any zombies. To be sure, the guards decide to test him again. What assumption(s) are the guards making about the disease and the testing process?

That each time the same person is tested is an independent event.

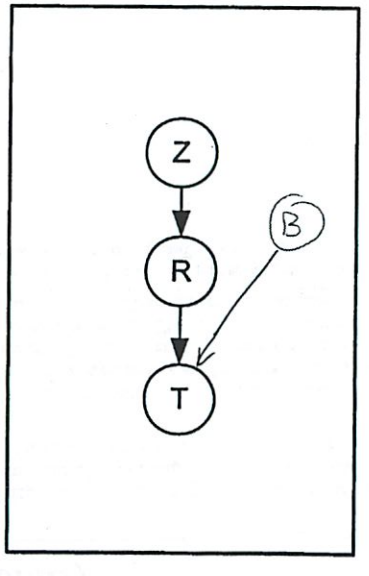
F (5 points) Victor and Wendy both arrive from Brainton, where Ulric came from too, and they both test positive, though they appear quite healthy. Ulric is doing fine in quarantine. The town scientists begin to consider two alternate hypotheses. Both of the alternate hypotheses allow being from Brainton (indicated with B) to explain away a positive test result. Both of the models extend the original three-node model with just the one extra node B. Draw the two possible models they should consider.

Draw only the new node B and its edges — you do not need conditional probability tables (yet):

Alternate Model A



Alternate Model B



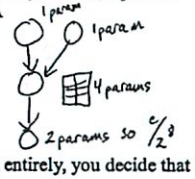
Despite **two positive test results each**, Ulric, Victor, and Wendy all turn out to be healthy. **An expedition to (apparently zombie-free) Brainton revealed that 80% of its population test positive using your town's test.** One of the alternate models is very likely!

G (6 points)
Assume that you think a model whose Bayes Net representation has k+1 parameters is about **half** as likely *a priori* as a model with k parameters, unless k = 0, in which case assume that the model probability is some constant, c.

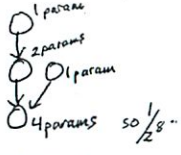
How likely is the original (three variable) model, *a priori*? $c/32$

How likely are each of the models you drew, *a priori*?

Alternate Model A $c/256$



Alternate Model B $c/256$



H (7 points)
To rule out your original model entirely, you decide that an alternate model has to explain the data not just better, but **at least 100 times better** than the original model does.

Was the expedition necessary to rule out the original model, or did we know enough once we knew the three visitors were actually healthy, despite the tests?

Circle one: Needed the Expedition (Ulric, Victor, and Wendy were enough)

Explain, stating any assumptions you need to make:

Assuming dependent events,
$P(UVW|orig\ model) = .03^3$
Assuming indep. events
$P(UVW|orig) = .03^6$
Let's assume dep. events for each test, as the guards should have.
$P(UVW|either\ alt\ model) = .8^3$

$P(UVW|orig)P(orig) \cdot 100 <?$
$P(UVW|alt)P(alt)$
$.03^3 \cdot 1/32 \cdot 100 <? .8^3 \cdot 1/256$
$.3\%$ $.3\%$
$.000003 \cdot 100 <? .24/8$
$.0003 <? .03$
yes!
So the alt model is preferred

Quiz 5, Problem 1, (40 points)

You have made it. You have graduated, started a thriving company, found a spouse, got a home in the suburbs, children. Now it is time to get a dog. Knowing nothing about dogs yourself, you ask a dog-loving, but inarticulate friend to characterize a series of dogs as Good or Bad. You hope to learn from your friend's characterizations.

Your first task is to pick some descriptors. You decide to look at intelligence, breed, exceptional characteristics, and gender. Male and female form a set. Yes and no form a set. Breeds are the leaves of a tree and form groups specified by the American Kennel Club: Bouvier and Collie are part of the Herding group, Beagle and Dachshund are part of the Hound group, and Chihuahua is part of the Toy group. Herding, Hound, Toy, Working, and Sporting all belong to the Recognized Dog category.

Using Arch learning, indicate in the table what is learned from each example and identify the heuristic involved by name, if known. If nothing is learned, put an x in the corresponding 2 columns.

Part A (32 points)

Candidate	Gender	Breed	Exceptional Quality	Intelligent	Heuristic	What is learned
Good	Male	Bouvier		Yes	—	INITIAL MODEL
Bad	Female	Bouvier		Yes	REQUIRE	MUST BE MALE
Bad	Male	Bouvier	Nasty	Yes	FORBID	MUST NOT BE NASTY
Good	Male	Collie		Yes	CLIMB TREE	HERDING
Good	Male	Collie		No	DROP LINK	INTELLIGENCE DOESN'T MATTE
Good	Male	Beagle		No	CLIMB TREE	RECOGNIZED DOG
Good	Male	Dachshund		No	—	—
Bad	Female	Chihuahua		No	—	—
Bad	Male	Chihuahua	Nasty	Yes	—	—

Part B (8 points)

Exhibit an example, which if used as the second example, would teach two characteristics at once.

Candidate	Gender	Breed	Exceptional quality	Intelligent
GOOD	FEMALE	COLLIE		NO

↑ DROP ↑ CLIMB ↑ DROP

Quiz 5, Problem 2, (21 points)

President Hockfield, eager to understand student dining preferences, decides to collect information in a self-organizing map. She starts by constructing some classification trees on a paper napkin. As you can see, she views students from the perspective of departments, such as EE, MechE, Physics, and XV:

Next, she develops an initial map with three relations:

Map element	Subject	Relation	Object
1	EE	Eats	Pizza
2	XV	Eats	Caviar
3	MechE	Drinks	Alcoholic beverage

Note that in this map, there is no concept of neighboring cell. She has decided to put that off for a while. Accordingly, each adjustment to the map alters only that cell judged closest to the new relation flowing into the map.

Part A

Now, President Hockfield observes the following:

Subject	Relation	Object
MechE	Eats	Nachos

Not fully remembering Professor Winston's lecture, she asks you to adjust the map to accommodate this information using the classification trees and edit distance (number of up and down steps in the classification tree needed to go from one combination to the other) to determine which element to change. You don't quite remember how to change the selected element, but fortunately, you took a photograph of Professor Winston's practice board:

If sample object classification extends map object classification
Then extend map classification object by one class
Else if sample object classification and map object classification are the same
Then do nothing
Else trim map object classification by one class

To save time, do not bother copying in any information that remains unchanged.

Map element	Subject	Relation	Object
1	ENGINEER	EATS	JUNK
2			
3			

Part B

Repeat with the following observation. Note that any change you made in Part A is to be carried into this part:

Subject	Relation	Object
EE	Drinks	Beer

Again, you need only enter the changed row. You need not write in any row that remains the same after the work you did in Part A.

Map element	Subject	Relation	Object
1			
2			
3	ENGINEER	DRINKS	BEER

Part C

Repeat with the following observation. Note that any changes you made in previous parts are to be carried into this part:

Subject	Relation	Object
Management	Drinks	Burgundy

Again, you need only enter the changed row. You need not write in any row that remains the same after the work you did in previous parts.

Map element	Subject	Relation	Object
1			
2			
3	EVERYBODY	DRINKS	ALCOHOLIC

Quiz 5, Problem 3, (15 points)

A: A Japanese Haiku is a poem that consists of 17 syllables arranged in 3 lines of 5, 7, and 5 syllables. These are very popular, so a publishing house has asked you to develop a system that can generate good ones automatically.

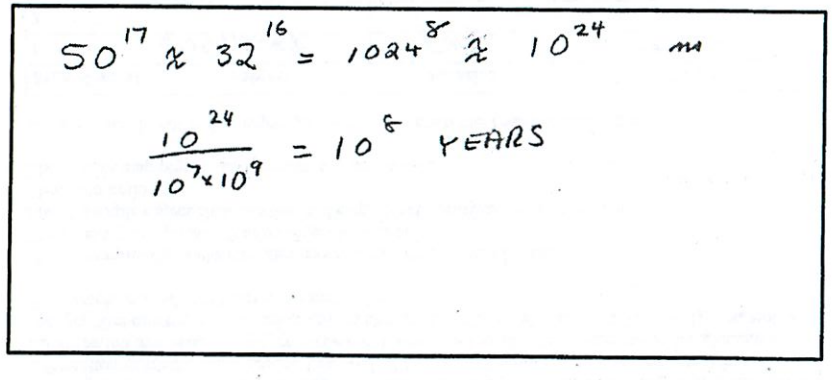
The syllables are drawn from the approximately 50 syllables that are found in Japanese.

James, who dropped 6.034 early in the term suggests that you just generate all possible Haikus and hand them to a panel of judges who will pick the best.

Part A (12 points)

Determine how long it would take to calculate all possible Haikus given a computer that produces one Haiku per nanosecond. So you won't think you need a calculator, you may make the following approximations:

$50 = 32$, $17 = 16$, and $1024 = 1000$, and seconds/year = 10^7



Part B (3 points)

Is James suggestion practical? Circle your answer

Yes (No) ONLY IF STARTED B4 THE DINOSAURS

Quiz 5, Problem 4, (24 points)

Circle the best answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

Genetic algorithms, without crossover, is best described as a kind of

1. Instance of General Problem Solver architecture
2. Instance of subsumption architecture
3. Instance of SOAR architecture
4. Hill climbing search
5. None of the above

Crossover is best described as

1. A product of means-ends analysis
2. A product of an abstraction barrier
3. A label for the strange mating behavior of Zebra Finches
4. A means to escape local maxima in a search space
5. None of the above

The SOAR architecture is best described as

1. A programming language
2. A descendant of the General Problem Solver architecture
3. An amalgam of several ideas
4. The design philosophy that led to the Stata center
5. None of the above

The Genesis architecture (Winston's research focus) is best described as

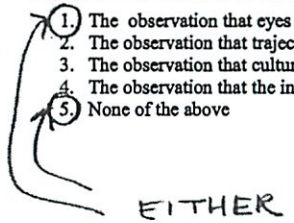
1. A search for a universal representation
2. A commitment to reasoning as the distinguishing feature of human intelligence
3. A search for principles that explain the intelligence of non human primates
4. A demonstration that natural language has little or no role in explaining human intelligence
5. None of the above

Minsky's Emotion Machine/Society of Mind architecture is best described as

1. Focused on explaining visual problem solving
2. Focused on reasoning on multiple levels
3. A commitment to the idea that language is the differentiating feature of human intelligence
4. A commitment to the idea that culture is reflected in the myths associated with the culture
5. None of the above

Intermediate features and the Goldilocks principle is best explained as

1. The observation that eyes are too small and faces are too large
2. The observation that trajectories are too small and transitions are too large
3. The observation that cultures are defined by their fairy tales
4. The observation that the intelligence of a songbird lies half way between insects and humans
5. None of the above



Name	
email	

6.034 Final Examination
December 15, 2010

Circle your TA and principle recitation instructor so that we can more easily identify with whom you have studied:

Martin Couturier	Kenny Donahue	Gleb Kuznetsov
Kendra Pugh	Mark Seifter	Yuan Shen

Robert Berwick	Randall Davis	Lisa Fisher

Indicate the approximate percent of the lectures, mega recitations, recitations, and tutorials you have attended so that we can better gauge their correlation with quiz and final performance and with attendance after OCW video goes on line. Your answers have no effect on your grade.

	Lectures	Recitations	Megas	Tutorials
Percent attended				

Quiz	**Score**	**Grader**
Q1		
Q2		
Q3		
Q4		
Q5		

There are 48 pages in this final examination, including this one. In addition, tear-off sheets are provided at the end with duplicate drawings and data. As always, open book, open notes, open just about everything.

Quiz 1, Problem 1, Rule Systems (50 points)

Kenny has designed two suits for the Soldier Design Competition, and he and Martin like to grapple in the prototypes on Kresge Oval.

- Kendra insists the suits qualify as **"deadly weapons"** and Kenny should give them to her for safekeeping.
- Kenny and Martin insist that they are examples of an **"enhanced prosthesis"** and that they should be able to keep them

The TAs decide to use Rule-Based Systems to resolve their dispute.

Rules:

P0	IF (AND ('(?x) is a Crazy Physicist', '(?x) is an Engineer') THEN ('(?x) builds a Weaponized Suit'))
P1	IF ('(?y)'s (?x) is Cardinal Red' THEN ('(?y)'s (?x) is not US Govt. Property'))
P2	IF (OR (AND ('(?y) is an Engineer', '(?y)'s (?x) is Really Heavy'), '(?y)'s (?x) is stolen by the Air Force') THEN ('(?y)'s (?x) is a Deadly Weapon'))
P3	IF (OR ('(?y) is not evil', '(?y) is a Roboticist') THEN ('(?y)'s suit research is not Evil'))
P4	IF (AND ('(?y)'s (?x) research is not Evil', '(?y)'s (?x) is not US Govt. Property') THEN ('(?y)'s (?x) is an Enhanced Prosthesis'))

Assertions:

A0: (Kenny is a Roboticist)

A1: (Martin is an Engineer)

A2: (Kenny's suit is Cardinal Red)

A3: (Martin's suit is Really Heavy)

Part A: Backward Chaining (30 points)

Make the following assumptions about backward chaining:

- The backward chainer tries to find a matching assertion in the list of assertions. If no matching assertion is found, the backward chainer tries to find a rule with a matching consequent. In case none are found, then the backward chainer assumes the hypothesis is false.
- The backward chainer never alters the list of assertions; it never derives the same result twice.
- Rules are tried in the order they appear.
- Antecedents are tried in the order they appear.

Simulate backward chaining with the hypothesis

Kenny's suit is an enhanced prosthesis

Write all the hypotheses the backward chainer looks for in the database in the order that the hypotheses are looked for. The table has more lines than you need. We recommend that you use the space provided on the next page to draw the goal tree that would be created by backward chaining from this hypothesis. **The goal tree will help us to assign partial credit** in the event you have mistakes on the list.

1 Kenny's suit is an enhanced prosthesis
2
3
4
5
6
7
8
9
10

Draw Goal Tree Here for Partial Credit

Part B: Forward Chaining (20 points)

Let's say, instead, our assertions list looked like this:

A0: Gleb is an Engineer
A1: Gleb's laptop is Really Heavy
A2: Gleb's suit is Really Heavy

B1 (4 points)

CIRCLE any and all rules that match in the first iteration of forward chaining

P0	P1	P2	P3	P4

B2 (4 points)

What assertion(s) are added or deleted from the database, as a consequence of this iteration?

B3 (4 points)

CIRCLE any and all rules that match in the second iteration of forward chaining

P0	P1	P2	P3	P4

B4 (4 points)

What assertion(s) are added or deleted from the database, as a consequence of this iteration?

B5 (4 points)

You take the same assertions as at the beginning of problem B, above, and re-order them:

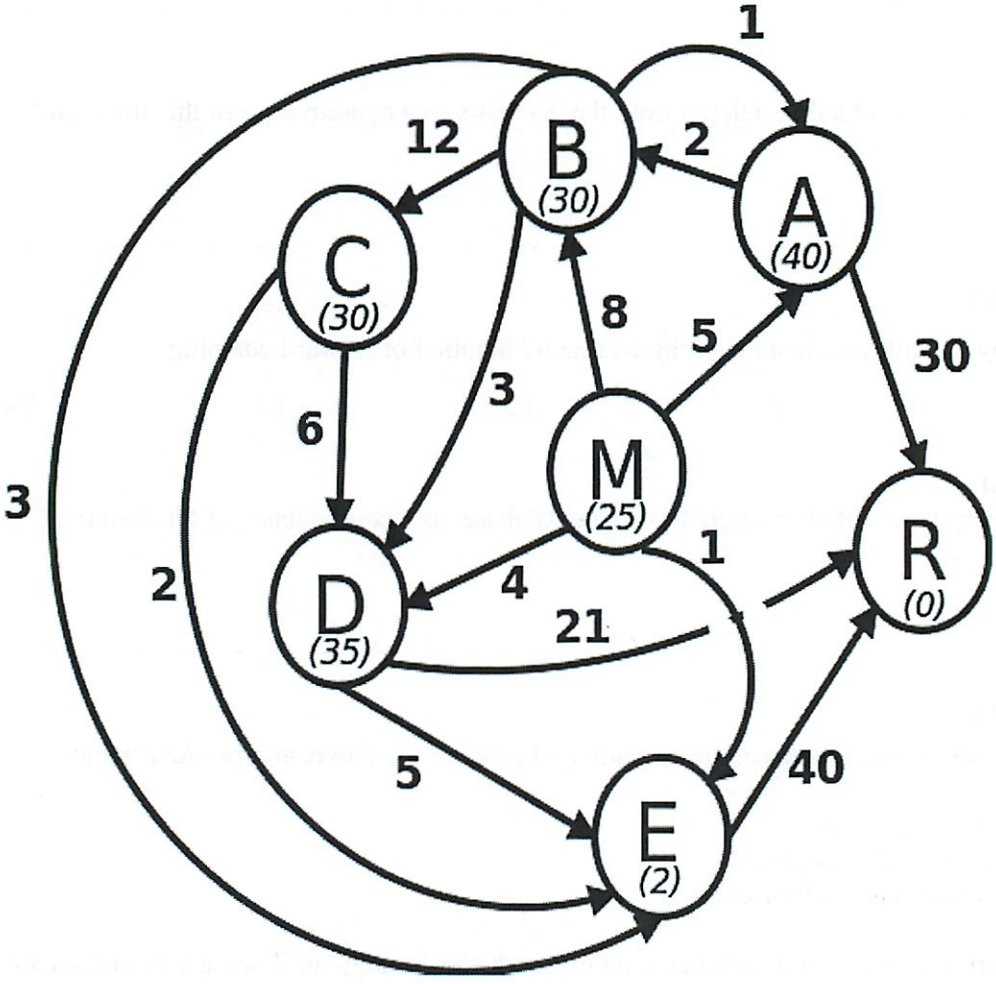
A0: Gleb is an Engineer
A1: Gleb's suit is Really Heavy
A2: Gleb's laptop is Really Heavy

Now, you start over, and run forward chaining from the beginning, until no new assertions are added to or deleted from the database. Is Gleb's laptop a Deadly Weapon?

Quiz 1, Problem 2, Search (50 points)

As you get close to graduating MIT, you decide to do some career planning. You create a graph of your options where the start node is **M = MIT** and your goal node is **R = Retire**, with a bunch of options in between. Your graph includes edge distances that represent, roughly, the "cost of transition" between these careers (don't think too hard about what this means). You also have heuristic node-to-goal distances which represent your preconceptions about how many more years you have to work until you retire. For example, you think it will take 25 years to go from MIT (M) to retirement (R), 30 years from Grad School (B), but only 2 years from Entrepreneur (E).

A = Wall Street | B = Grad School | C = Professor | D = Government | E = Entrepreneur



Part A: Basic search (25 points)

In all search problems, use alphabetical order to break ties when deciding the priority to use for extending nodes.

A1 (3 points)

Assume you want to retire after doing the least number of different jobs. Of all the basic search algorithms you learned about(that is, excluding branch and bound and A*) which one should you apply to the graph in order to find a path, **with the least search effort**, that has the minimum number of nodes from M to R?

A2 Basic Search Chosen Above (7 points)

Perform the search you wrote down in A1 (with an Extended List). Draw the search tree and give the final path.

Tree:

Path:

A3 Beam Search with w=2 (15 points)

Now you are interested in finding a path and the associated distance. Try a **Beam Search** with a width **w=2**, ***with*** an extended list. As before, you are looking for a path from **M** to **R**. Use the "preconceptions" heuristic distances indicated in parentheses at each node.

Tree:

Path, if any:

Extended nodes in order extended:

Part B: Advanced Search (25 points)

B1 Branch and Bound with Extended List (15 points)

Use Branch and Bound search with an Extended List to find a path from **M** to **R,** as well as the extended node list. Use this space to draw the corresponding tree and show your work.

Tree:

Path:

Extended nodes in order extended:

B2 Thinking about Search (9 points)

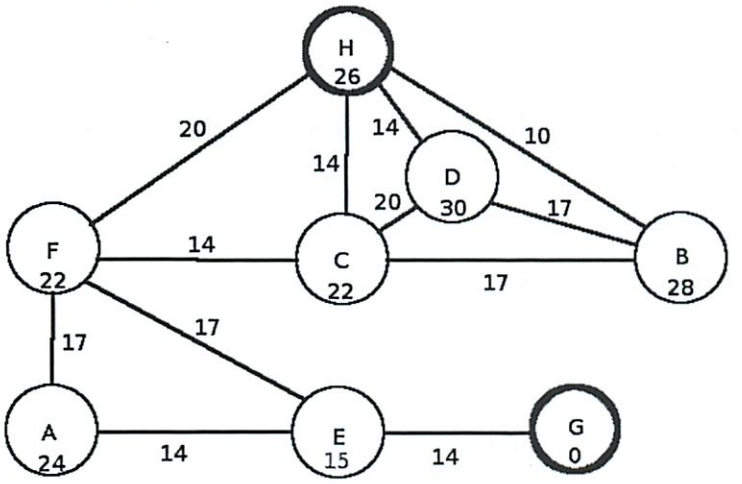
Concisely explain why Branch and Bound with Extended List yields a different result than Beam Search in this problem.

What can we say about the path found by Branch and Bound with Extended List? (We're looking for a fairly strong statement here.)

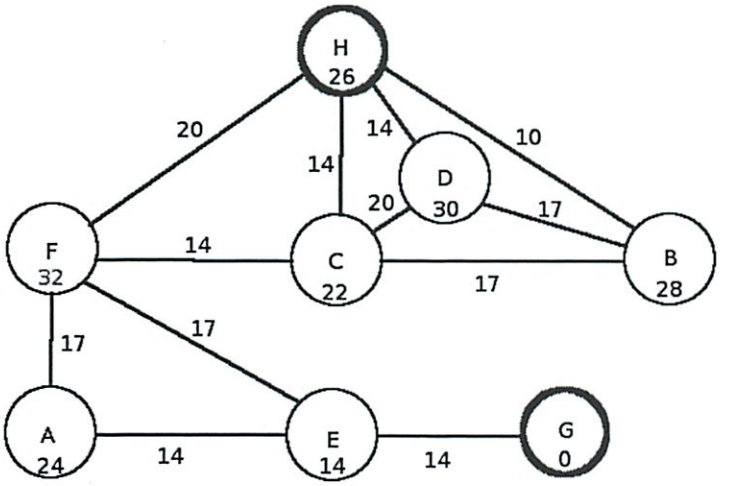
Is there an algorithm that guarantees the same answer as Branch and Bound ***for the graph in this problem,*** but can find the answer with fewer extended paths. If Yes, what is that algorithm? If No, explain why not.

B3 Permissible Heuristics (6 points)

Suppose you are asked to find the shortest path from H to G in the graphs below. For both of the graphs explain why the heuristic values shown are not valid for A*. Note the differences in the graphs at nodes F and E.



Reason(s):



Reason(s):

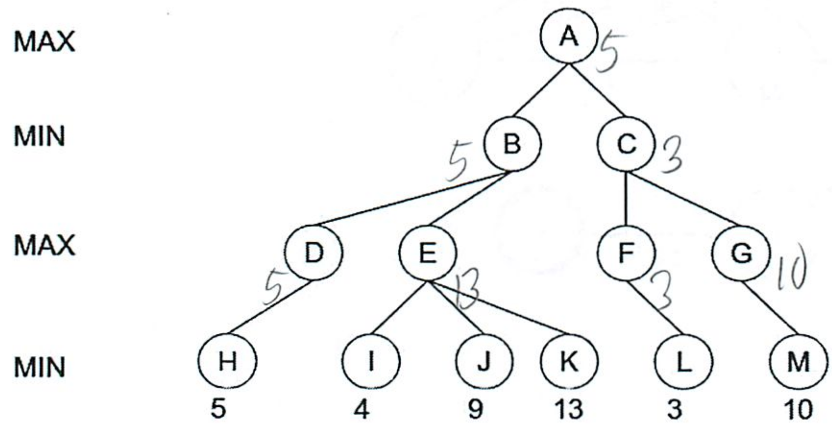
12/15

Quiz 2, Problem 1, Games (50 points)

Part A: Basics(15 points)

A1 Plain Old Minimax(7 points)

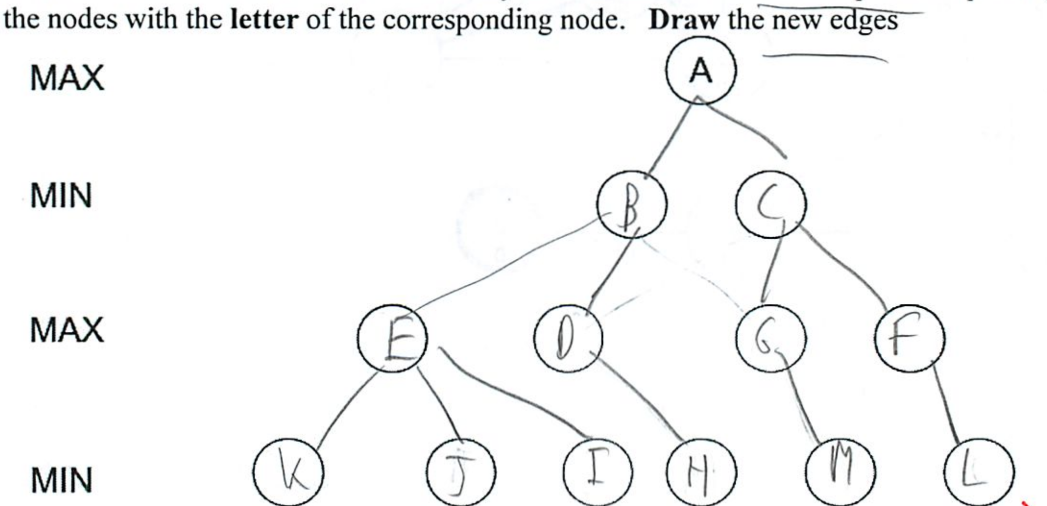
Perform Minimax on this tree. **Write the minimax value associated with each node in the box below, next to its corresponding node letter.**



A= 5	B= 5	C= 3	D= 5	E= 13	F= 3	G= 10

A2 Tree Rotations (8 points)

Using the minimax calculations from part A1, **without** performing any alpha-beta calculation, rotate the children of each node in the above tree at every level to ensure maximum alpha-beta pruning.
- **Fill in** the nodes with the **letter** of the corresponding node. **Draw** the new edges



can't cross

Low to high! would have looked it up

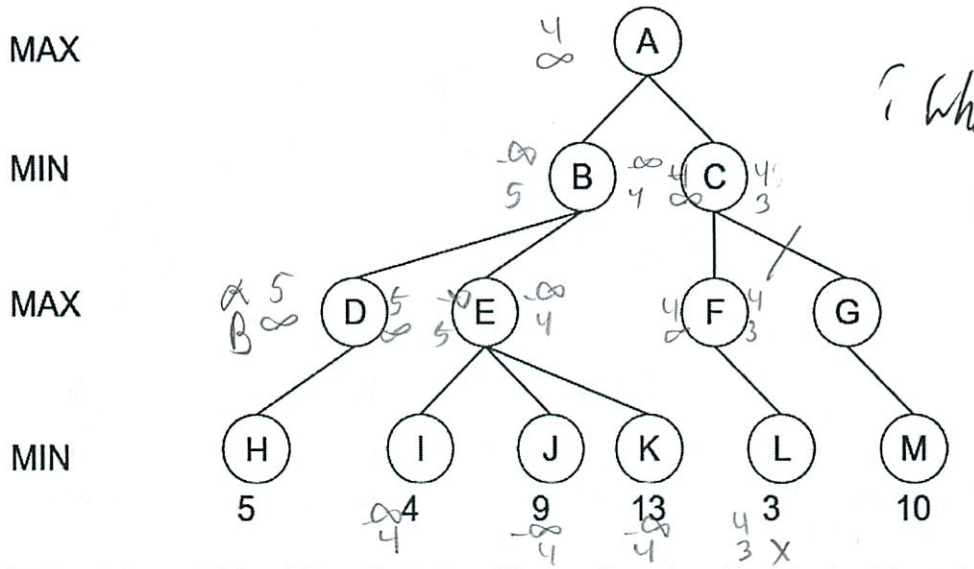
Urg forgot
Max to min?
can do new edges!

Part B: Alpha Beta (35 points)

Here we go

B1: Straight-forward Alpha Beta(15 points)

Perform Alpha Beta search on the following tree.
- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



What's with the 2 types of lines?

Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

H, I, J, K, L

# of evaluations: 9 **List:** H, I, J, K, L

Next Move: B ✓ **Moving towards:** F → I

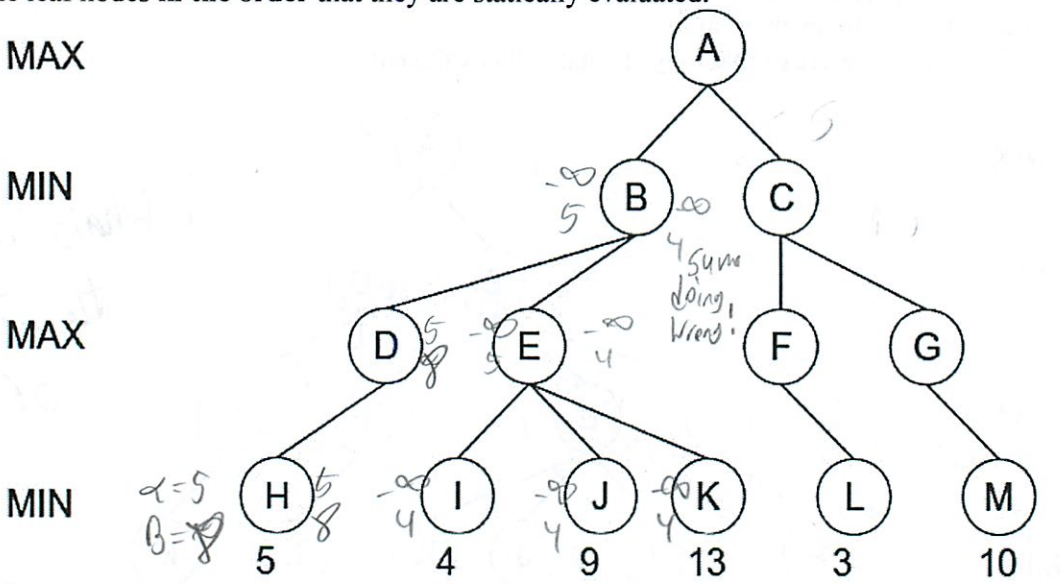
H

Yeah did it wrong

B2: Preset Alpha-Beta (15 points)

Perform alpha-beta search, using **initial values** of **alpha = 5 and beta = 8.**

- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

# of evaluations: 2 **List:** H, C

Next Move: B **Moving towards:** H ← always same?

B3: Alpha-Beta Properties (5 points)

If you were able to maximally prune a tree while performing Alpha-Beta search, approaximately how many static evaluations would you end up doing for a tree of depth d and branching factor b?

$O(b^{\frac{d}{2}})$

? I guess just something to know

Part B: Applying Constraints (42 points)

You decide to switch to a new representation that uses the courses as variables and the times as values.

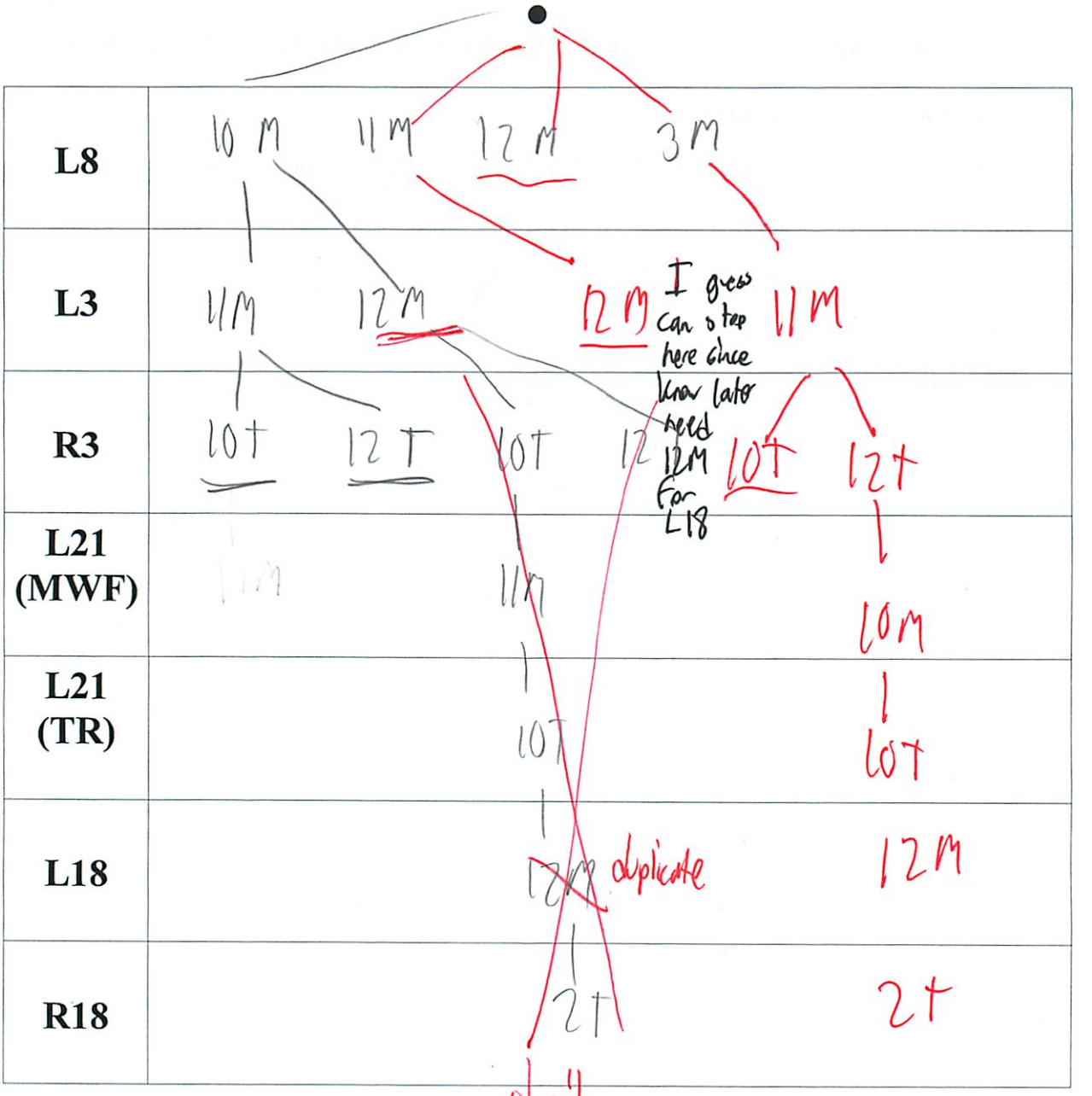
B1 (5 points)

The initial domains are given below. Cross out the values that are incompatible with Constraint (3).

Variable	Domain
L8	10M 11M 12M 1M 2M 3M
L3	11M 12M
R3	10T 11T 12T 1T
L21 (MWF)	10M 11M
L21 (TR)	10T 11T
L18	12M 1M
R18	11T 1T 2T

B2 (16 Points)

Run the DFS with **forward checking only** on your defined variables and the reduced domains you found in Part B1 by applying Constraint(3).



Part A: Picking a Representation (8 points)

In order to fill in this schedule, you decide to set it up as a CSP using meeting times as variables and courses as the values of in the domains. After filling in the domain table, this is what you see:

Variable	Domain
10M	L8, L21
11M	L3, L8, ~~L21~~
12M	L3, L18, L8
1M	
2M	
3M	L8
10T	R3, L21
11T	
12T	R3
1T	
2T	R18
3T	

What is wrong with the way that this problem is set up and why?

L21 M, T must =

No value for 3T

So drop it?

Quiz 2, Problem 2,
Constraint Propagation (50 points)

After taking 6.034, your decide to offer your services to incoming freshman to help them set up their course schedules. One particular freshman comes to you with four classes as well as an availability schedule (grayed out boxes represent reserved times).

Course	Lecture Times Offered	Recitation Times Offered
3.091	MWF 11,12	TR 10,11,12,1
18.01	MWF 12, 1	TR 11,1,2
8.01T	MWF 10, 11, 12, 1, 2, 3	NONE
21F.301	MTWRF 10, 11	NONE

Time	MWF	TR
10		
11		
12		
1		
2		
3		

For easier bookkeeping you adopt the folowing naming convention (L = Lecture, R = Recitation, # = course number):

3.091 Lecture	→ **L3**	MWF10	→**10M**
3.091 Recitation	→ **R3**	MWF11	→**11M**
8.01T Lecture	→ **L8**		
18.01 Lecture	→ **L18**	TR10	→**10T**
18.01 Recitation	→ **R18**	TR11	→**11T**
21F.301 Lecture	→ **L21**		

You also devise this set of constraints for yourself:

(1) Each class must be assigned to exactly one timeslot
(2) Each timeslot can be assigned to a maximum of one class
(3) No classes can be scheduled during the grayed out time periods
(4) The TR selection for 21F.301 must occur at the same time as the MWF selection.

B3 (5 Points)

How many times did the algorithm need to backtrack?

~~5~~ I think L8 does not count

B4 (10 Points)

It occurs to you that you may be able to accelerate the process of finding a solution if you were to perform domain reduction with **propagation through singletons** *before* running the DFS. Fill in your updated domain table with the results of your computations.

? how do before running well can all constraints

Variable	Domain
L8	~~10M~~ ~~11M~~ ~~12M~~ 3M
L3	11 M ~~12M~~
R3	~~10 T~~ 12 T
L21 (MWF)	10 M ~~11M~~
L21 (TR)	10 T
L18	12M
R18	2 T

? So only 1 possible sol ?

B5 (6 Points)

Run DFS with **constraint checking only** on your updated domain table:

L8	3M
L3	11M
R3	12T
L21 (MWF)	10M
L21 (TR)	10T
L18	12M ✓
R18	2T

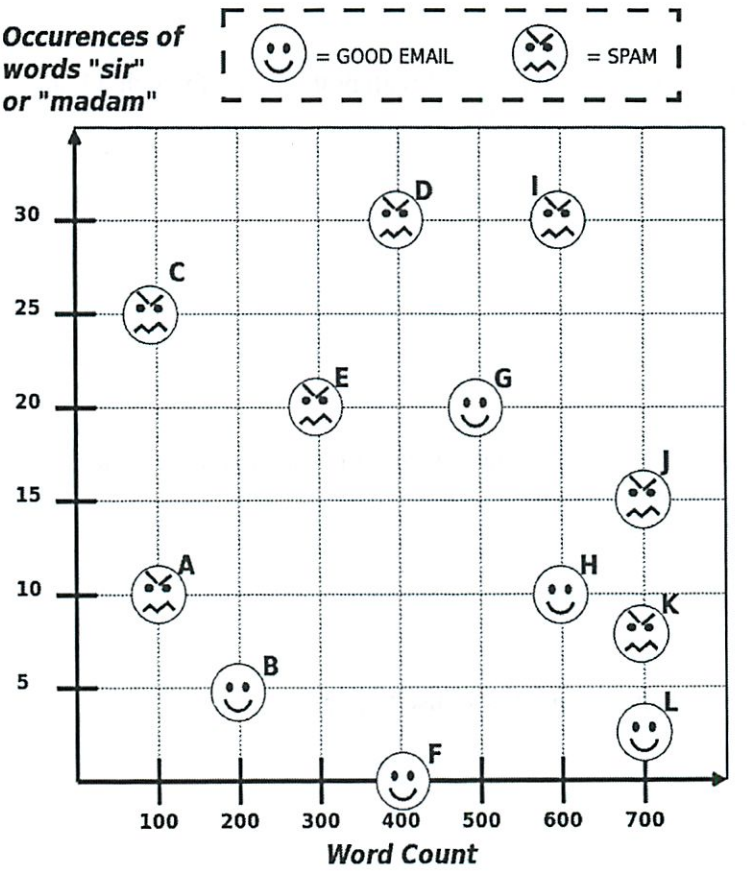
Quiz 3, Problem 1
KNN and ID Trees (50 points)

After receiving yet another "Dear sir or madam.." email, you decide to construct a spam filter.

Part A: Nearest Neighbors (25 points)

For your first attempt you decide to try using a k Nearest Neighbors model. You decide to classify spam according to 2 features: email word count and occurrences of the words "sir" or "madam".

A1 (10 points)



Draw the decision boundary for **1-nearest-neighbor** on the above diagram of the given training data. **Use the center of the faces as the positions of the training data points.**

A2 (8 points)

How will **1-nearest-neighbor** classify an email with **200** words of which **9** are the word "sir"?
Plot this point on the graph as X? (2pts)

How will **3-nearest-neighbors** classify an email with **600** words of which **7** are the word "madam"?
Plot this point on the graph as Y? (3pts)

How will **5-nearest-neighbors** classify an email with **500** words of which **25** are the word "madam"?
Plot this on the graph as Z? (3pts)

A3 (7 points)

List which points yield errors when performing leave-one-out cross validation using **1-nearest-neighbor** classification. (3 pts)

How would one go about selecting a good **k** to use? (4 pts)

Part B: ID Trees (25 points)

Realizing nearest neighbors may not be the best tool for building a spam filter, you decide to try another classifier you learned about in 6.034: Identification Trees.

B1 (8 points)

It appears that the over-use of the words "sir or madam" seems to be a strong hint at an email being spam.

What is the minimum disorder and minimum-disorder decision boundary when you consider only the dimension of "Sir or Madam"? You can use fractions, real numbers, and logarithms in your answer.

Approximate boundary:

Associated Disorder:

B2 (8 points)

Suppose we were given the following additional information about our training set:

Emails I, G, J, H, K and L are from Anne Hunter.

One of those emails might be important so you don't want to risk missing a single one so you re-label all Anne Hunter emails in the training data set to be good emails. You are to find the best axis-parallel test given the revised labellings of good email and spam.

(NOTE: Use the unlabeled graphs in the tear-off sheets section if you need it to visualize the modified data).

B2.1 Which emails does your new test missclassify on the **modified** data? **(4pts)**

B2.2 What is the disorder of your new test on the **modified** training data set?
Leave your answer as a function of fractions, real numbers, and logarithms. **(4pts)**

B3 (9 points)

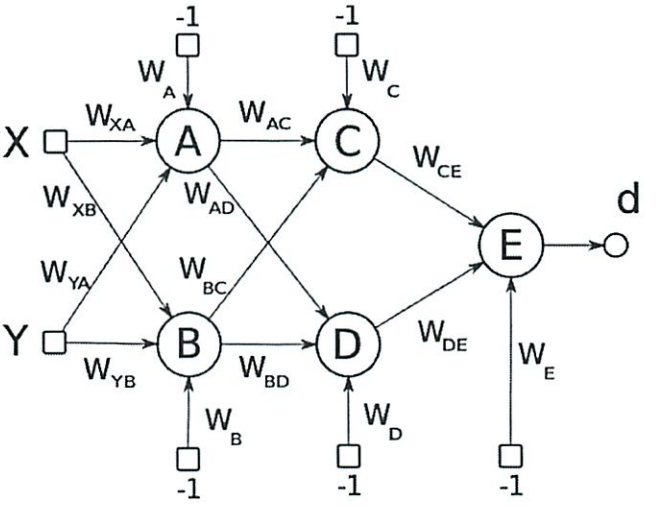
Soon, you decide that your life goal is no longer to be a tool for a Harvard or Sloanie startup so you decide that **all emails from Anne Hunter should be marked as spam. (Again, use the unlabeled graphs in the tear-off sheets if you need them).**

Given the revised determination of what is good email and spam, draw the **disorder minimizing identification tree** that represents your fully trained ID-tree spam filter. You may use any horizontal and vertical classifiers in the dimensions of word count and "sir or madam" occurrences. Ties should be broken in order of horizontal then vertical classifiers.

Quiz 3, Problem 2, Neural Nets (50 Points)

Note that this problem has three completely independent parts.

Part A: Derivations (14 pts)



A1. (7 pts) Using what you've learned from doing lab 5, write out the equation for $\frac{dP}{dw_{CE}}$ expressed in terms of o_i , d, and/or any weights and constants in the network. (o_i refers to the output of any neuron in the network.)

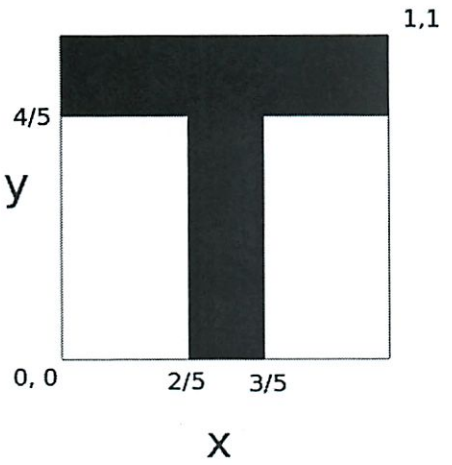
A2. (7 pts) Write out the equation for $\frac{do_E}{dw_{XA}}$ expressed in terms of $\frac{do_i}{dw_{XA}}$, o_i, and/or any weights and constants in the network.

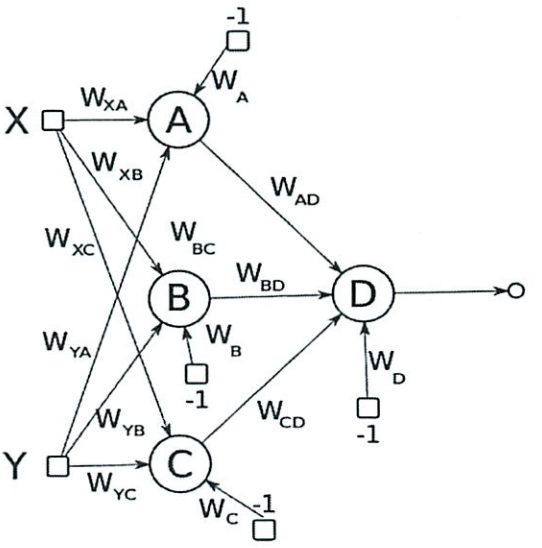
Part B: Letter Recognition (20 pts)

You propose to use a neural network to recognize characters from a scanned page. Letters are represented binary images on a 1x1 unit grid. Assume that scaling and rotation are all done.

Because you want to start with something easy, you start with the problem of recognizing a character as either possibly a T or definitely not a T. During training, each training sample consists of a random point, (x, y), along with the desired 0 or 1 value: 1 if the underlying pixel at (x, y) is part of a T; 0 if the pixel is part of a T's background.

You want to find the most **compact** network that will correctly handle the T problem, so you decide to analytically work out the minimal network that will correctly classify a character as possibly a T or definitely not a T.





Assume you decide to have the above network architecture, fill in the 7 missing weights in the table that are required to accurately classify all points in the image for T. Your weights must be **integer** weights or **integer** constraints on weights! Show your work for partial credit:

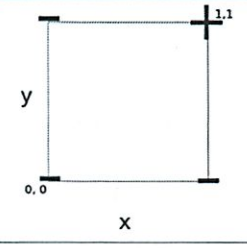
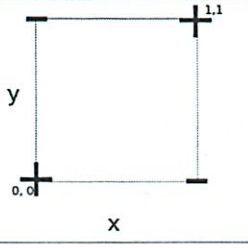
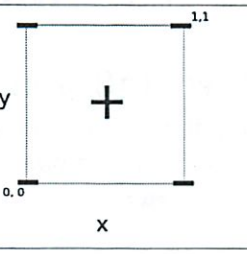
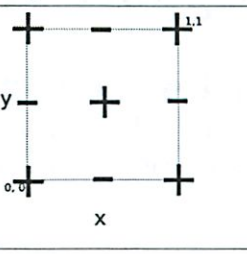
Answers:

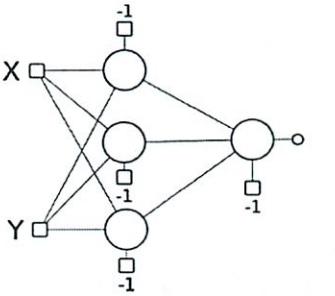
W_{XA}	**0**	W_{XC}	
W_{YA}		W_{YC}	**0**
W_A		W_C	
W_{XB}		W_{AD}	
W_{YB}		W_{BD}	**2**
W_B	**2**	W_{CD}	**2**
		W_D	**3**

Show work here for partial credit:

Part C: Expressive Power of Neural Nets (16 pts)

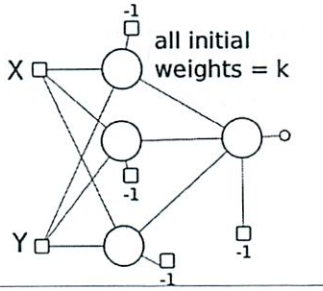
Circle all the functions that these networks are theoretically able to fully learn, and if your answer is No indicate the lowest possible error rate. **List of Functions:**

X AND Y	X = Y	X = 0.5 AND Y = 0.5	X-Shape
			



C1	Can be fully learned?	The minimum # error if No
X AND Y	Yes No	
X = Y	Yes No	
X = 0.5 AND Y = 0.5	Yes No	
X Shape	Yes No	

How about when all initial weights are set to the same constant k?

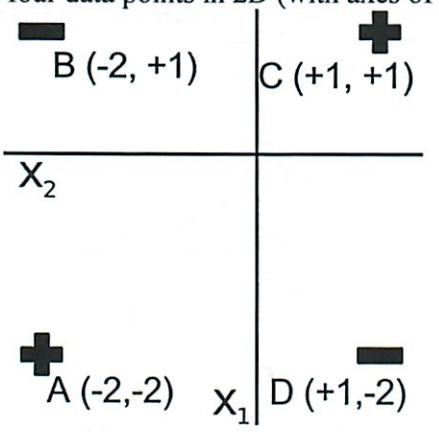


C2	Can be fully learned?	The minimum # error if No
X AND Y	Yes No	
X = Y	Yes No	
X = 0.5 AND Y = 0.5	Yes No	
X Shape	Yes No	

Quiz 4, Problem 1,
Support Vector Machines (50 points)

Part A: Solving SVMs (40 pts)

You decided to manually work out the SVM solution to the CORRELATES-WITH function. First you reduce your problem to looking at four data points in 2D (with axes of x_1 and x_2).



You ultimately want to come up with a classifier of the form below. **More formulas are provided in tear off sheets.**

$$h(\bar{x}): \sum a_i y_i K(\bar{x}, \bar{x}_i) + b \geq 0$$

Your TA, Yuan suggests that you use this kernel:

$$K(\vec{u}, \vec{v}) = u_1 v_1 + u_2 v_2 + |u_2 - u_1||v_2 - v_1|$$

A1(5 pts): Note that $\Phi(\vec{u})$ is a vector that is a transform of $\vec{u}$ and the kernel is the dot product, $\Phi(\vec{u}) \cdot \Phi(\vec{v})$. Determine the number of dimensions in $\Phi(\vec{u})$, and then write the components of $\Phi(\vec{u})$ in terms of $\vec{u}$'s x_1 and x_2 components, u_1 and u_2. Explain why it is better to use $\Phi(\vec{u})$ rather than $\vec{u}$.

$\Phi(\vec{u})$'s components: u_1 u_2 u_1 u_2 $|u_2 - u_1|$

Why better: Since now not linearly separable

A2(10 pts): Fill in the unshaded portion in the following table of Kernel values. they just had 1 value but I think ans can vary

K(A,A) = $-2,-2,-2-2+16$	K(A,B) = $-2-2-21$ -8	K(A,C) = $-2-211$ 4	K(A,D) = $-2-21-2$ 8
K(B,A) =	K(B,B) = $-21-21$ 4	K(B,C) = -2111 -2	K(B, D) = $-2\ 11-2$ 4
K(C,A) =	K(C,B) =	K(C,C) = 111 4	K(C, D) = $111-2$ -2
K(D,A) =	K(D,B) =	K(D,C) =	K(D, D) = $1-21-2$ 4

A3 (10 pts): Now write out the full **constraints equations** you'll need to solve this SVM problem They should be in terms of α_i, b, and constants. **(Hint:** The four data points lie on their appropriate gutters).

	Constraint Equations
1	$\alpha_A + \alpha_C = \alpha_B + \alpha_D$
2	$\alpha_A(1)\begin{pmatrix}-2\\-2\end{pmatrix} + \alpha_B(-1)\begin{pmatrix}-2\\1\end{pmatrix} + \alpha_C(1)\begin{pmatrix}1\\1\end{pmatrix} + \alpha_D(-1)\begin{pmatrix}1\\-2\end{pmatrix} = \begin{bmatrix}w_1\\w_2\end{bmatrix}$
3	$cw_1 \times + cw_2 + cb \leq 0 = h(\vec{x})$
4	$d = \frac{2}{\sqrt{2c^2}}$
5	$h(\vec{x}) = y_A \cdot k(A,A)\alpha_A + y_B k(A,B)\alpha_B + y_C k(A,C)\alpha_C + y_D k(A,D)\alpha_D = +1$ Oh they expand them at and $= 1, 1, -1, -1, 8$ I think mine is = 1x as valid but I should note this both way

A4 (5 pts): Instead of solving the system of equations for alphas, suppose the alphas were magically given to you as:

α_A=1/9	α_B=1/9	α_C=1/9	α_D=1/9	b=1

Compute $\vec{w}$ Given your alphas. **Hint: The number of dimensions in $\vec{w}$ is the same as the number of dimensions of $\Phi(\vec{x})$**

2

$\vec{w} = \begin{matrix}0\\0\\2/3\end{matrix}$ ← 3rd dimension (wait do we need all 3 - why?)

$-2(\frac{1}{9}) + 2(\frac{1}{9}) + \frac{1}{9} - \frac{1}{9} = w_1$

$-2(\frac{1}{9}) - \frac{1}{9} + \frac{1}{9} + 2(\frac{1}{9}) = w_2$

this one hard to visualize

A5 (5 pts): What is the equation of the optimal SVM decision boundary using the answers and values from A5?

So plug in values from before

$h(\vec{x}) = \vec{w}\,\phi(x) + b = \begin{bmatrix} 0 & 0 & -\frac{2}{3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ (x_2 - x_1) \end{bmatrix} + b = -\frac{2}{3}(x_2 - x_1) + 1$

A6 (5 pts): What is the width of the road defined by the optimal SVM decision boundary?

Width = $\frac{2}{\|w\|} = \frac{2}{\sqrt{(\frac{2}{3})^2}} = 3$

Oh can use that eqn

Ok remember how this problem was done

out of scope - or connected is ok

Part B: Kernel for k-Nearest Neighbors (10 pts)

A student noticed that Kernel methods can be used in other classification methods, such as k-nearest neighbors. Specifically, one could classify an unknown point by summing up the fractional votes of all data points, each weighted using a Radial Basis Kernel. The final output of an unknown point $\vec{x}$ depends on the sum of **weighed** votes of data points $\vec{x}_i$ in the training set, computed using the function:

$$K(\vec{x}, \vec{x}_i) = \exp\left(\frac{-\|\vec{x} - \vec{x}_i\|}{s^2}\right)$$

Negatively labeled data points contribute -1 times the kernel value and positively labeled training data points contribute +1 times the kernel value.

You may find the graphs provided in the tear off sheets helpful.

B1 (6 pts)

Using this approach, as **s** increases to infinity, what k does this correspond to in k-NN?

oh - can look this up

gets larger - more k all pts get = vote

well ∩

As **s** decreases to zero, what k does this approximately correspond to in k-NN?

gets smaller - less k only close pts get strong votes

Basically correct - they where more explicit

B2 (4 pts):

State a reason why you would prefer to use the SVM with Radial Basis Kernel solution rather than the method of Weighted Nearest Neighbors with a Gaussian Kernel.

Sometimes easier to seperate like that

They seem to have some diff reasons

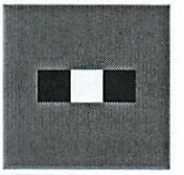
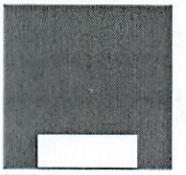
1. Performance. Faster
2. SVM sol has less terms - so shorter, more compact sol
3. Convience - only needs SV pts

oh it was SVM vs NN - not which kernal to use!

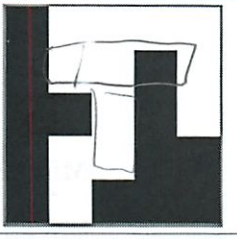
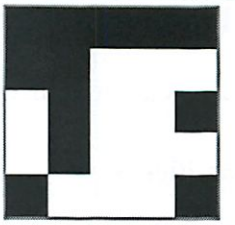
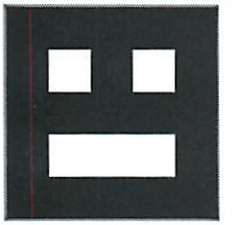
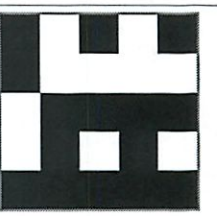
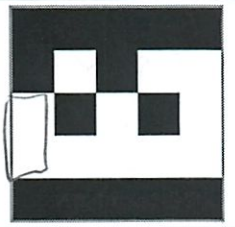
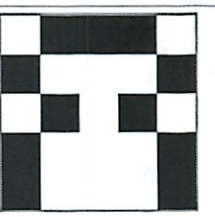
Quiz 4, Problem 2, Boosting (50 points)

Kenny wants to recognize faces in images. He comes up with a few things that he thinks will probably work well as weak classifiers and decides to create an amalgam classifier based on his training set. Then, given an image, he should be able to classify it as a FACE or NOT FACE. When matched against faces, the GRAY part of a classifier can be either WHITE or BLACK)

Classifiers:

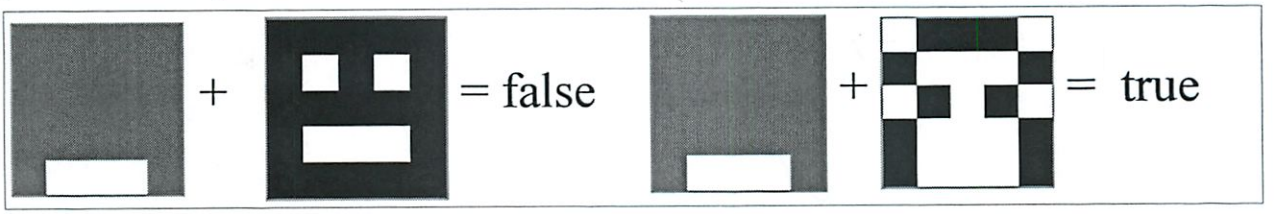
Name	Image Representation
A Has Hair	
B Has Forehead	
C Has Eyes	
D Has Nose	
E Has Smile	
F Has Ear	

Data:

Index	Classification	Image Representation	Index	Classification	Image Representation
1	NOT FACE		5	FACE	
2	NOT FACE too high		6	FACE	
3	NOT FACE		7	FACE	
4	FACE		8	FACE	

Examples:

Here is how classifier E works. Note that gray means "don't care," that is, it doesn't matter whether the pixel in the same position is black or white.



Part A: Pre-Boosting (10 points)

Seems long & hard

A1 Finding the Errors(5 points)

For each classifier, list the data points it gets wrong.

Classifier Name	General Idea	Misclassified
A	Has hair	2, 6, 7 ✓
B	Has forehead	4, 5
C	Has eyes	1, 5, 7
D	Has nose	1, 3, 7 ✓
E	Has smile	3, 4, 7
F	Has ear	8 ✓
G	TRUE (FACE)	1, 2, 3 ✓
H	NOT(A)	1, 3, 4, 5, 8 ✓
I	NOT(B)	1, 2, 3, 6, 7, 8 ✓
J	NOT(C)	2, 3, 4, 6, 8 ✓
K	NOT(D)	2, 4, 5, 6, 8 ✓
L	NOT(E)	1, 2, 5, 6, 8 ✓
M	NOT(F)	1, 2, 3, 4, 5, 6, 7 ✓
N	FALSE (NOT_FACE)	4, 5, 6, 7, 8 ✓

Confused

Opposite seems like it

A2 Weeding out Classifiers (5 points)

Which, if any, of the classifiers in A1 can never be useful and why?

H, I, J, K, L, N - Subset of F

H, I, J, K, L, M, N, - greater than $\frac{1}{2}$

Other look good

M has 6

But they did not include $= \frac{1}{2}$

I'll take it

Surprised I got it

B2 Amalgam Classifier (5 points)

What is the overall classifier after 4 rounds of boosting?

$\text{Sign}\left(\frac{1}{2}\ln(3)[B] + \frac{1}{2}\ln(3)[A] + \frac{1}{2}\ln(2)[B] + \frac{1}{2}\ln\left(\frac{5}{3}\right)[A]\right)$ ✓ basically

What is its error rate? ;; need to actually test; or shortcut?

$\frac{2}{8} = 25\%$

Part C: Miscellaneous (15 points)

C1 Using Classifiers (5 points)

Assume the boosting setup in Part B occurs for 100 rounds whether or not the overall classifier gets to the point where all training samples are correctly classified. Place the four classifiers in the following two bins

Frequently selected: A, B	Infrequently selected: C, E ✓ ;

C2 More using Classifiers (5 points)

Which **three** classifiers from A-N would you use so that you would not have to do boosting to get a perfect classifier for the samples. ;; fit togeta

F ~~B~~ A, B, F

2, 4, 5, 6, 7, 8 ;; randomly picked 3 ;

C3 Even more using Classifiers (5 points)

Now, Suppose you are working with **just two classifiers,** neither of which has a zero error rate. Will boosting converge to an overall classifier that perfectly classifies all the training samples?

Yes No

Explain: Yes Can train training samples

Still same error rate as higher weighted weak classifier. So together imperfect

↳ Thought perfect fit in there

Part B: Boosting (25 points)

B1 Synthesis (20 points)

Synthesize boosting using **only classifiers A, B, C, and E.** For ties, choose alphabetically.

	Round1		Round2		Round3		Round 4	
w1	1/8	h_1= B	1/12	h_2= A	1/18	h_3= B	1/24	h_4= A
w2	1/8	e_1= 2/8	1/12 X	e_2= 3/12	1/6	e_3= 6/18	1/8 X	e_4= 3/8
w3	1/8	a_1= $\frac{1}{2}\ln 3$	1/12	a_2= $\frac{1}{2}\ln 3$	1/18	a_3= $\frac{1}{2}\ln 2$	1/24	a_4= $\frac{1}{2}\ln 5/3$
w4	1/8 X		1/4		1/6 X		1/4	
w5	1/8 X		1/4		1/6 X		1/4	
w6	1/8		1/12 X		1/6		1/8 X	
w7	1/8		1/12 X		1/6		1/8 X	
w8	1/8		1/12		1/18		1/24	
e_A	3/8		(3/12)		3/6		(3/8)	
e_B	(2/8)		2/4		(2/6)		1/2	
e_C	3/8		5/12		7/18		5/12	
e_E	3/8		5/12		7/18		5/12	

still need to try but how

Wow!

$\frac{1}{2}\ln\left(\frac{6/8}{2/8}\right)$ $\frac{1}{2}\ln(3)$

$\frac{1}{2}\ln\left(\frac{9/12}{3/12}\right)$ $\frac{1}{2}\ln(3)$

$\frac{1}{2}\ln\left(\frac{12/18}{6/18}\right)$ $\frac{1}{2}\ln(2)$

$\frac{1}{2}\ln\left(\frac{5/8}{3/8}\right)$ $\frac{1}{2}\ln\left(\frac{5}{3}\right)$

Correct $\frac{1}{2}\frac{1}{6/8} = \frac{2}{3}w_i = \frac{1}{12}$ inc $\frac{1}{4}$

Correct $\frac{2}{3}w_i$ inc $\frac{1}{18}$ $\frac{1}{6}$ $\frac{1}{6}$

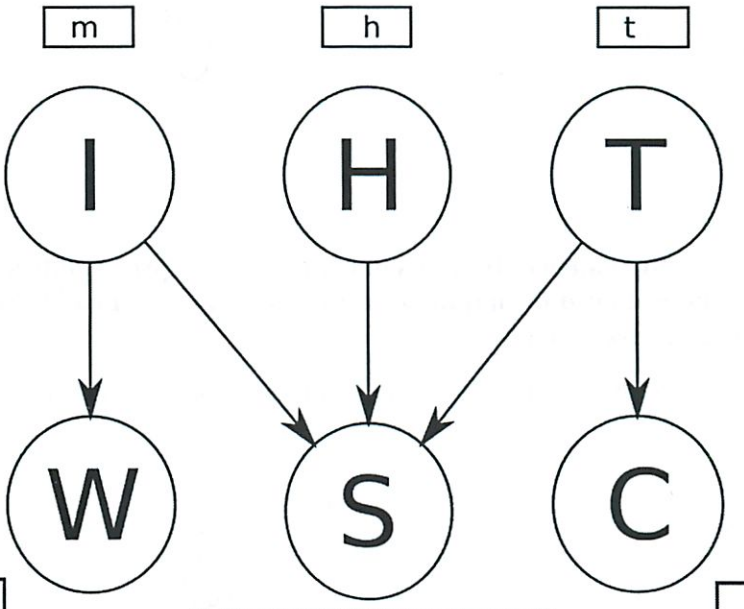
Correct $\frac{3}{4}w_i \to \frac{1}{24}$ inc $\frac{1}{4}$

12/15

Quiz 5, Problem 1, Probability (50 points)

Part A: Life Lessons in Probability (22 pts)

Consider the following inference net developed for students who graduate from MIT:



I	W
F	v
T	w

I	H	T	S
F	F	F	a
F	F	T	b
F	T	F	c
F	T	T	d
T	F	F	e
T	F	T	f
T	T	F	g
T	T	T	s

T	C
F	j
T	k

I = Had quality **I**nstruction

H = **H**ard working ethic

T = Raw **T**alent

S = **S**uccessful in life

C = **C**onfidence

W = Took 6.034

A1: What is the probability that a 6.034 student (W = true) had quality instruction (I = true) and became successful in life (S = true), but did not have raw talent (T = false) yet was hardworking H = true) and confident (C = true). Leave your answer unsimplified in terms of constants from the probability tables. **(6pts)**

not conditional - remember from tutorial

$P(W \wedge I \wedge S \wedge \neg T \wedge H \wedge C)$

$= P(W|I)\,P(I)\,P(S|I \wedge H \wedge \neg T)\,P(\cancel{I})\,P(H)\,P(\neg T)\,P(C|\neg T)$ T already have $P(\neg T)$

$= w \cdot m \cdot g \cdot h \cdot (1-x) \cdot j$ ✓ I already included

For A2-A3: Express your final answer in terms of expressions of probabilities that could be read off the Bayes Net. You do not need to simplify down to constants defined in the Bayes Net tables. You may use summations as necessary.

A2: What is probability of success in life (S = true) **given** that a student has high quality instruction (I = true)? **(6 pts)**

$P(S|I) = P(S|I \wedge H \wedge T)\,P(H)P(T)$

$\bar{H}\,T \quad \bar{H}\,T$

$H\,\bar{T} \quad H\,\bar{T}$

$\bar{H}\,\bar{T} \quad \bar{H}\,\bar{T}$

✓

A3: What is the probability a student is hardworking (H = true) , **given** that s/he was a 6.034 student (W = true)? **(10 pts)**

$P(H|W) = ind = P(H)$ ✓

Did in tutorial too :)

Part B: The Naïve Crime Fighter (12 pts)

A murder has occurred in quiet town of South Park. Your have been provided the following table by forensic experts:

Suspect	Has a motive	Location	Murder Weapon	Degree of Suspicion
Professor Chaos	0.4	Fair = 0.1 School = 0.7 CW= 0.2	Kindness = 0.3 Chili = 0.3 Moral Fiber = 0.2 Pure Evil = 0.2	0.3
Mint Berry Crunch	0.3	Fair = 0.4 School = 0.4 CW = 0.3	Kindness = 0.4 Chili = 0.1 Moral Fiber = 0.4 Pure Evil = 0.1	0.1
The Dark Lord Cthulu	0.1	Fair = 0.3 School = 0.3 CW = 0.4	Kindness = 0 Chili = 0.5 Moral Fiber = 0 Pure Evil = 0.5	0.4
Scott Tenorman	0.9	Fair = 0.8 School = 0.1 CW = 0.1	Kindness = 0.2 Chili = 0.5 Moral Fiber = 0.1 Pure Evil = 0.2	0.2

You have determined that the murder did not have a motive, the murder took place at the Fair, and the murder weapon was a bowl of chili. You've decided to use what you learned about Naïve Bayes to help you determine who committed the murder.

The murderer is most likely: $\arg\max_S P(S) P(\bar{M} \cap L=F \cap W=C \mid S)$

DLC

Show your work:

Chaos .3 · (1-.4) · .1 · .3 = .0054
MBC .1 (1-.3) · .4 · .1 = .0028
DLC .4 (1-.1) · .3 · .5 = .054 ← guilty
ST .2 (1-.9) · .8 · .5 = .008

New

Part C: Coin Tosses (16 pts)

You decide to reexamine the coin toss problem using model selection.

You have **2** types of coins:

1) **Fair:** P(heads) = 0.5
2) **All-heads:** P(heads) = 1.0

Oh still prob

You have **2** possible models to describe your observed sequences of coin types and coin tosses.

1) **Model 1:** You have both types of coins and you draw one of them at random, with equal probability, and toss it exactly once. You repeat both the drawing and tossing 3 times total. draw 3 toss 3
2) **Model 2:** You have both types of coins and you draw one of them at random, with equal probability, and toss it exactly 3 times. draw 1 toss 3

Finally, you have the following observed data.

Toss 1	Toss 2	Toss 3
H	H	T

The following questions use your knowledge of model selection to determine which is most likely.

You decide to use the following criterion to weigh models:

$$P(M)=\frac{1}{Z}\frac{1}{|parameters|}$$ where Z is a normalization constant.

No. didn't understand q

$|parameters|$ is defined as the number of cell entries in the CPTs of the Bayes Net representation.

C1. What is P(Model 1)? **(3 pts)** (Partial credit if you sketch the Models as Bayes Nets)

~~P(Model 1 | coin=fair)P(coin fair) + P(Model 1 | coin=Heads)~~ P(coin = heads)

C1 C2 C3 → T1 T2 T3

|params| = 3 + 6 = 9

P(model 1) = $\frac{1}{9z}$

What is P(Model 2)? **(3 pts)**

~~P(Model 2 | coin fair) P(coin fair) + P(Model 2 | coin heads) P(coin heads)~~

C → T1 T2 T3

|params| = 3·2 + 1 = 7

P(model 2) = $\frac{1}{7z}$

what was goal of that problems

So I think

3 ▢▢▢ vs 1 ▢

6 ▢▢▢ 9 vs 6 ▢▢▢ 7

but how is that P(M)

well they tell you - but doesn't make sense

You've decided that the a priori model probability P(Model) to use should be uniform.

P(Model 1) = P(Model 2) Right .5 each

Under this assumption you decide to work out the most likely model, given the data, P(Model | Data).

C2 What is the most-likely model based on this fully observed data set: **(10 pts)**

P(Model 1 | Data)? can be either coin

P(Model 1 | H∩H∩T) = Bayes rule = P(1∩H∩H∩T) / P(H∩H∩T) = P(H∩H∩T | 1) P(1) / P(H∩H∩T)

swap

So just P(Data | Model 1)

each flip ind

= P(data | 1) P(1) ← equal

no was right

P(Model 2 | Data)?

P(H∩H∩T | 2)

Therefore the most likely model is: (circle one) Model 1 Model 2

But if coin 1 p = .5
2 p = 1

equal prob

P(H | 1) P(1) + P(H | 2) P(2)
.5 · .5 + 1 · .5
.75

P(T | 1) P(1) + P(T | 2) P(2)
.5 .5 + 0
.25

So .75 · .75 · .25
.140625 ~~.5~~ / .5 ✓
c ← bigger

if coin 1

.5 · .5 · .5 have that seq

if coin 2

0 prob

So (.5³) · .5 + .5 · 0
.0625 · .5 / c ✓

So fairly figured it out - had to peak a bit - if thought harder would have just figured out

Quiz 5, Problem 2, Near Miss (20 points)

Having missed many tutorials, lectures, and recitations, Stew is stuck on trying to figure out who are the TAs in 6.034. You, who is more faithfully attending, knows who is who. Armed with your knowledge about Near Miss concept learning. You decide to build a model that will help Stew figure out who the TAs are.

? what was this again? -Oh quiz 3

The following table summarizes the training data about the current staff of 6.034. The **Title** attribute is organized as a tree, with MEng and PhDs both a type of Student. Students and Faculty are grouped under the type People-You-See-On-Campus

Name	TA	Hair Color	Title	Glasses	Gender	Experience # Years	Taken 6.034
Kendra	Yes	Brown	MEng	yes	female	1	Yes
Kenny	Yes	Brown	MEng	no	male	1	Yes
Martin	Yes	Black	MEng	no	male	1	Yes
Mark	Yes	Black	PhD	no	male	4	Yes
Mich	Yes	Blonde	MEng	yes	male	10	No
Gleb	Yes	Brown	MEng	no	male	2	Yes
Yuan	Yes	Black	PhD	yes	male	3	No
Lisa	No	Blond	Professor	yes	female	10	No
Bob	No	Brown	Professor	no	male	10	No
Randy	No	Brown	Professor	no	male	10	No

Fill in the table to build a model of a TA. Mark an attributes as "?" if the property has been dropped.

Example	Heuristics Used	Model Description TAs					
		Hair Color	Title	Glasses	Gender	Experience	Taken
Kendra	Initial Model	Brown	MEng	yes	female	1	Yes
Kenny							
Martin							
Yuan							
Bob							

Fill in the following table to build a model of a Faculty Member (FM). You may assume that Patrick, Randy, Bob, and Lisa are faculty members.

Example	Heuristics Used	Model Description FMs					
		Hair Color	Title	Glasses	Gender	Experience	Taken
Patrick	Initial Model	Blond	Professor	Yes	Male	10	No
Randy							
Bob							
Mich							
Lisa							

What class(es) would match these people given your TA model and your FM model. If neither, write N in the Class(es) column.

Name	Class(es)	Hair Color	Title	Glasses	Gender	Experience	Taken 6.034
Olga		Blond	MEng	no	female	1	Yes
Patricia		Blond	Professor	Yes	female	10	No

Quiz 5, Problem 3, Big Ideas (30 points)

Circle the **best** answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

1 Ullman's alignment method for object recognition

1. Is an effort to use neural nets to detect faces aligned with a specified orientation
2. Relies on a presumed ability to put visual features in correspondence
3. Uses A* to calculate the transform needed to synthesize a view
4. Uses a forward chaining rule set
5. None of the above

2 Ullman's intermediate-features method for object recognition

1. Is an effort to use boosting with a classifier count not too small and not too large
2. Is an example of the Rumpelstiltskin principle
3. Is a demonstration of the power of large data sets drawn from the internet
4. Uses libraries containing samples (such as nose and mouth combinations) to recognize faces
5. None of the above

3 The SOAR architecture is best described as

1. A commitment to the strong story hypothesis
2. A commitment to rule-like information processing
3. An effort to build systems with parts that fail
4. The design philosophy that led to the Python programming language
5. None of the above

4 The Genesis architecture (Winston's research focus) is best described as, in part, as

1. A commitment to the strong story hypothesis
2. Primarily motivated by a desire to build more intelligent commercial systems
3. A commitment to rule-like information processing
4. A belief that the human species became gradually smarter over 100s of thousands of years.
5. None of the above

5 A transition frame

1. Focuses on movement along a trajectory
2. Focuses on the movement from childlike to adult thinking
3. Focuses on a small vocabulary of state changes
4. Provides a mechanism for inheriting slots from abstract frames, such as the disaster frame
5. None of the above

6 Reification is

1. The attempt to develop a universal representation
2. The tendency to attribute magical powers to particular mechanisms
3. The process by which ways of thinking are determined by macro and micro cultures
4. The process of using perceptions to answer questions too hard for rule-based systems
5. None of the above

7 Arch learning includes

1. A demonstration of how to combine the benefits of neural nets and genetic algorithms
2. A commitment to bulldozer computing using 100's of examples to learn concepts
3. The near miss concept
4. A central role for the Goldilocks principle
5. None of the above

8 Arch learning benefits importantly from

1. An intelligent teacher
2. Exposure to all samples at the same time
3. Use of crossover
4. Sparse spaces
5. None of the above

9 Experimental evidence indicates

1. People who talk to themselves more are better at physics problems than those who talk less
2. Disoriented rats look for hidden food in random corners of a rectangular room
3. Disoriented children combine color and shape information at about the time they start walking
4. Disoriented children combine color and shape information at about the time they start counting
5. None of the above

10 Goal trees

1. Enable rule-based systems to avoid logical inconsistency
2. Enable rule-based systems answer questions about behavior
3. Are central to the subsumption architecture's ability to operate without environment models
4. Are central to the subsumption architecture's ability to cope with unreliable hardware
5. None of the above

Name	6034 TAs
email	6034tas@csail.mit.edu

6.034 Final Examination
December 15, 2010

Circle your TA and principle recitation instructor so that we can more easily identify with whom you have studied:

Martin Couturier	Kenny Donahue	Gleb Kuznetsov
Kendra Pugh	Mark Seifter	Yuan Shen

Robert Berwick	Randall Davis	Lisa Fisher

Indicate the approximate percent of the lectures, mega recitations, recitations, and tutorials you have attended so that we can better gauge their correlation with quiz and final performance and with attendance after OCW video goes on line. Your answers have no effect on your grade.

	Lectures	Recitations	Megas	Tutorials
Percent attended	100	100	100	100

Quiz	Score	Grader
Q1	100	
Q2	100	
Q3	100	
Q4	100	
Q5	100	

There are 48 pages in this final examination, including this one. In addition, tear-off sheets are provided at the end with duplicate drawings and data. As always, open book, open notes, open just about everything.

Quiz 1, Problem 1, Rule Systems (50 points)

Kenny has designed two suits for the Soldier Design Competition, and he and Martin like to grapple in the prototypes on Kresge Oval.

- Kendra insists the suits qualify as **"deadly weapons"** and Kenny should give them to her for safekeeping.
- Kenny and Martin insist that they are examples of an **"enhanced prosthesis"** and that they should be able to keep them

The TAs decide to use Rule-Based Systems to resolve their dispute.

Rules:

P0	IF (AND ('(?x) is a Crazy Physicist', '(?x) is an Engineer') THEN ('(?x) builds a Weaponized Suit'))
P1	IF ('(?y)'s (?x) is Cardinal Red' THEN ('(?y)'s (?x) is not US Govt. Property'))
P2	IF (OR (AND ('(?y) is an Engineer', '(?y)'s (?x) is Really Heavy'), '(?y)'s (?x) is stolen by the Air Force') THEN ('(?y)'s (?x) is a Deadly Weapon'))
P3	IF (OR ('(?y) is not evil', '(?y) is a Roboticist') THEN ('(?y)'s suit research is not Evil'))
P4	IF (AND ('(?y)'s (?x) research is not Evil', '(?y)'s (?x) is not US Govt. Property') THEN ('(?y)'s (?x) is an Enhanced Prosthesis'))

Assertions:

A0: (Kenny is a Roboticist)

A1: (Martin is an Engineer)

A2: (Kenny's suit is Cardinal Red)

A3: (Martin's suit is Really Heavy)

Part A: Backward Chaining (30 points)

Make the following assumptions about backward chaining:

- The backward chainer tries to find a matching assertion in the list of assertions. If no matching assertion is found, the backward chainer tries to find a rule with a matching consequent. In case none are found, then the backward chainer assumes the hypothesis is false.
- The backward chainer never alters the list of assertions; it never derives the same result twice.
- Rules are tried in the order they appear.
- Antecedents are tried in the order they appear.

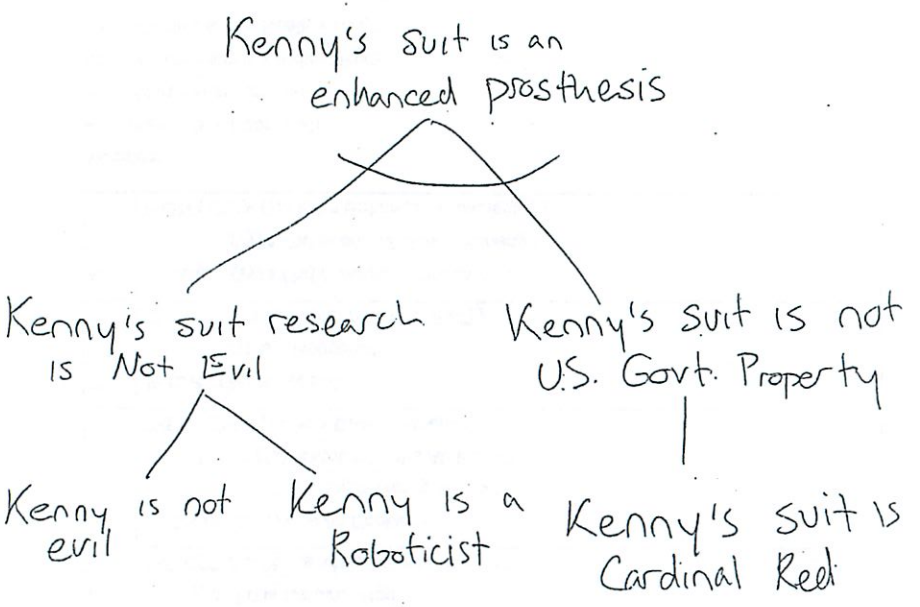
Simulate backward chaining with the hypothesis

Kenny's suit is an enhanced prosthesis

Write all the hypotheses the backward chainer looks for in the database in the order that the hypotheses are looked for. The table has more lines than you need. We recommend that you use the space provided on the next page to draw the goal tree that would be created by backward chaining from this hypothesis. **The goal tree will help us to assign partial credit** in the event you have mistakes on the list.

1	Kenny's suit is an enhanced prosthesis
2	Kenny's suit research is not Evil
3	Kenny is not evil
4	Kenny is a Roboticist
5	Kenny's suit is not U.S. Govt. Property
6	Kenny's suit is Cardinal Red
7	
8	
9	
10	

Draw Goal Tree Here for Partial Credit



Part B: Forward Chaining (20 points)

Let's say, instead, our assertions list looked like this:

A0: Gleb is an Engineer
A1: Gleb's laptop is Really Heavy
A2: Gleb's suit is Really Heavy

B1 (4 points)

CIRCLE any and all rules that match in the first iteration of forward chaining

P0	P1	P2	P3	P4

B2 (4 points)

What assertion(s) are added or deleted from the database, as a consequence of this iteration?

Gleb's laptop is a Deadly Weapon

B3 (4 points)

CIRCLE any and all rules that match in the second iteration of forward chaining

P0	P1	P2	P3	P4

B4 (4 points)

What assertion(s) are added or deleted from the database, as a consequence of this iteration?

Gleb's suit is a Deadly Weapon

B5 (4 points)

You take the same assertions as at the beginning of problem B, above, and re-order them:

A0: Gleb is an Engineer
A1: Gleb's suit is Really Heavy
A2: Gleb's laptop is Really Heavy

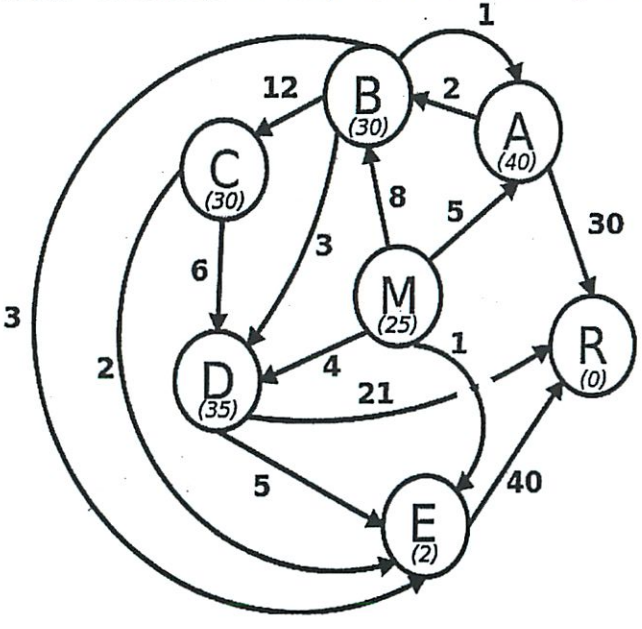
Now, you start over, and run forward chaining from the beginning, until no new assertions are added to or deleted from the database. Is Gleb's laptop a Deadly Weapon?

Yes

Quiz 1, Problem 2, Search (50 points)

As you get close to graduating MIT, you decide to do some career planning. You create a graph of your options where the start node is **M = MIT** and your goal node is **R = Retire**, with a bunch of options in between. Your graph includes edge distances that represent, roughly, the "cost of transition" between these careers (don't think too hard about what this means). You also have heuristic node-to-goal distances which represent your preconceptions about how many more years you have to work until you retire. For example, you think it will take 25 years to go from MIT (M) to retirement (R), 30 years from Grad School (B), but only 2 years from Entrepreneur (E).

A = Wall Street | B = Grad School | C = Professor | D = Government | E = Entrepreneur



Part A: Basic search (25 points)

In all search problems, use alphabetical order to break ties when deciding the priority to use for extending nodes.

A1 (3 points)

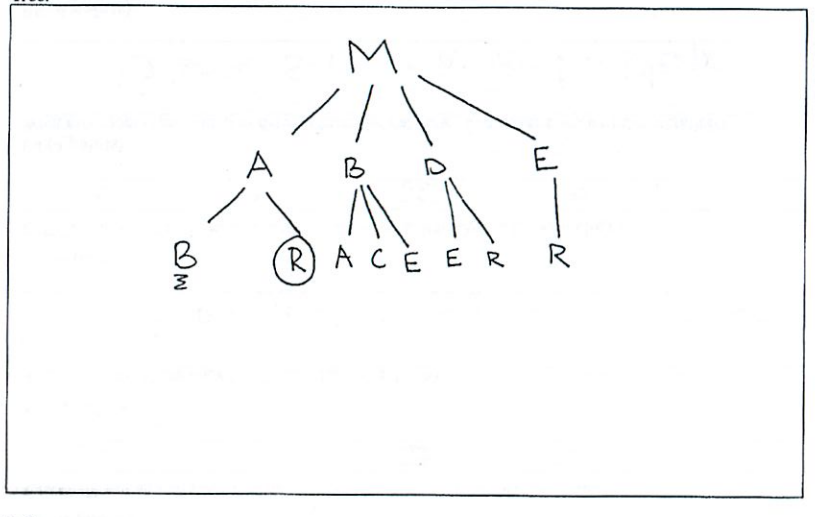
Assume you want to retire after doing the least number of different jobs. Of all the basic search algorithms you learned about(that is, excluding branch and bound and A*) which one should you apply to the graph in order to find a path, **with the least search effort**, that has the minimum number of nodes from M to R?

Breadth First Search

A2 Basic Search Chosen Above (7 points)

Perform the search you wrote down in A1 (with an Extended List). Draw the search tree and give the final path.

Tree:



x5

Path:

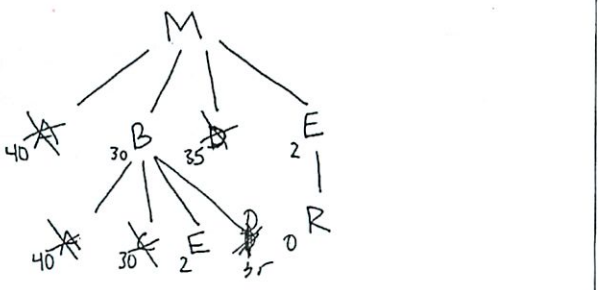
M - A - R

x2

A3 Beam Search with w=2 (15 points)

Now you are interested in finding a path and the associated distance. Try a **Beam Search** with a width **w=2**, ***with*** an extended list. As before, you are looking for a path from **M** to **R**. Use the "preconceptions" heuristic distances indicated in parentheses at each node.

Tree:



Path, if any:

M - E - R

Extended nodes in order extended:

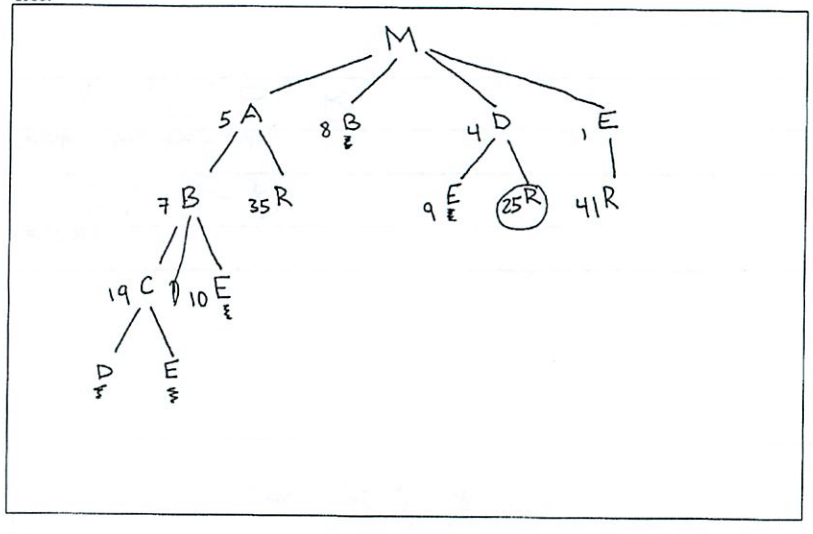
M, E, B, R

Part B: Advanced Search (25 points)

B1 Branch and Bound with Extended List (15 points)

Use Branch and Bound search with an Extended List to find a path from **M** to **R**, as well as the extended node list. Use this space to draw the corresponding tree and show your work.

Tree:



10

Path:

M - D - R

c

3

Extended nodes in order extended:

M, ~~E~~, D, A, B, C, R

B2 Thinking about Search (6 points)

Concisely give the reason why Branch and Bound with Extended List yields a different result than Beam Search in this problem.

Heuristics pull path toward local minimum and beam prunes other, potentially shorter, paths. x2

What can we say about the path found by Branch and Bound with Extended List? (We're looking for a fairly strong statement here.)

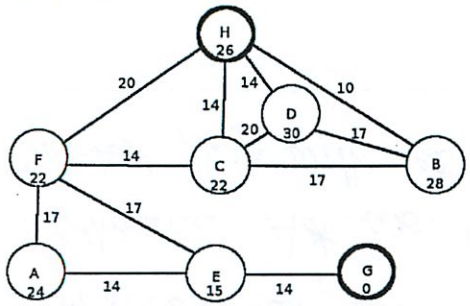
Shortest Path x2

Is there an algorithm guarantees the same answer as Branch and Bound ***for the graph in this problem***, but can find the answer with fewer extended paths. If Yes, what is that algorithm? If No, state the possible problem.

No, the heuristic is neither admissable nor consistent. (Otherwise A* would give Shortest path with less extended paths) x2

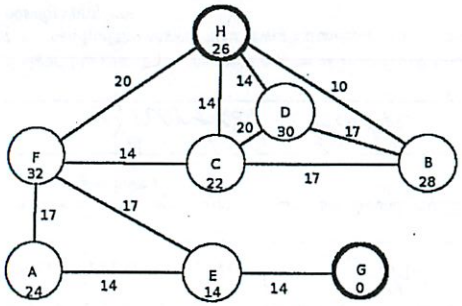
B3 Permissible Heuristics (6 points)

Suppose you are asked to find the shortest path from H to G in the graphs below. For both of the graphs explain why the heuristic values shown are not valid for A*. Note the differences in the graphs at nodes F and E.



Reason(s):

not admissable (e.g. E → G)



Reason(s):

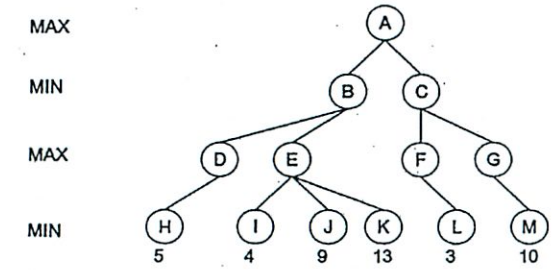
not consistent (e.g. F → E)

Quiz 2, Problem 1, Games (50 points)

Part A: Basics(15 points)

A1 Plain Old Minimax(7 points)

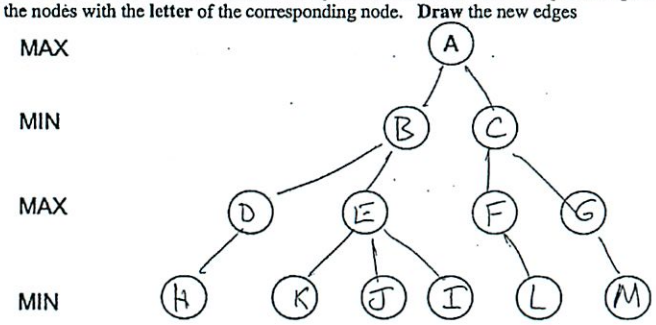
Perform Minimax on this tree. **Write the minimax value associated with each node in the box below, next to its corresponding node letter.**



A= 5 B= 5 C= 3 D= 5 E= 13 F= 3 G= 10

A2 Tree Rotations (8 points)

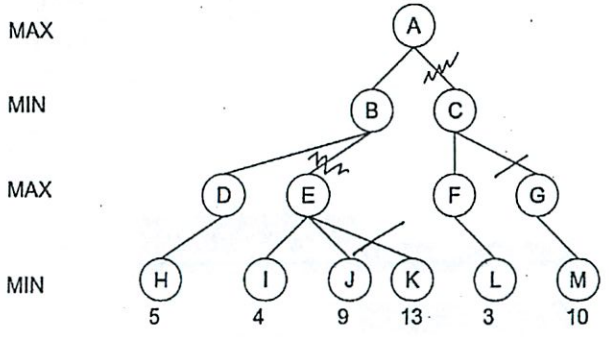
Using the minimax calculations from part A1, **without** performing any alpha-beta calculation, rotate the children of each node in the above tree at every level to ensure maximum alpha-beta pruning.
- **Fill in** the nodes with the **letter** of the corresponding node. **Draw** the new edges



Part B: Alpha Beta (35 points)

B1: Straight-forward Alpha Beta(15 points)

Perform Alpha Beta search on the following tree.
- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

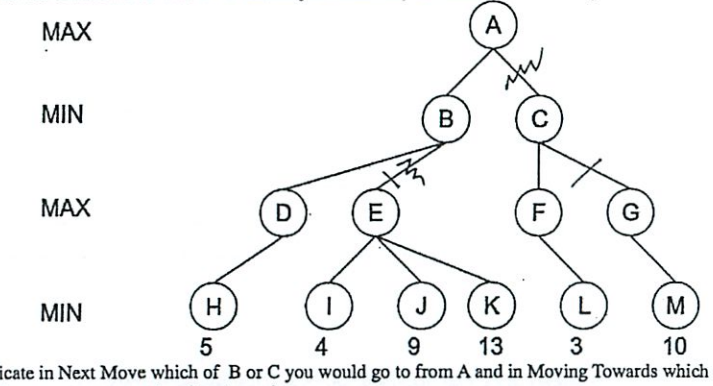
# of evaluations: 4 **List:** H I J L

Next Move: B **Moving towards:** H

B2: Preset Alpha-Beta (15 points)

Perform alpha-beta search, using **initial values of alpha = 5 and beta = 8.**

- **Indicate** pruning by striking through the appropriate edge(s).
- **Mark** your steps for partial credit.
- **Fill in** the number of static evaluations.
- **List** the leaf nodes **in the order** that they are statically evaluated.



Indicate in Next Move which of B or C you would go to from A and in Moving Towards which node in the bottom row you are heading toward.

# of evaluations: 2 **List:** H L

Next Move: B **Moving towards:** H

B3: Alpha-Beta Properties (5 points)

If you were able to maximally prune a tree while performing Alpha-Beta search, approximately how many static evaluations would you end up doing for a tree of depth *d* and branching factor *b*?

$O(b^{\frac{d}{2}})$

Quiz 2, Problem 2, Constraint Propagation (50 points)

After taking 6.034, your decide to offer your services to incoming freshman to help them set up their course schedules. One particular freshman comes to you with four classes as well as an availability schedule (grayed out boxes represent reserved times).

Course	Lecture Times Offered	Recitation Times Offered
3.091	MWF 11,12	TR 10,11,12,1
18.01	MWF 12, 1	TR 11,1,2
8.01T	MWF 10, 11, 12, 1, 2, 3	NONE
21F.301	MTWRF 10, 11	NONE

Time	MWF	TR
10		
11		
12		
1		
2		
3		

For easier bookkeeping you adopt the folowing naming convention (L = Lecture, R = Recitation, # = course number):

3.091 Lecture	→ **L3**	MWF10	→**10M**
3.091 Recitation	→ **R3**	MWF11	→**11M**
8.01T Lecture	→ **L8**		
18.01 Lecture	→ **L18**	TR10	→**10T**
18.01 Recitation	→ **R18**	TR11	→**11T**
21F.301 Lecture	→ **L21**		

You also devise this set of constraints for yourself:

(1) Each class must be assigned to exactly one timeslot
(2) Each timeslot can be assigned to a maximum of one class
(3) No classes can be scheduled during the grayed out time periods
(4) The TR selection for 21F.301 must occur at the same time as the MWF selection.

Part A: Picking a Representation (8 points)

In order to fill in this schedule, you decide to set it up as a CSP using meeting times as variables and courses as the values of in the domains. After filling in the domain table, this is what you see:

Variable	Domain
10M	L8, L21
11M	L3, L8, L21
12M	L3, L18, L8
[illegible]	
[illegible]	
3M	L8
10T	R3, L21
[illegible]	
12T	R3
[illegible]	
2T	R18
3T	

What is wrong with the way that this problem is set up and why?

VARIABLE 3T CANNOT HAVE A VALUE

Part B: Applying Constraints (42 points)

You decide to switch to a new representation that uses the courses as variables and the times as values.

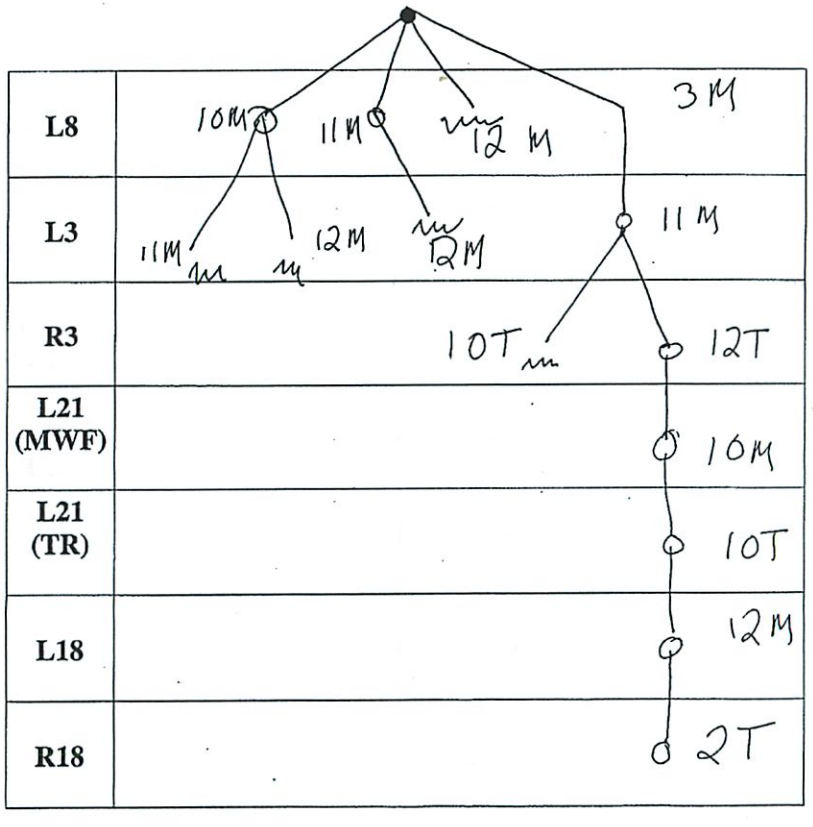
B1 (5 points)

The initial domains are given below. Cross out the values that are incompatible with Constraint (3).

Variable	Domain
L8	10M 11M 12M ~~1M~~ ~~2M~~ 3M
L3	11M 12M
R3	10T ~~11T~~ 12T ~~1T~~
L21 (MWF)	10M 11M
L21 (TR)	10T ~~11T~~
L18	12M ~~1M~~
R18	~~1T~~ ~~1T~~ 2T

B2 (16 Points)

Run the DFS with **forward checking only** on your defined variables and the reduced domains you found in Part B1 by applying Constraint(3).



B3 (5 Points)

How many times did the algorithm need to backtrack?

5

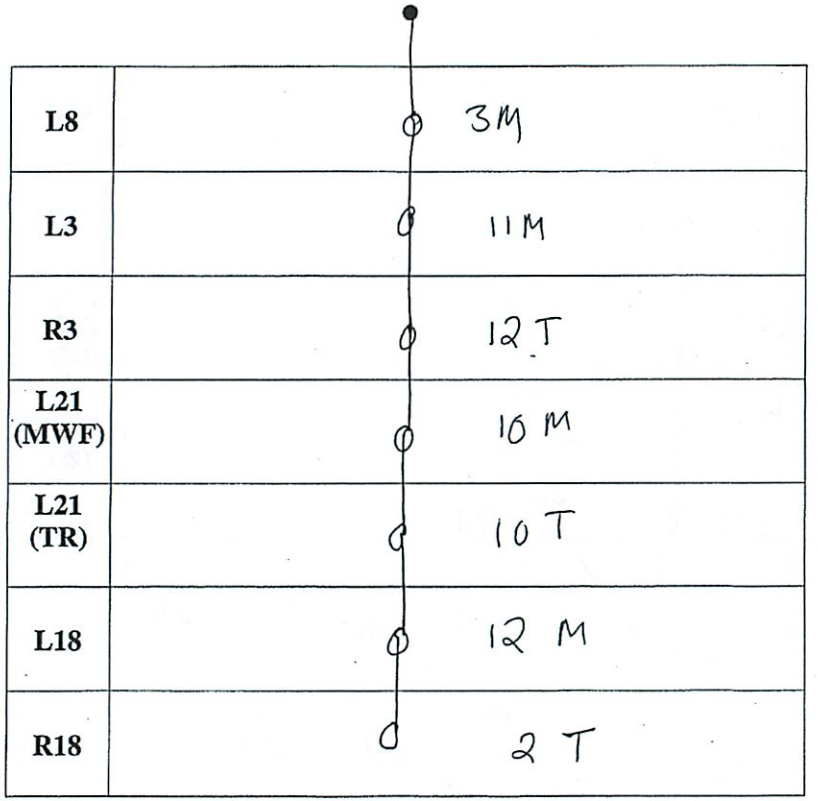
B4 (10 Points)

It occurs to you that you may be able to accelerate the process of finding a solution if you were to perform domain reduction with **propagation through singletons** *before* running the DFS. Fill in your updated domain table with the results of your computations.

Variable	Domain
L8	3M
L3	11M
R3	12T
L21 (MWF)	10M
L21 (TR)	10T
L18	12M
R18	2T

B5 (6 Points)

Run DFS with **constraint checking only** on your updated domain table:



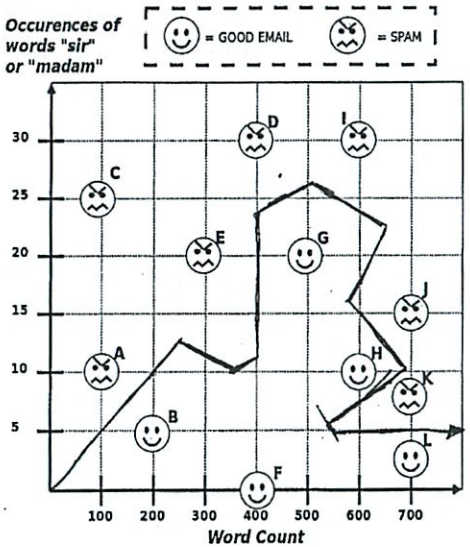
Quiz 3, Problem 1
KNN and ID Trees (50 points)

After receiving yet another "Dear sir or madam.." email, you decide to construct a spam filter.

Part A: Nearest Neighbors (25 points)

For your first attempt you decide to try using a k Nearest Neighbors model. You decide to classify spam according to 2 features: email word count and occurrences of the words "sir" or "madam".

A1 (10 points)



Draw the decision boundary for 1-nearest-neighbor on the above diagram of the given training data. **Use the center of the faces as the positions of the training data points.**

A2 (8 points)

How will **1-nearest-neighbor** classify an email with **200** words of which **9** are the word "sir"?
Plot this point on the graph as X? (2pts)

GOOD EMAIL

How will **3-nearest-neighbors** classify an email with **600** words of which **7** are the word "madam"?
Plot this point on the graph as Y? (3pts)

GOOD EMAIL

How will **5-nearest-neighbors** classify an email with **500** words of which **25** are the word "madam"?
Plot this on the graph as Z? (3pts)

SPAM

A3 (7 points)

List which points yield errors when performing leave-one-out cross validation using **1-nearest-neighbor** classification. (3 pts)

A, B, E, G, H, J, K, L

How would one go about selecting a good **k** to use? (4 pts)

You can perform leave-one-out cross validation for different values of K to determine which K yields the least number of errors.

Part B: ID Trees (25 points)

Realizing nearest neighbors may not be the best tool for building a spam filter, you decide to try another classifier you learned about in 6.034: Identification Trees.

B1 (8 points)

It appears that the over-use of the words "sir or madam" seems to be a strong hint at an email being spam.

What is the minimum disorder and minimum-disorder decision boundary when you consider only the dimension of "Sir or Madam"? You can use fractions, real numbers, and logarithms in your answer.

Approximate boundary:

"sir or madam" < 6

Associated Disorder:

$$\frac{9}{12}\left(-\frac{2}{9}\log_2\frac{2}{9} - \frac{7}{9}\log_2\frac{7}{9}\right)$$

B2 (8 points)

Suppose we were given the following additional information about our training set:

Emails I, G, J, H, K and L are from Anne Hunter.

One of those emails might be important so you don't want to risk missing a single one so you re-label all Anne Hunter emails in the training data set to be good emails. You are to find the best axis-parallel test given the revised labellings of good email and spam.

(NOTE: Use the unlabeled graphs in the tear-off sheets section if you need it to visualize the modified data).

B2.1 Which emails does your new test missclassify on the **modified** data? (4pts)

B, F

(word count > 450)

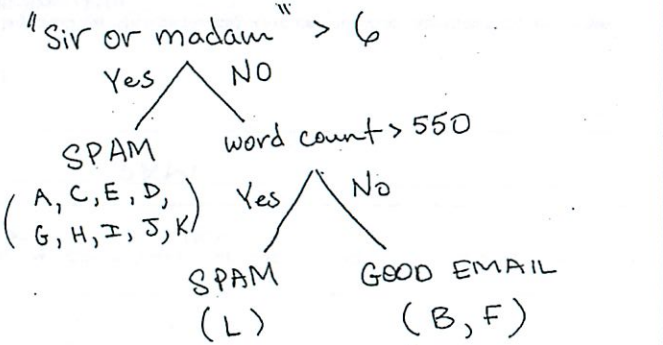
B2.2 What is the disorder of your new test on the **modified** training data set? Leave your answer as a function of fractions, real numbers, and logarithms. (4pts)

$$\frac{6}{12}\left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right)$$

B3 (9 points)

Soon, you decide that your life goal is no longer to be a tool for a Harvard or Sloanie startup so you decide that **all emails from Anne Hunter should be marked as spam.** (Again, use the **unlabeled graphs in the tear-off sheets if you need them).**

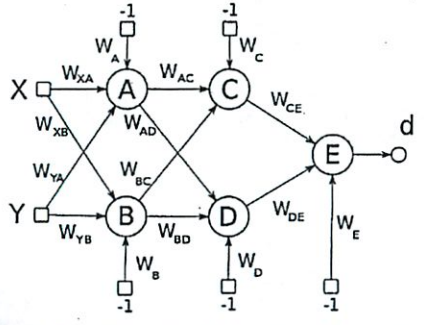
Given the revised determination of what is good email and spam, draw the **disorder minimizing identification tree** that represents your fully trained ID-tree spam filter. You may use any horizontal and vertical classifiers in the dimensions of word count and "sir or madam" occurrences. Ties should be broken in order of horizontal then vertical classifiers.



Quiz 3, Problem 2, Neural Nets (50 Points)

Note that this problem has three completely independent parts.

Part A: Derivations (14 pts)



A1. (7 pts) Using what you've learned from doing lab 5, write out the equation for $\frac{dP}{dw_{CE}}$ expressed in terms of o_i, d, and/or any weights and constants in the network. (o_i refers to the output of any neuron in the network.)

$$\frac{dP}{dw_{CE}} = \underbrace{(d - O_E) \cdot O_E(1-O_E) \cdot O_C}_{\frac{dP}{dO_E}} \cdot \underbrace{\frac{dO_E}{dz_E}}_{} \cdot \underbrace{\frac{d[w_{CE}O_C + w_{DE}O_D - w_E]}{dw_{CE}}}_{z_E}$$

2 pts each component

If had chain rule +3 pts but wrong

A2. (7 pts) Write out the equation for $\frac{do_E}{dw_{XA}}$ expressed in terms of $\frac{do_i}{dw_{XA}}$, o_i, and/or any weights and constants in the network. **NOTE: o_i may not be o_E!**

$$\frac{dO_E}{dw_{XA}} = \frac{dO_E}{dz_E} \cdot \frac{dz_E}{dw_{XA}} \quad \} \text{ chain rule +1}$$

$$= O_E(1-O_E) \frac{d[w_{CE}O_C + w_{DE}O_D - w_E]}{dw_{XA}}$$

$$= O_E(1-O_E)\left[w_{CE}\frac{dO_C}{dw_{XA}} + w_{DE}\frac{dO_D}{dw_{XA}} \right]$$

2 pts 2 pts 2 pts

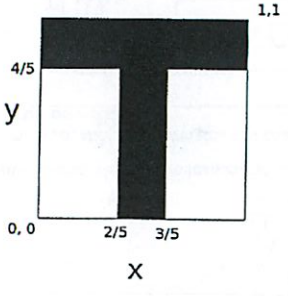
If completely wrong but had some semblance of chain rule +3 pts

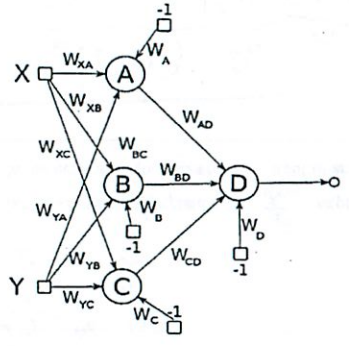
Part B: Letter Recognition (20 pts)

You propose to use a neural network to recognize characters from a scanned page. Letters are represented binary images on a 1x1 unit grid. Assume that scaling and rotation are all done.

Because you want to start with something easy, you start with the problem of recognizing a character as either possibly a T or definitely not a T. During training, each training sample consists of a random point, (x, y), along with the desired 0 or 1 value: 1 if the underlying pixel at (x, y) is part of a T; 0 if the pixel is part of a T's background.

You want to find the most **compact** network that will correctly handle the T problem, so you decide to analytically work out the minimal network that will correctly classify a character as possibly a T or definitely not a T.





Assume you decide to have the above network architecture, fill in the 7 missing weights in the table that are required to accurately classify all points in the image for T. Your weights must be integer weights or integer constraints on weights! Show your work for partial credit:

Answers:

W_{XA}	0	W_{XC}	-5
W_{YA}	5	W_{YC}	0
W_A	4	W_C	-3
W_{XB}	5	W_{AD}	4 (or >3)
W_{YB}	0	W_{BD}	2
W_B	2	W_{CD}	2
		W_D	3

Show work here for partial credit:

① We have 3 regions represented by neurons A, B, C. Their line eqns are:

$y > 4/5 \quad x > 2/5 \quad x < 3/5$

② Next express eqn in a more familar sum of weights form.

$0x + 1y > 4/5$
$1x + 0y > 2/5$
$-1x + 0y > -3/5$

③ multiply by 5 to get integer weights

$0x + 5y > 4$
$5x + 0y > 2$
$-5x + 0y > -3$

④ Recall that generic binding equation for neurons are

$W_{xn}x + W_{yn}y > W_n \quad n \in A, B, C.$

⑤ By correspondence and constraints given

$(W_{xA}=0)x + (W_{yA}=5)y > (W_A=4)$ Neuron A ↑
$(W_{xB}=5)x + (W_{yB}=0)y > (W_B=2)$ Neuron B ←
$(W_{xC}=-5)x + (W_{yC}=0)y > (W_C=-3)$ Neuron C →

For Neuron D; it must be doing some logic on A, B, C.
Namely A or (B and C)

Use Logic table

	A	B	C	output
	0	0	0	0
	0	0	1	0
	0	1	0	0
	0	1	1	1
①	1	0	0	1
②	1	0	1	1
③	1	1	0	1
④	1	1	1	1

expand only these since they have W_{AD}.

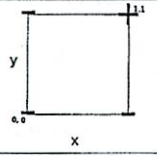
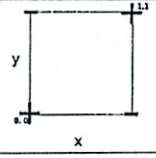
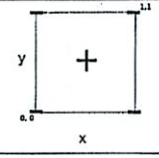
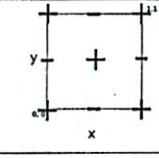
① $W_{AD} > -3 > 0$
② $W_{AD} + 2 > -3 > 0$
③ $W_{AD} + 2 > -3 > 0$
④ $W_{AD} + 2 + 2 > -3 > 0$

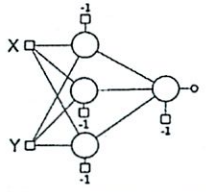
① $W_{AD} > 3$ ← strongest constraint
② $W_{AD} > 1$
③ $W_{AD} > 1$ so $W_{AD} = 4$
④ $W_{AD} > -1$ n/a >3

Partial credit breakdown. +3 for anything resembling steps ①-④ +3 for a truth table

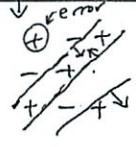
Part C: Expressive Power of Neural Nets (16 pts)

Circle all the functions that these networks are theoretically able to fully learn, and if your answer is No indicate the lowest possible error rate. **List of Functions:**

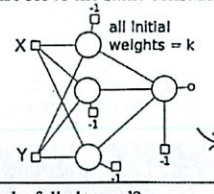
X AND Y	X = Y	X = 0.5 AND Y = 0.5	X-Shape
			



C1	Can be fully learned?	The minimum # error if No
X AND Y	(Yes) No	
X = Y	(Yes) No	
X = 0.5 AND Y = 0.5	(Yes) No	
X Shape	Yes (No)	1

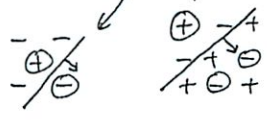


How about when all initial weights are set to the same constant k?



becomes 1 neuron because of symmetry.

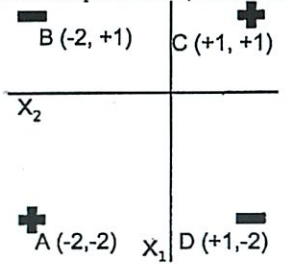
C2	Can be fully learned?	The minimum # error if No
X AND Y	(Yes) No	
X = Y	Yes (No)	1
X = 0.5 AND Y = 0.5	Yes (No)	2
X Shape	Yes (No)	3



Quiz 4, Problem 1, Support Vector Machines (50 points)

Part A: Solving SVMs (40 pts)

You decided to manually work out the SVM solution to the CORRELATES-WITH function. First you reduce your problem to looking at four data points in 2D (with axes of x_1 and x_2).



You ultimately want to come up with a classifier of the form below. **More formulas are provided in tear off sheets.**

$$h(\vec{x}): \sum a_i y_i K(\vec{x}, \vec{x}_i) + b \geq 0$$

Your TA, Yuan suggests that you use this kernel:

$$K(\vec{u}, \vec{v}) = u_1 v_1 + u_2 v_2 + |u_2 - u_1||v_2 - v_1|$$

A1(5 pts): Note that $\Phi(\vec{u})$ is a vector that is a transform of $\vec{u}$ and the kernel is the dot product, $\Phi(\vec{u}) \cdot \Phi(\vec{v})$. Determine the number of dimensions in $\Phi(\vec{u})$, and then write the components of $\Phi(\vec{u})$ in terms of $\vec{u}$'s x_1 and x_2 components, u_1 and u_2. Explain why it is better to use $\Phi(\vec{u})$ rather than $\vec{u}$.

$\Phi(\vec{u})$'s components: $[\ u_1 \quad u_2 \quad |u_2 - u_1|\ ]$

Why better: We are converting a 2D problem into 3D.
A linearly unseperable problem in 2D is now linearly seperable in 3D

A2(10 pts): Fill in the unshaded portion in the following table of Kernel values.

K(A,A) = 4+4·0·0=8	K(A,B) = 4-2+0·3=2	K(A,C) = -2-2+0·0=-4	K(A,D) = -2+4+0·3=2
K(B,A) =	K(B,B) = 4+1+3·3=14	K(B,C) = -2+1+3·0=-1	K(B, D) = -2-2+3·3=5
K(C,A) =	K(C,B) =	K(C,C) = 1+1+0·0=2	K(C, D) = 1-2+0·3=-1
K(D,A) =	K(D,B) =	K(D,C) =	K(D, D) = 1+4+3·3=14

A3 (10 pts): Now write out the full **constraints equations** you'll need to solve this SVM problem They should be in terms of α_i, b, and constants. **(Hint:** The four data points lie on their appropriate gutters).

	Constraint Equations ($y_A \cdot k(A,A)$, $y_B k(A,B)$)
1	$h(\vec{A}) = +8\,\alpha_A + -2\,\alpha_B + -4\,\alpha_C + -2\,\alpha_D + b = +1$ (from +ve gutter eqn)
2	$h(\vec{B}) = 2\,\alpha_A + -14\,\alpha_B + -1\,\alpha_C + -5\,\alpha_D + b = -1$ (from -ve gutter)
3	$h(\vec{C}) = -4\,\alpha_A + 1\,\alpha_B + 2\,\alpha_C + 1\,\alpha_D + b = +1$ (from +ve gutter)
4	$h(\vec{D}) = 2\,\alpha_A + -5\,\alpha_B + -1\,\alpha_C + -14\,\alpha_D + b = -1$ (from -ve gutter)
5	$+1\,\alpha_A \quad -1\,\alpha_B \quad +1\,\alpha_C \quad -1\,\alpha_D + 0b = 0$ (from $\frac{\partial L}{\partial b}=0$)

A4 (5 pts): Instead of solving the system of equations for alphas, suppose the alphas were magically given to you as:

α_A=1/9	α_B=1/9	α_C=1/9	α_D=1/9	b=1

Compute $\vec{w}$ Given your alphas. **Hint: The number of dimensions in $\vec{w}$ is the same as the number of dimensions of $\Phi(\vec{x})$**

$\vec{w} = \sum_i \alpha_i y_i \Phi(\vec{x}_i)$ (from $\frac{\partial L_D}{\partial \vec{w}}=0$ for decompose Kernels)

$$\vec{w} = \frac{1}{9}\cdot(+1)\begin{bmatrix}-2\\-2\\0\end{bmatrix} + \frac{1}{9}(-1)\begin{bmatrix}-2\\1\\3\end{bmatrix} + \frac{1}{9}(+1)\begin{bmatrix}1\\1\\0\end{bmatrix} + \frac{1}{9}(-1)\begin{bmatrix}+1\\-2\\3\end{bmatrix} = \begin{bmatrix}0\\0\\-2/3\end{bmatrix}$$

A5 (5 pts): What is the equation of the optimal SVM decision boundary using the answers and values from A5?

$h(\vec{x}) = \vec{w}\cdot\Phi(\vec{x}) + b = [0, 0, -\frac{2}{3}]\cdot[x_1, x_2, |x_2 - x_1|] + b$

$h(\vec{x}) = -\frac{2}{3}|x_2 - x_1| + 1$

A6 (5 pts): What is the width of the road defined by the optimal SVM decision boundary?

Width = $\frac{2}{\|\vec{w}\|} = \frac{2}{\sqrt{(\frac{2}{3})^2}} = 3$

Part B: Kernel for k-Nearest Neighbors (10 pts)

A student noticed that Kernel methods can be used in other classification methods, such as k-nearest neighbors. Specifically, one could classify an unknown point by summing up the fractional votes of all data points, each weighted using a Radial Basis Kernel. The final output of an unknown point $\vec{x}$ depends on the sum of **weighed** votes of data points $\vec{x}_i$ in the training set, computed using the function:

$$K(\vec{x},\vec{x}_i)=\exp\left(\frac{-\|\vec{x}-\vec{x}_i\|}{s^2}\right)$$

Negatively labeled data points contribute -1 times the kernel value and positively labeled training data points contribute +1 times the kernel value.

You may find the graphs provided in the tear off sheets helpful.

B1 (6 pts)

Using this approach, as s increases to infinity, what k does this correspond to in k-NN?

K = n a n-Nearest neighbors where n = total # of data points
Because when $s \to \infty$, Gaussian goes fm $\wedge \to$ ——, So all points get equal votes

As s decreases to zero, what k does this approximately correspond to in k-NN?

K = 1 or 1-Nearest neighbor, Because when $s \to 0$
Gaussian goes fm $\wedge \to \wedge$ (sharp), and only closest points get strong votes.

B2 (4 pts):

State a reason why you would prefer to use the SVM with Radial Basis Kernel solution rather than the method of Weighted Nearest Neighbors with a Gaussian Kernel.

Eqn for weighted N.N. $h(\vec{x}) = \sum_i y_i K(\vec{x}_i, \vec{x})$ $i \in$ all n points
Eqn fn SVM boundry $h(\vec{x}) = \sum_i \alpha_i y_i K(\vec{x}_i, \vec{x})$ $i \in$ only support vectors

#SV $\ll$ n.

possible Reasons:

① performance, classifying a new point require $O(\#SV)$ time vs $O(n)$ time kernel calculations so SVM soln is faster.

② SVM soln has less terms, so a shorter/ more compact soln → AKA Occam's Razor.

③ Convenience; SVM soln only need SV points vs all training points. If it is a product, it is more portable. Imagine putting all your train data in a cell phone, not a good idea.

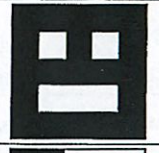
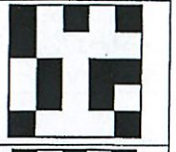
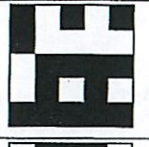
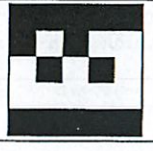
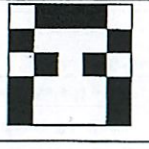
Quiz 4, Problem 2, Boosting (50 points)

Kenny wants to recognize faces in images. He comes up with a few things that he thinks will probably work well as weak classifiers and decides to create an amalgam classifier based on his training set. Then, given an image, he should be able to classify it as a FACE or NOT FACE. When matched against faces, the GRAY part of a classifier can be either WHITE or BLACK)

Classifiers:

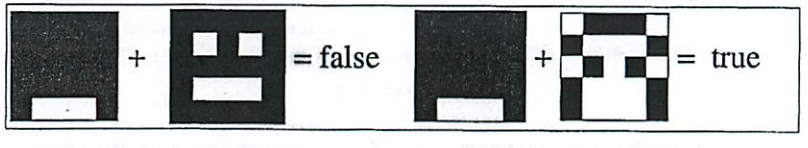
Name	Image Representation
A Has Hair	
B Has Forehead	
C Has Eyes	
D Has Nose	
E Has Smile	
F Has Ear	

Data:

Index	Classification	Image Representation	Index	Classification	Image Representation
1	NOT FACE		5	FACE	
2	NOT FACE		6	FACE	
3	NOT FACE		7	FACE	
4	FACE		8	FACE	

Examples:

Here is how classifier E works. Note that gray means "don't care," that is, it doesn't matter whether the pixel in the same position is black or white.



Part A: Pre-Boosting (10 points)

A1 Finding the Errors(5 points)

For each classifier, list the data points it gets wrong.

Classifier Name	General Idea	Misclassified
A	Has hair	2, 6, 7
B	Has forehead	4, 5
C	Has eyes	1, 5, 7
D	Has nose	1, 3, 7
E	Has smile	3, 4, 7
F	Has ear	8
G	TRUE (FACE)	1, 2, 3
H	NOT(A)	1, 3, 4, 5, 8
I	NOT(B)	1, 2, 3, 6, 7, 8
J	NOT(C)	2, 3, 4, 6, 8
K	NOT(D)	2, 4, 5, 6, 8
L	NOT(E)	1, 2, 5, 6, 8
M	NOT(F)	1, 2, 3, 4, 5, 6, 7
N	FALSE (NOT_FACE)	4, 5, 6, 7, 8

A2 Weeding out Classifiers (5 points)

Which, if any, of the classifiers in A1 can never be useful and why?

H contains errors of F
I contains errors of F
J contains errors of F
K contains errors of F
L contains errors of F
N contains errors of F

M contains errors of G

Part B: Boosting (25 points)

B1 Synthesis (20 points)

Synthesize boosting using **only classifiers A, B, C, and E.** For ties, choose alphabetically.

	Round1		Round2		Round3		Round 4	
w1	1/8	h_1= B	1/12	h_2= A	1/18	h_3= B	1/24	h_4=A
w2	1/8	e_1=2/8	(1/12)	e_2=3/12	1/6=3/18	e_3=6/18	(3/24)	e_4=9/24
w3	1/8	$a_1=\frac{1}{2}\ln 3$	1/12	$a_2=\frac{1}{2}\ln 3$	1/18	$a_3=\frac{1}{2}\ln 2$	1/24	$a_4=\frac{1}{2}\ln\frac{5}{3}$
w4	(1/8)		1/4=3/12		(3/18)		1/4=6/24	
w5	(1/8)		1/4=3/12		(3/18)		1/4=6/24	
w6	1/8		(1/12)		1/6=3/18		(3/24)	
w7	1/8		(1/12)		1/6=3/18		(3/24)	
w8	1/8		1/12		1/18		1/24	
e_A	3/8		(3/12)		9/18		(9/24)	
e_B	(2/8)		6/12		(6/18)		12/24	
e_C	3/8		5/12		7/18		10/24	
e_E	3/8		5/12		7/18		10/24	

B2 Amalgam Classifier (5 points)

What is the overall classifier after 4 rounds of boosting?

$H(\vec{x}) = \text{sign}\left(\frac{1}{2}\ln 6\, B(\vec{x}) + \frac{1}{2}\ln 5\, A(\vec{x})\right)$

What is its error rate?

$H(\vec{x}) = B(\vec{x}) \therefore$ error rate $= 2/8 = 25\%$

Part C: Miscellaneous (15 points)

C1 Using Classifiers (5 points)

Assume the boosting setup in Part B occurs for 100 rounds whether or not the overall classifier gets to the point where all training samples are correctly classified. Place the four classifiers in the following two bins

Frequently selected: A B	Infrequently selected: C E

C2 More using Classifiers (5 points)

Which **three** classifiers from A-N would you use so that you would not have to do boosting to get a perfect classifier for the samples.

A B F

C3 Even more using Classifiers (5 points)

Now, Suppose you are working with **just two classifiers,** neither of which has a zero error rate. Will boosting converge to an overall classifier that perfectly classifies all the training samples?

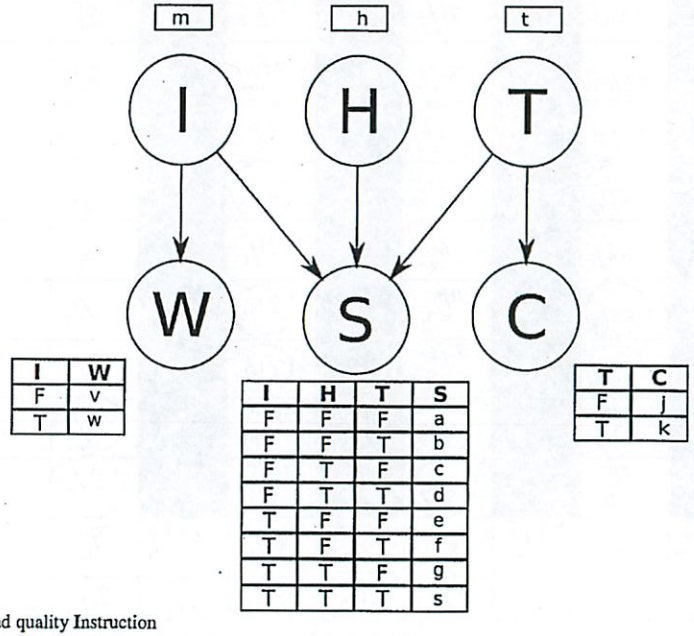
Yes (No)

Explain: Amalgam classifier will be equivalent to the higher-weighted weak classifier, so it will have the same error rate as that weak classifier. Neither weak classifier has a zero error rate, so the amalgam must be imperfect as well.

Quiz 5, Problem 1, Probability (50 points)

Part A: Life Lessons in Probability (22 pts)

Consider the following inference net developed for students who graduate from MIT:



I	W
F	v
T	w

I	H	T	S
F	F	F	a
F	F	T	b
F	T	F	c
F	T	T	d
T	F	F	e
T	F	T	f
T	T	F	g
T	T	T	s

T	C
F	j
T	k

I = Had quality Instruction

H = Hard working ethic

T = Raw Talent

S = Successful in life

C = Confidence

W = Took 6.034

A1: What is the probability that a 6.034 student (W = true) had quality instruction (I = true) and became successful in life (S = true), but did not have raw talent (T = false) yet was hardworking H = true) and confident (C = true). Leave your answer unsimplified in terms of constants from the probability tables. **(6pts)**

$$P(I, H, \bar{T}, W, S, C) = P(I)\,P(H)\,P(\bar{T})\,P(W|I)\,P(S|I,H,\bar{T})\,P(C|\bar{T})$$
$$= m \cdot h \cdot (1-t) \cdot w \cdot g \cdot j$$

For A2-A3: Express your final answer in terms of expressions of probabilities that could be read off the Bayes Net. You do not need to simplify down to constants defined in the Bayes Net tables. You may use summations as necessary.

A2: What is probability of success in life (S = true) **given** that a student has high quality instruction (I = true)? **(6 pts)**

$$\sum_{H \in \{T,F\}} \sum_{T \in \{T,F\}} P(S=true \mid I=true, H, T)\,P(H)P(T)$$

4 term expansion of above also acceptable.

A3: What is the probability a student is hardworking (H = true) , **given** that s/he was a 6.034 student (W = true)? **(10 pts)**

$$P(H=True \mid W=True) = \frac{P(H=true, W=true)}{P(W=true)} = \frac{\sum_I P(I)\,P(H=true)\,P(W=true|I)}{\sum_I P(I)\,P(W=true|I)}$$

$$= P(H=true)$$

H and W are independent!

Part B: The Naïve Crime Fighter (12 pts)

A murder has occurred in quiet town of South Park. Your have been provided the following table by forensic experts:

Suspect	Has a motive	Location	Murder Weapon	Degree of Suspicion
Professor Chaos	0.4	Fair = 0.1 School = 0.7 CW= 0.2	Kindness = 0.3 Chili = 0.3 Moral Fiber = 0.2 Pure Evil = 0.2	0.3
Mint Berry Crunch	0.3	Fair = 0.4 School = 0.4 CW = 0.3	Kindness = 0.4 Chili = 0.1 Moral Fiber = 0.4 Pure Evil = 0.1	0.1
The Dark Lord Cthulu	0.1	Fair = 0.3 School = 0.3 CW = 0.4	Kindness = 0 Chili = 0.5 Moral Fiber = 0 Pure Evil = 0.5	0.4
Scott Tenorman	0.9	Fair = 0.8 School = 0.1 CW = 0.1	Kindness = 0.2 Chili = 0.5 Moral Fiber = 0.1 Pure Evil = 0.2	0.2

You have determined that the murder did not have a motive, the murder took place at the Fair, and the murder weapon was a bowl of chili. You've decided to use what you learned about Naïve Bayes to help you determine who committed the murder.

The murderer is most likely:

The Dark Lord Cthulu.

Show your work:

$\arg\max_C P(C \mid M=\text{False}, L=\text{Fair}, W=\text{chili}) \propto P(M=\text{False}\mid C)\,P(L=\text{Fair}\mid C)\,P(W=\text{chili}\mid C)\,P(C)$

P.C. $(1-0.4)(0.1)(0.3)(0.3) = 56/10^4$

M.B.C. $(1-0.3)(0.4)(0.1)(0.1) = 28/10^4$

C $(1-0.1)(0.3)(0.5)(0.4) = 27\cdot 20/10^4$ ← max

T $(1-0.9)(0.8)(0.5)(0.2) = 80/10^4$

Part C: Coin Tosses (16 pts)

You decide to reexamine the coin toss problem using model selection.

You have **2** types of coins:

1) **Fair:** P(heads) = 0.5
2) **All-heads:** P(heads) = 1.0

You have **2** possible models to describe your observed sequences of coin types and coin tosses.

1) **Model 1:** You have both types of coins and you draw one of them at random, with equal probability, and toss it exactly once. You repeat both the drawing and tossing 3 times total.
2) **Model 2:** You have both types of coins and you draw one of them at random, with equal probability, and toss it exactly 3 times.

Finally, you have the following observed data.

Toss 1	Toss 2	Toss 3
H	H	T

The following questions use your knowledge of model selection to determine which is most likely.

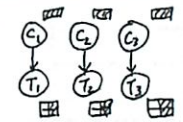
You decide to use the following criterion to weigh models:

$$P(M)=\frac{1}{Z}\frac{1}{|parameters|}$$ where Z is a normalization constant.

$|parameters|$ is defined as the number of cell entries in the CPTs of the Bayes Net representation.

C1. What is P(Model 1)? **(3 pts)** (Partial credit if you sketch the Models as Bayes Nets)

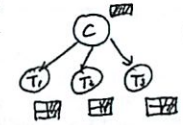
model 1



|params| = 3+6=9 $P(\text{model 1}) = \frac{1}{9Z}$

What is P(Model 2)? **(3 pts)**

model 2



|params| = 3·2+1 = 7 $P(\text{model 2}) = \frac{1}{7Z}$

You've decided that the a priori model probability P(Model) to use should be uniform.

P(Model 1) = P(Model 2)

Under this assumption you decide to work out the most likely model, given the data, P(Model | Data).

C2 What is the most-likely model based on this fully observed data set: **(10 pts)**

P(Model 1 | Data)?

$$= \frac{P(\text{Data} \mid \text{model1})\,P(\text{model1})}{c} = \left[\left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^1\right]^2 \left(\frac{1}{2}\right)^2\right] \cdot \frac{1}{2} = \frac{9}{64} \cdot \frac{1}{2}$$

($P(\text{model1})$ — Better)

P(Model 2 | Data)?

$$= \frac{P(\text{data} \mid \text{model 2})\,P(\text{model 2})}{c} = \left[\left(\frac{1}{2}\right)^3 \frac{1}{2}\right] \frac{1}{2} = \frac{1}{16} \cdot \frac{1}{2}$$

Therefore the most likely model is: (circle one) (Model 1) Model 2

Quiz 5, Problem 2, Near Miss (20 points)

Having missed many tutorials, lectures, and recitations, Stew is stuck on trying to figure out who are the TAs in 6.034. You, who is more faithfully attending, knows who is who. Armed with your knowledge about Near Miss concept learning. You decide to build a model that will help Stew figure out who the TAs are.

The following table summarizes the training data about the current staff of 6.034. The **Title** attribute is organized as a tree, with MEng and PhDs both a type of Student. Students and Faculty are grouped under the type People-You-See-On-Campus

Name	TA	Hair Color	Title	Glasses	Gender	Experience # Years	Taken 6.034
Kendra	Yes	Brown	MEng	yes	female	1	Yes
Kenny	Yes	Brown	MEng	no	male	1	Yes
Martin	Yes	Black	MEng	no	male	1	Yes
Mark	Yes	Black	PhD	no	male	4	Yes
Mich	Yes	Blonde	MEng	yes	male	10	No
Gleb	Yes	Brown	MEng	no	male	2	Yes
Yuan	Yes	Black	PhD	yes	male	3	No
Lisa	No	Blond	Professor	yes	female	10	No
Bob	No	Brown	Professor	no	male	10	No
Randy	No	Brown	Professor	no	male	10	No

Fill in the table to build a model of a TA. Mark an attributes as "?" if the property has been dropped.

Example	Heuristics Used	Model Description TAs					
		Hair Color	Title	Glasses	Gender	Experience	Taken
Kendra	Initial Model	Brown	MEng	yes	female	1	Yes
Kenny	Drop Link	{Brown, Black}	MEng	?	?	1	yes
Martin	Extend Set	{Brown, Black}	MEng	?	?	1	yes
Yuan	Climb tree close Interval Drop Link	{Brown, Black}	Student	?	?	[1-3]	?
Bob	None Not a nerd miss	{Brown, Black}	Student	?	?	[1-3]	?

Fill in the following table to build a model of a Faculty Member (FM).

Example	Heuristics Used	Model Description FMs					
		Hair Color	Title	Glasses	Gender	Experience	Taken
Patrick	Initial Model	Blond	Professor	Yes	Male	10	No
Randy	Extend set Drop Link	Brown, Blond	professor	?	Male	10	No
Bob	None (No difference)	Brown, Blond	professor	?	Male	10	No
Mick	Require Link	Brown, Blond	must professor	?	Male	10	No
Lisa	Drop Link	Brown, Blond	must professor	?	?	10	No

What class(es) would match these people given your TA model and your FM model. If neither, write N in the Class(es) column.

Name	Class(es)	Hair Color	Title	Glasses	Gender	Experience	Taken 6.034
Olga	N	Blond	MEng	no	female	1	Yes
Patricia	FM	Blond	Professor	Yes	female	10	No

not matched.

Quiz 5, Problem 3, Big Ideas (30 points)

Circle the **best** answer for each of the following question. There is no penalty for wrong answers, so it pays to guess in the absence of knowledge.

1 Ullman's alignment method for object recognition

1. Is an effort to use neural nets to detect faces aligned with a specified orientation
2. Relies on a presumed ability to put visual features in correspondence
3. Uses A* to calculate the transform needed to synthesize a view
4. Uses a forward chaining rule set
5. None of the above

2 Ullman's intermediate-features method for object recognition

1. Is an effort to use boosting with a classifier count not too small and not too large
2. Is an example of the Rumpelstiltskin principle
3. Is a demonstration of the power of large data sets drawn from the internet
4. Uses libraries containing samples (such as nose and mouth combinations) to recognize faces
5. None of the above

3 The SOAR architecture is best described as

1. A commitment to the strong story hypothesis
2. A commitment to rule-like information processing
3. An effort to build systems with parts that fail
4. The design philosophy that led to the Python programming language
5. None of the above

4 The Genesis architecture (Winston's research focus) is best described as, in part, as

1. A commitment to the strong story hypothesis
2. Primarily motivated by a desire to build more intelligent commercial systems
3. A commitment to rule-like information processing
4. A belief that the human species became gradually smarter over 100s of thousands of years.
5. None of the above

5 A transition frame

1. Focuses on movement along a trajectory
2. Focuses on the movement from childlike to adult thinking
3. Focuses on a small vocabulary of state changes
4. Provides a mechanism for inheriting slots from abstract frames, such as the disaster frame
5. None of the above

6 Reification is

1. The attempt to develop a universal representation
2. The tendency to attribute magical powers to particular mechanisms
3. The process by which ways of thinking are determined by macro and micro cultures
4. The process of using perceptions to answer questions too hard for rule-based systems
5. None of the above

7 Arch learning includes

1. A demonstration of how to combine the benefits of neural nets and genetic algorithms
2. A commitment to bulldozer computing using 100's of examples to learn concepts
3. The near miss concept
4. A central role for the Goldilocks principle
5. None of the above

2 3
4 5

8 Arch learning benefits importantly from

1. An intelligent teacher
2. Exposure to all samples at the same time
3. Use of crossover
4. Sparse spaces
5. None of the above

2 3
1 1
1
2

9 Experimental evidence indicates

1. People who talk to themselves more are better at physics problems than those who talk less
2. Disoriented rats look for hidden food in random corners of a rectangular room
3. Disoriented children combine color and shape information at about the time they start walking
4. Disoriented children combine color and shape information at about the time they start counting
5. None of the above

10 Goal trees

1. Enable rule-based systems to avoid logical inconsistency
2. Enable rule-based systems answer questions about behavior
3. Are central to the subsumption architecture's ability to operate without environment models
4. Are central to the subsumption architecture's ability to cope with unreliable hardware
5. None of the above

Tear off sheets

Duplicate data and drawings

6.034 Final, 2011

Do not hand these in

Q1, P1

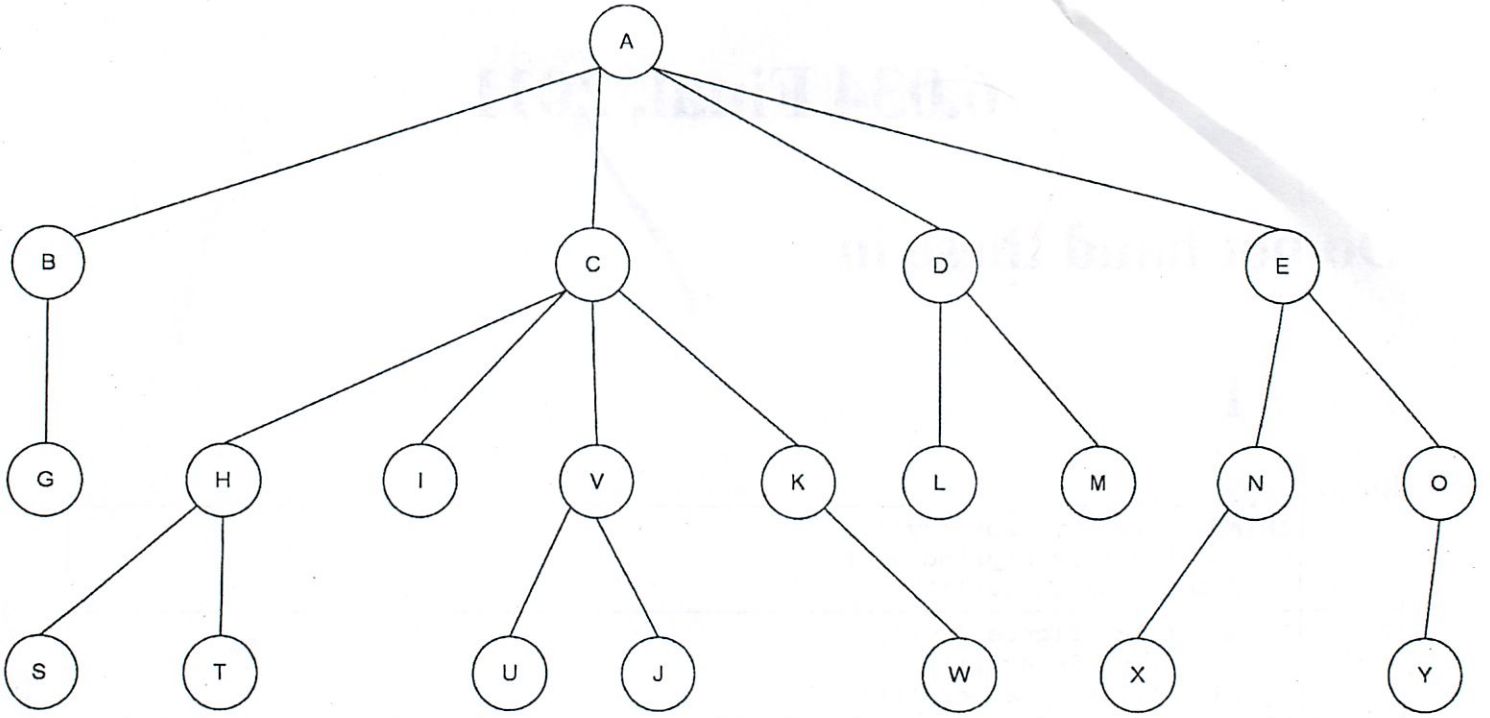
Rules:

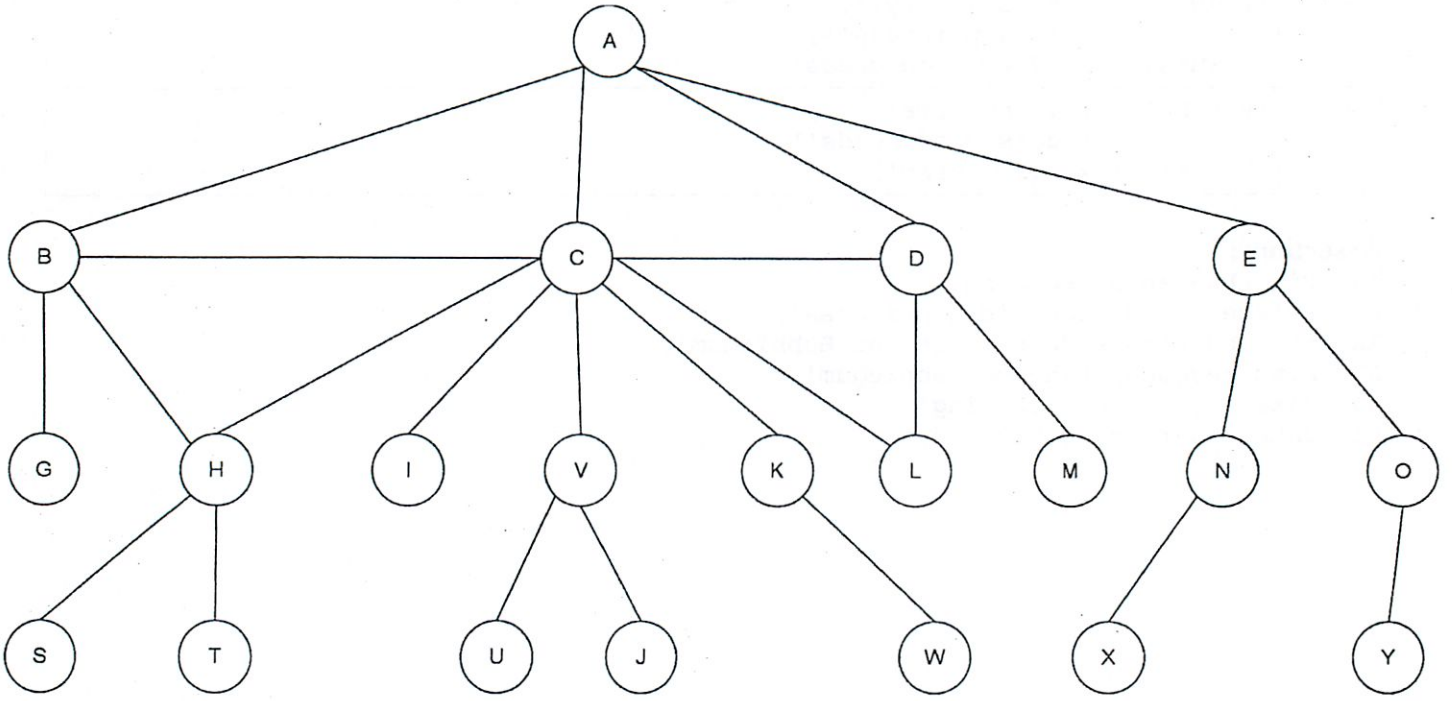
P0	IF(AND('(?x) kidnaps (?y)', '(?y) is a princess') THEN('(?x) is evil'))
P1	IF(AND('(?x) fights (?y)', '(?y) is evil'), THEN('(?x) is a hero'))
P2	IF(OR('(?x) has an awesome hat', '(?x) can reshape his body' THEN('(?x) is awesome'))
P3	IF(AND('(?x) rescues (?y)', '(?y) is a princess'), THEN('(?x) does good deeds'))
P4	IF(AND('(?x) is awesome', '(?x) does good deeds'), THEN('(?x) is a hero'))

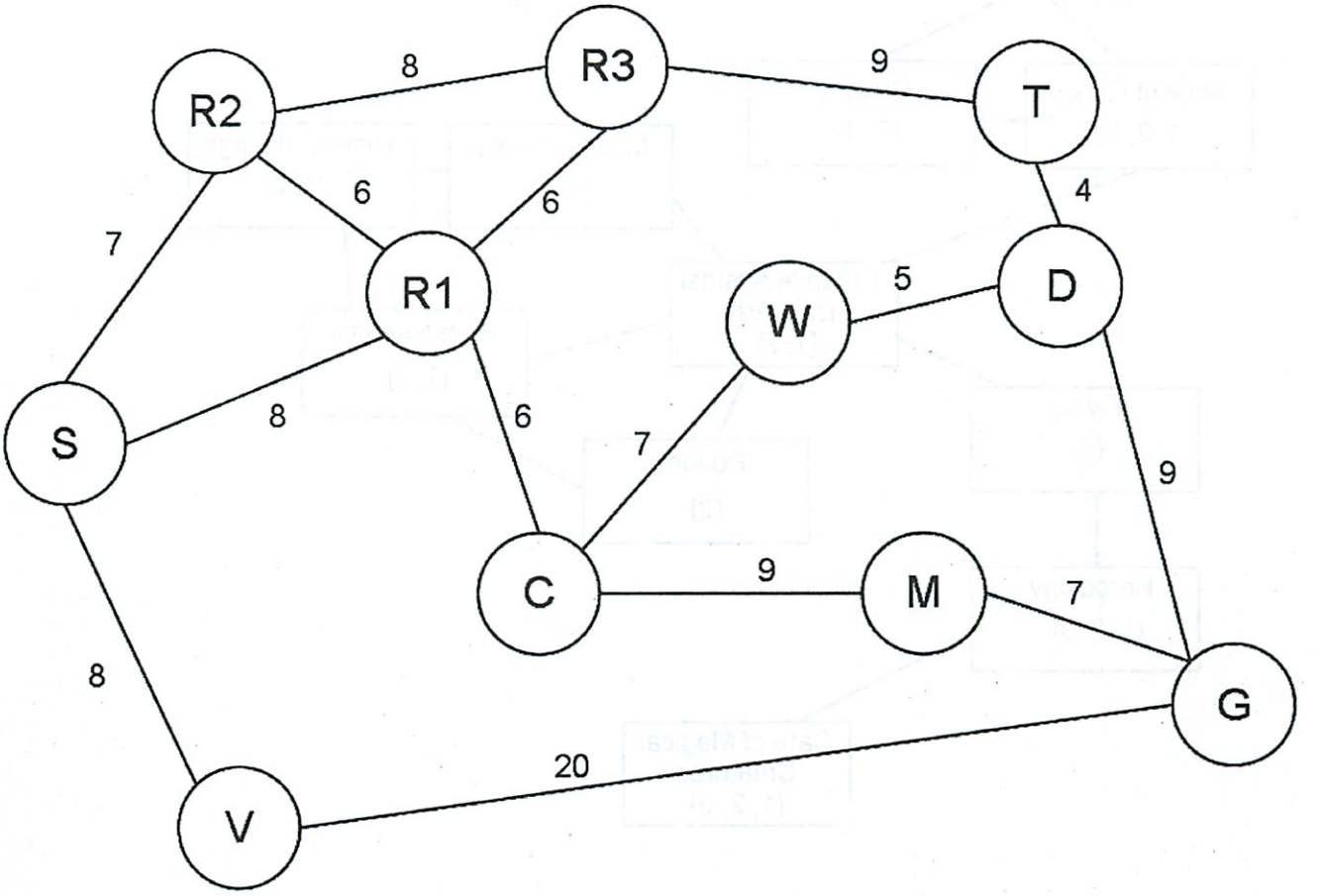
Assertions:

```
A0 'Finn has an awesome hat'
A1 'Princess Bubblegum is a princess'
A2 'the Ice King kidnaps Princess Bubblegum'
A3 'Finn rescues Princess Bubblegum'
A4 'Jake fights the Ice King'
A5 'Jake can reshape his body'
```

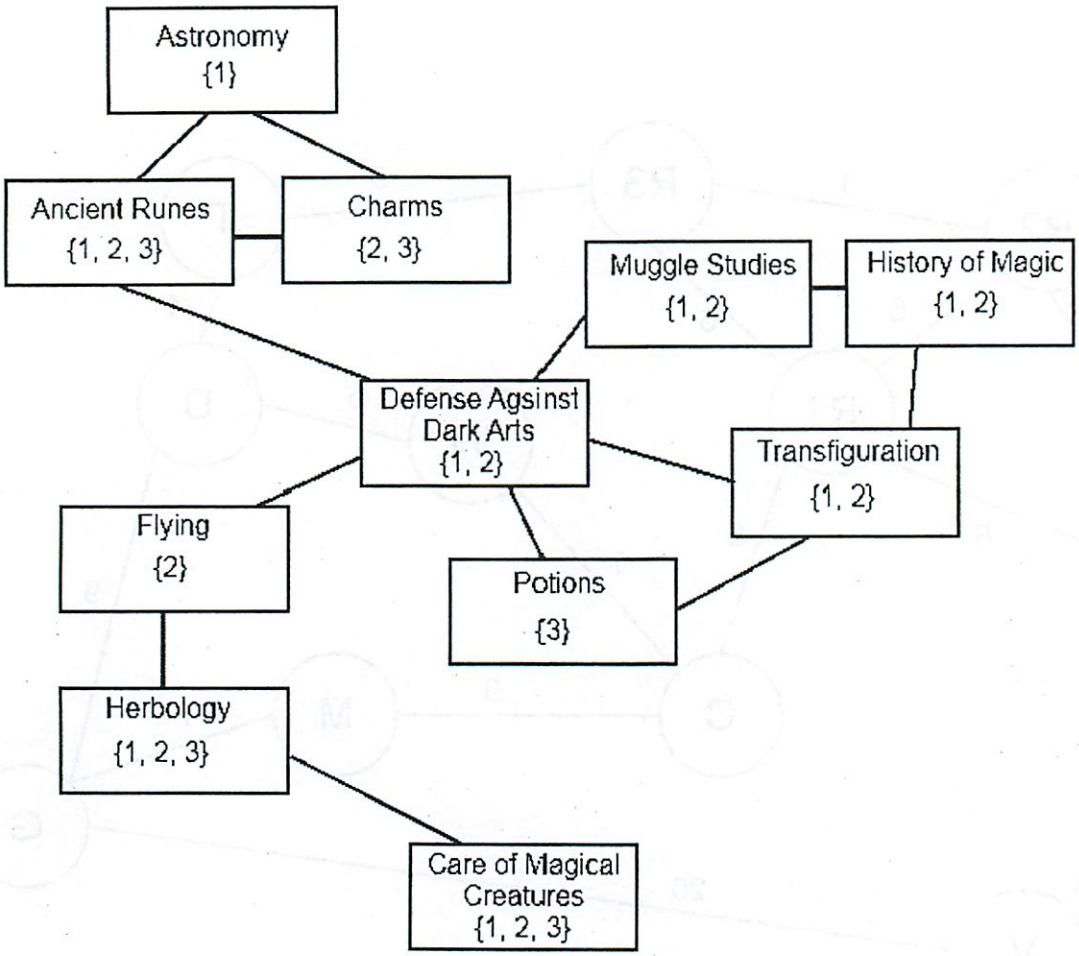
Q1, P2



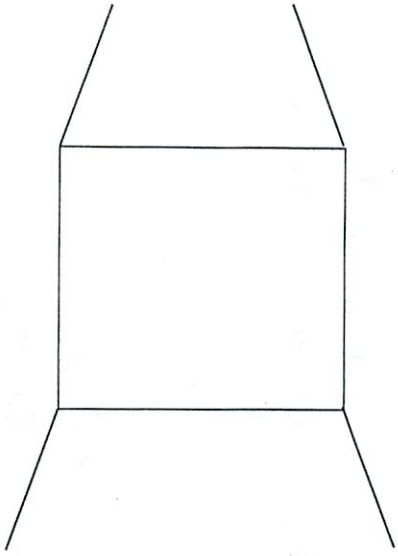


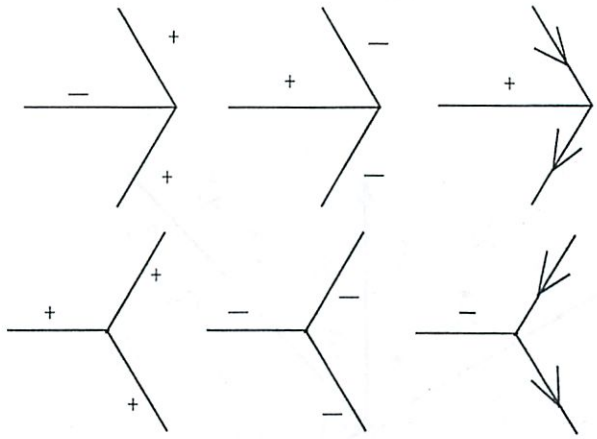

R2
R3
T
R1
W
D
S
C
M
G
V
8
9
6
6
4
7
5
8
6
7
9
9
7
8
20

Q2, P2

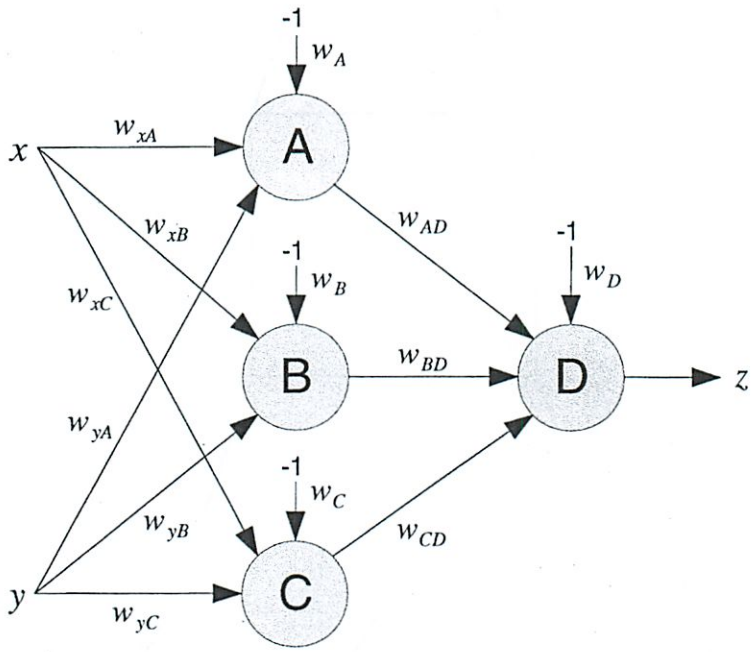

Astronomy
{1}
Ancient Runes
{1, 2, 3}
Charms
{2, 3}
Muggle Studies
{1, 2}
History of Magic
{1, 2}
Defense Agsinst
Dark Arts
{1, 2}
Transfiguration
{1, 2}
Flying
{2}
Potions
{3}
Herbology
{1, 2, 3}
Care of Magical
Creatures
{1, 2, 3}

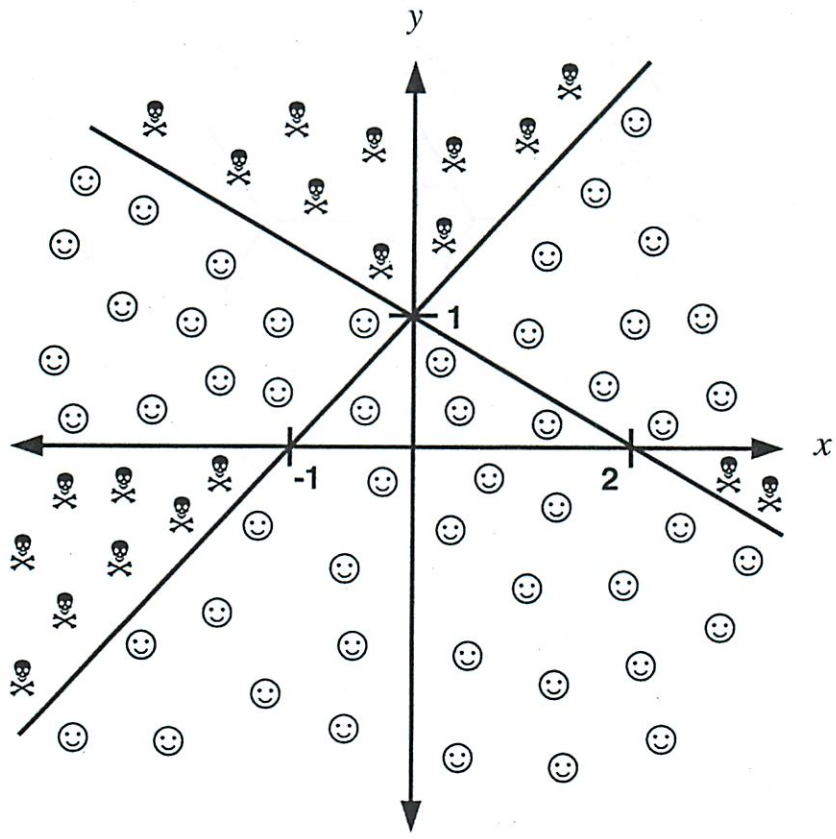
Q2, P3



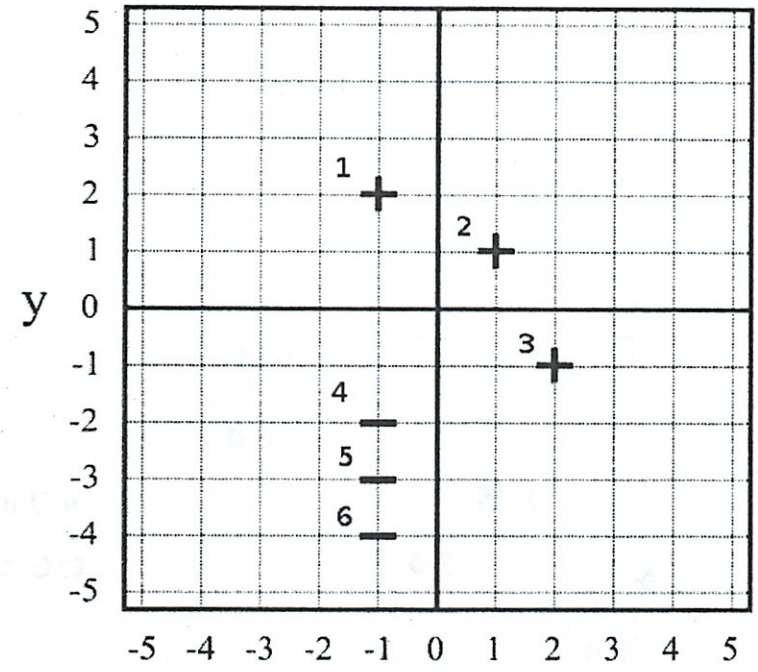


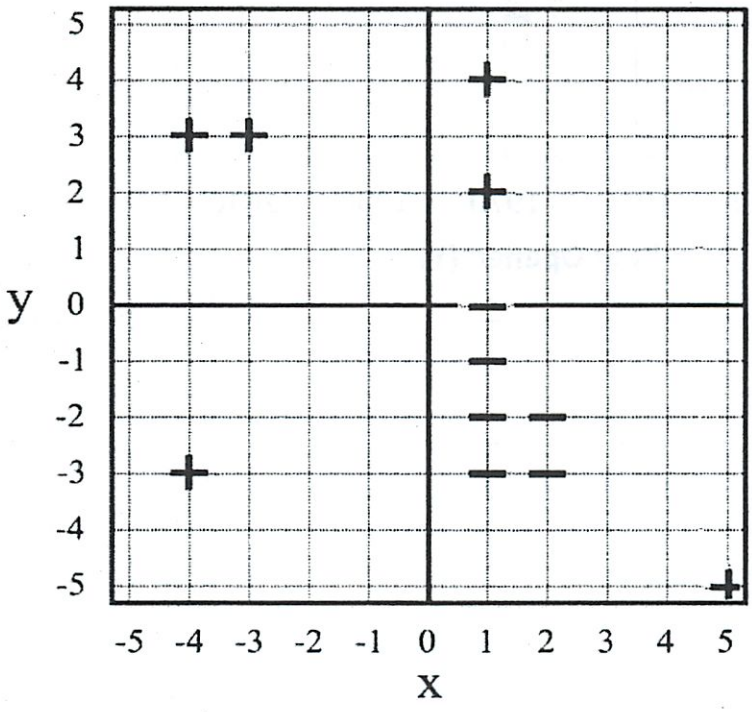
Q3, P2


-1
w_A
w_{xA}
x
A
w_{xB}
w_{AD}
-1
w_D
-1
w_B
w_{xC}
B
w_{BD}
D
z
w_{yA}
-1
w_C
w_{yB}
w_{CD}
y
C
w_{yC}

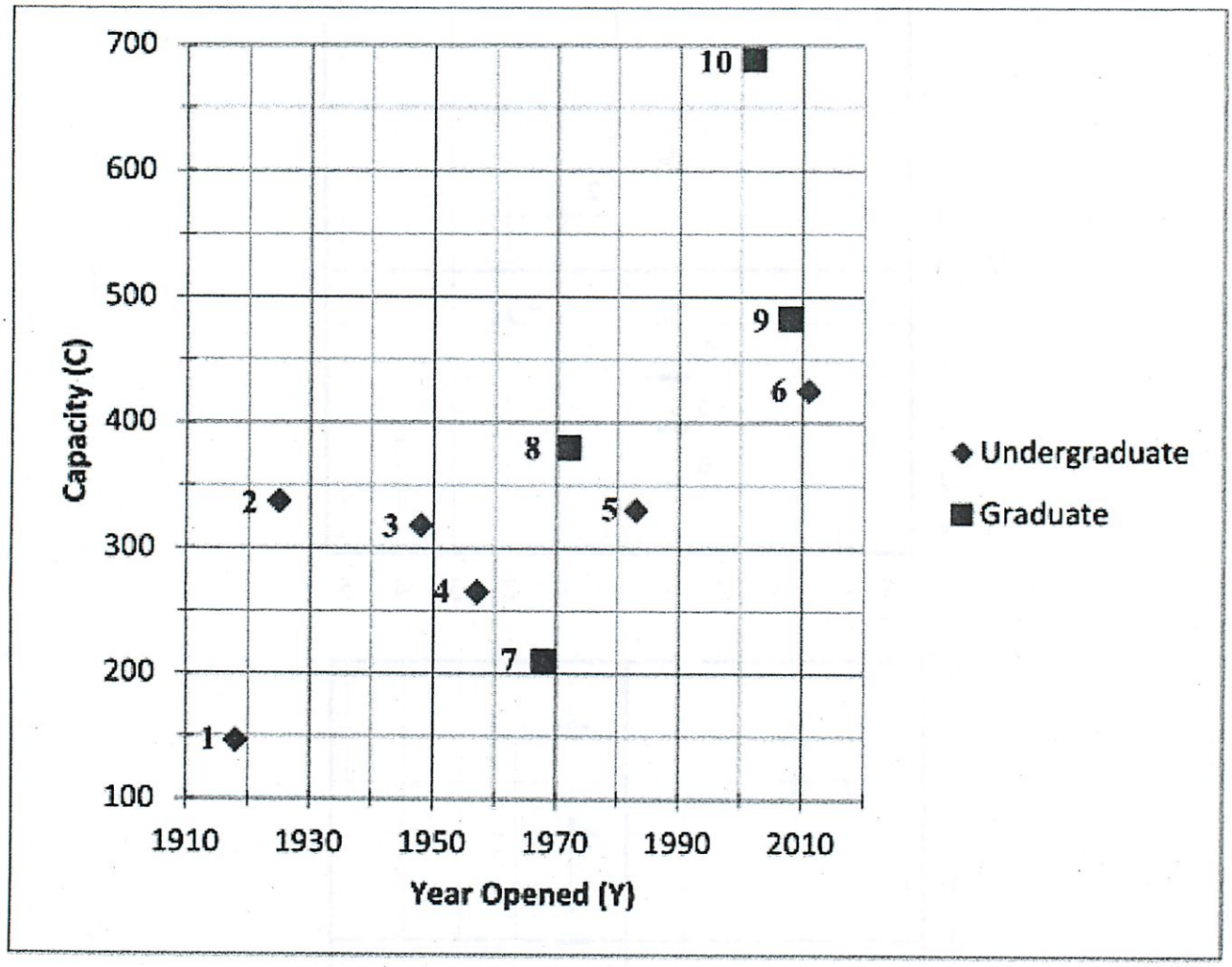

y
1
-1
2
x

Q4, P1





Q4, P2


700
600
500
400
300
200
100
Capacity (C)
1910
1930
1950
1970
1990
2010
Year Opened (Y)
1
2
3
4
5
6
7
8
9
10
Undergraduate
Graduate

Q5, P1

- **S:** True if you got enough sleep the night before the final.
- **C:** True if you drank coffee before the final.
- **R:** True if you review the material for quiz 5 for at least four hours.
- **A:** True if you get a 5 on quiz 5.

Furthermore, you know the following statements about each random variable to be true:

- If you did not sleep enough the night before the final, you definitely drink coffee. You do not drink coffee otherwise.
- You are awake enough for the final exactly if you have had enough sleep or if you've had coffee.
- The number of hours you review for the final does not depend on whether or not you drink coffee.
- If you are awake and you reviewed for at least four hours for the final, you have an 80% chance to solve quiz 5 well enough to get a 5.
- If you are awake but you did not review for at least four hours, you still have a 20% chance of guessing enough right answers to get a 5.
- If you are not fully awake, you stand no chance at getting a 5.

6.034 Debrief

12/4

Prob av really hard

But I think I got everything else

Hopefully got an A in class

↑ perhaps if a slight amt of time more would help get it...