6.034 Quiz 4

6 December 2017

Name	Millie Bobby Brown
Email	not-a-demogorgon@mit.edu

For 1 extra credit point: Circle the TA whose recitations you attend, so that we can more easily enter your score in our records and return your quiz to you promptly:

Suri Bandler

Jake Barnwell

Abigail Choe

Francesca Cicileo

Erin Hong

Nathan Landman

Amanda Liu

Nick Matthews

Samarth Mohan

Michael Shum

Jackie Xu

Problem	Maximum	Score	Grader
1 – Bayes	50		
2 – Boosting	50		
Total	**100**		

SRN	6		

Survey

We are investigating the effects of sleep on performance. We will **not** consider your answer to these questions when determining grades.

How many hours of sleep did you get last night? □

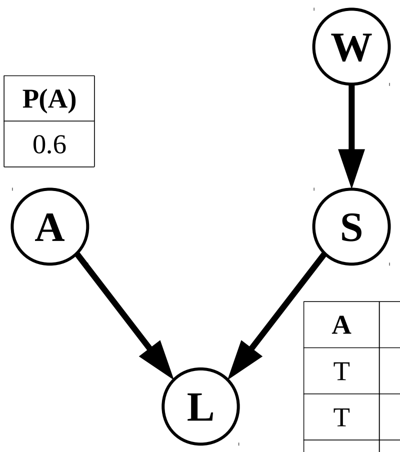
How many hours of sleep did you get **two** nights ago? □

There are 12 pages in this quiz, including this one, but not including tear-off sheets. A tear-off sheet with equations and duplicate data is located after the final page of the quiz.

This quiz is open book, open notes, open just about everything, including a calculator, but no computers.

Problem 1: Bayes (50 points)

Patrick Winston is interested in modeling 6.034 lecture attendance. He theorizes that students' attendance at ***L***ecture is dependent on whether their ***A***larm clock is functional and whether they ***S***lept the night before, which itself depends on whether other classes assigned ***W***ork the previous day. Based on some polling, he establishes the probabilities shown in the Bayes net on the right.



P(W)
0.3

P(A)
0.6

W	P(S | W)
T	0.2
F	0.9

A	S	P(L | AS)
T	T	0.9
T	F	0.7
F	T	0.6
F	F	0.2

Part A: Probable Cause (22 points)

For questions A1 and A2, you can show your work for partial credit in the boxes below.

A1 (5 points) Structurally, is **A** conditionally independent of **S** given **W**?

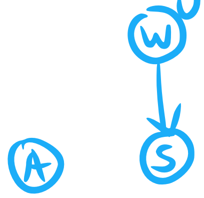
[✓ YES] **NO** **Can't Tell**

A2 (5 points) Structurally, are **A** and **S** marginally independent?

[✓ YES] **NO** **Can't Tell**

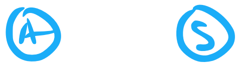
For partial credit on A1, you can show your work here.

① ancestral graph



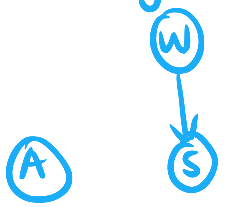
② moralize (no change)

③ disorient + delete givens



For partial credit on A2, you can show your work here.

① ancestral graph



② moralize, disorient, delete givens (no change)

A3 (6 points) What is the probability that a student ***S***lept last night? *Below, circle the* ***one*** *number that is closest to the marginal probability* ***P(S)***.

0.1 **0.3** **0.5** [✓ **0.7**]

For partial credit on A3, you can show your work here.

$$P(S) = \sum_{w} P(Sw) = P(SW) + P(S\bar{W})$$

$$= P(S|W)P(W) + P(S|\bar{W})P(\bar{W})$$

$$= (0.2)(0.3) + (0.9)(0.7)$$

$$= 0.69$$

A4 (6 points) Patrick has been wondering about Kyla the Therapy Dog's sleeping schedule. Though Kyla is a dog, Patrick thinks that her sleeping tendencies can be modeled in the same way as those for students, **except he instead assumes P(S) = 0.5 for Kyla**. Given the assumption P(S) = 0.5, what is the probability that Kyla ***S***lept last night and her ***A***larm clock is functional?

P(S,A) = $\boxed{0.3}$

For partial credit on A4, you can show your work here.

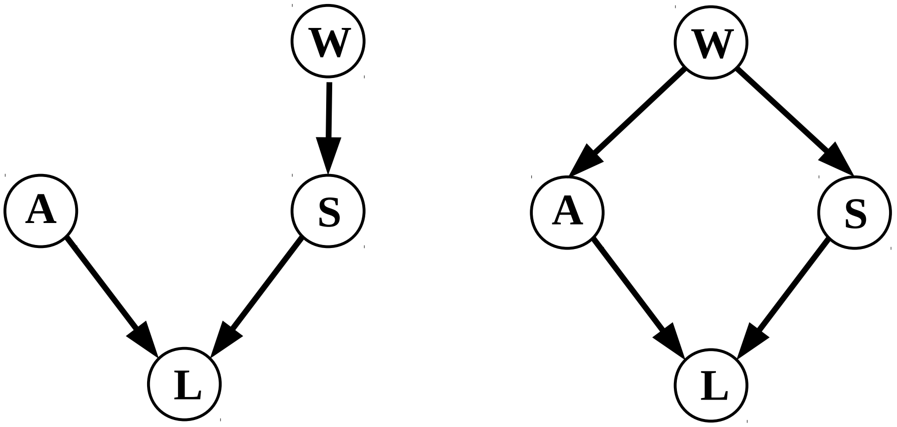
$S \perp\!\!\!\perp A$ so $P(SA) = P(S)P(A)$

$$= (0.5)(0.6)$$

$$= 0.3$$

Part B: A Small Change… (8 points)

Kimberle suggests an *alternative* Bayes network to model lecture attendance, as shown below to the right. For your convenience, we have also reproduced Patrick's Bayes net, from part A, on the left. (The conditional probability tables have been omitted; they are not relevant to this problem.)


Patrick's model (from part A)

Kimberle's proposed model

B1 (4 points) Given Kimberle's new model, is **A** structurally independent of **S** given **W**? Circle the **one** best answer below.

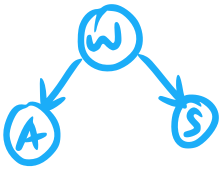
[✓ **YES**]

NO

Can't Tell

For partial credit on B1, you can show your work here.

① ancestral graph



② moralize (no change)

③ disorient + delete givens

B2 (4 points) Patrick claims that Kimberle's model has one *undesirable* trait. What could he be referring to? Circle the **one** best answer.

(a) Her model is symmetric; symmetric models are more difficult to train.

(b) Her model is not a true Bayes net, as it contains loops.

(c) Her model encodes more independence assumptions: it is more likely to underfit.

[✓ (d)] Her model claims **A** is structurally dependent on **W**, which seems less reasonable.

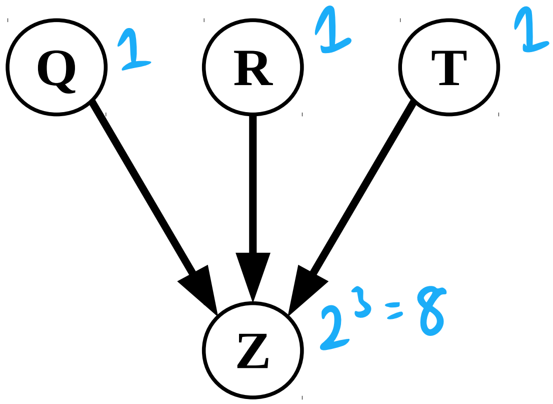
fewer

Part C: Model Madness (20 points)

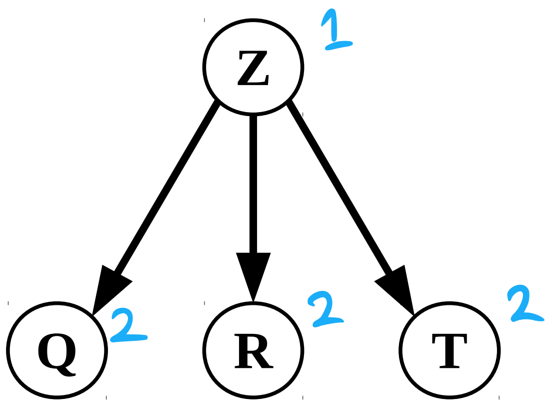
This section is not related to parts A or B.

Consider two different Bayes nets, each using the same **boolean** variables to model some data.

Bayes net 1



Bayes net 2



C1 (6 points) For each Bayes net above, what is the number of entries (parameters) necessary to be stored in that net's lookup tables to fully recreate the joint probability table? *For partial credit, on the Bayes nets above, you may draw the conditional probability tables associated with each variable.*

	Bayes net 1	**Bayes net 2**
Number of Entries Necessary:	11	7

C2 (4 points) To train a Bayes net model, one must use many data points to populate the conditional probability tables. *Assuming that both Bayes nets, once fully trained, accurately model the data*, which is likely to be **easier** to train? Below, circle the **one** best answer.

more

(a) Bayes net 1, because it encodes fewer independence assumptions.

(b) Bayes net 1, because it can make use of explaining away.

[✓ (c)] Bayes net 2, because there are fewer parameters that have to be tuned.

(d) Bayes net 2, because Q, R, and T are all marginally independent.

(e) None of the above statements is true.

conditionally

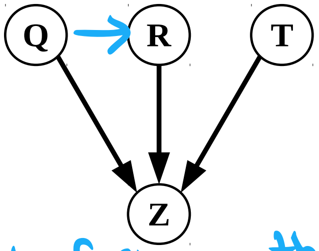
C3 (6 points) Is it possible to **add additional arrows** (links) to Bayes net 1, in order to *increase* the minimum number of entries (parameters) necessary stored in its probability tables? How about to *decrease* the minimum number of entries?

In each row below, circle either **NO** or **YES**. If you circle **YES**, draw the additional links between variables on the diagram given.

Possible to ***increase*** *minimum number of entries by adding arrows?*

NO, it is not possible.	[✓ **YES**, it is possible. For example:]

(draw arrows on this diagram)

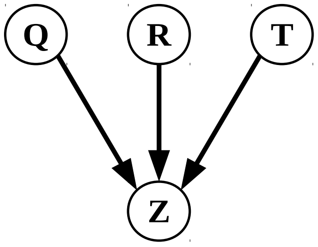


any set of arrows that doesn't result in a cycle is correct

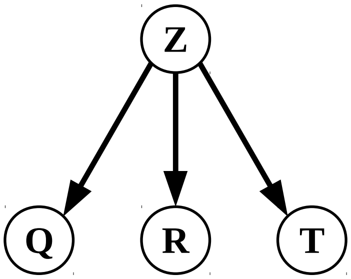
Possible to ***decrease*** *minimum number of entries by adding arrows?*

[✓ **NO**, it is not possible.]	**YES**, it is possible. For example:

(draw arrows on this diagram)



C4 (4 points) You are asked to estimate **P(Q=True | Z=True)** using **Bayes net 2** (reproduced to the right for your convenience). As soon as you write down your estimate, you are given *new information*: **R=True**. Given this information, how does the probability of **Q=True** change? Circle the **one** best answer below.



$P(Q|Z) = P(Q|ZR)$ because

$Q \perp\!\!\!\perp R \mid Z$

Decreases ↓ [✓ **Stays the same**] **Increases ↑** **Need more information**

Problem 2: Stranger Boosting (50 points)

Part A: Save Hawkins! (31 points)

The Demogorgon infection is quickly spreading through Hawkins. The Hawkins Lab has retrieved samples from six (6) individuals and needs to check whether or not they are Demogorgons. They have five (5) unreliable forms of "Is Demogorgon?" tests (***h**$_i$*), and have hired you, the local Adaboost expert, to figure out how to combine the results from all 5 tests to determine who is a **Demogorgon (DG)** and who is a **Human (H)**.

	Training points: six (6) individuals					
	Will	**Max**	**Nancy**	**Steve**	**Mews**	**Billy**
***h**$_1$*	H	H	H	DG	DG	DG
***h**$_2$*	DG	DG	H	H	H	H
***h**$_3$*	H	H	DG	DG	DG	DG
***h**$_4$*	H	DG	DG	DG	DG	DG
***h**$_5$*	DG	DG	H	DG	H	DG
Actual Classification	**DG**	**H**	**H**	**H**	**DG**	**DG**

A1 (3 points) Complete the table below by filling in which training points above are misclassified by the fifth weak classifier, ***h**$_5$*:

Weak Classifier	**Misclassified Training Points**
***h**$_1$*	Will, Steve
***h**$_2$*	Max, Mews, Billy
***h**$_3$*	Will, Nancy, Steve
***h**$_4$*	Will, Max, Nancy, Steve
***h**$_5$*	Max, Steve, Mews

A2 (24 points) On the next page, perform 2.5 rounds of boosting using these classifiers and training data. In each round, pick the classifier with the **error rate furthest from** ½. Break ties by choosing earlier classifiers. In any round, if Adaboost would terminate instead of choosing a classifier, write **NONE** for that round's weak classifier (**h**), then leave all remaining spaces blank.

For your convenience, a copy of the data, as well as an equation sheet for Adaboost, is provided on a tear-off sheet at the end of the quiz.

	Round 1	Round 2	Round 3
weight of Will	1/6 X	1/4 X	2/12
weight of Max	1/6	1/8 X	1/12
weight of Nancy	1/6	1/8 X	1/12
weight of Steve	1/6 X	1/4 X	2/12
weight of Mews	1/6	1/8	1/4
weight of Billy	1/6	1/8	1/4
error rate of h_1	$\boxed{2/6}$	4/8	
error rate of h_2	3/6	3/8	
error rate of h_3	3/6	5/8	
error rate of h_4	4/6	$\boxed{6/8}$	
error rate of h_5	3/6	4/8	
weak classifier chosen (**h**)	h_1	h_4	
weak classifier error (**ε**)	2/6	6/8	
voting power (**α**)	$\frac{1}{2}\ln 2$	$\frac{1}{2}\ln\frac{1}{3}$ or $-\frac{1}{2}\ln 3$	

For partial credit on A2, you can show your work here.

A3 (4 points) Suppose we continue running the Adaboost algorithm for a large number of rounds. Is it possible for $\boldsymbol{h_3}$ to be chosen as the "best" classifier is some round of Adaboost? Circle the **one** best answer below, and briefly explain your answer. (Hint: What does $\boldsymbol{h_3}$ misclassify compared to other classifiers?)

YES

[✓ **NO**]

Not enough information

Briefly explain:
- points misclassified by h_3 are a subset of points misclassified by h_4 and a superset of points misclassified by h_1

Part B: Oh, Joyce! (6 points)

As always, Joyce Byers is panicking. She doesn't trust your Adaboost skills, so she creates her own ensemble classifier *J(x)*:

$$J(x) \;=\; 3\cdot\left(\boldsymbol{h_1}'s \;\; vote\right) + 5\cdot\left(\boldsymbol{h_2}'s \;\; vote\right) + 2\cdot\left(\boldsymbol{h_5}'s \;\; vote\right)$$

Given the data for 3 unknown test points (Mike, Eleven, and Dustin), how would Joyce's classifier classify each point? In each box below, write either **Demogorgon** or **Human**; or, if the classifier can't tell, instead write **CAN'T TELL**.

	Test Points		
	Mike	**Eleven**	**Dustin**
$\boldsymbol{h_1}$	H	H	H
$\boldsymbol{h_2}$	DG	H	DG
$\boldsymbol{h_5}$	H	DG	DG
Prediction	Can't tell	H	DG

For partial credit on part B, you can show your work here.

$J(\text{Mike}) = 3(H) + 5(DG) + 2(H) \rightarrow 5DG = 5H$

$J(\text{Eleven}) = 3(H) + 5(H) + 2(DG) \rightarrow 7H > 2DG$

$J(\text{Dustin}) = 3(H) + 5(DG) + 2(DG) \rightarrow 3H < 7DG$

Part C: Conceptual Questions (13 points)

This section consists of questions about Adaboost *in general*—they are not related to either of the previous sections. Circle the **one** best answer for each question.

C1 (3 points) The weight of a training point can be the same across two consecutive rounds.

TRUE [✓ **FALSE**]

C2 (3 points) A weak classifier that misclassifies exactly half the training points will always have an error rate of ½.

TRUE [✓ **FALSE**]

C3 (3 points) If some classifier $\boldsymbol{h}_i$ was chosen in some round of Adaboost, then it can be chosen again in a future round.

[✓ **TRUE**] **FALSE**

C4 (3 points) When "best" means "smallest error," the best weak classifier $\boldsymbol{h}$ can have a negative voting power.

TRUE [✓ **FALSE**]

C5 (1 point) If all weak classifiers correctly classify all points, then it's possible to manually select voting powers in a way to make an ensemble classifier misclassify a point.

[✓ **TRUE**] **FALSE**

Spiritual and Right Now (6 points)

For each of the following questions, circle the **one** best answer. There is **no penalty for wrong answers**, so it pays to guess in the absence of knowledge.

1. To motivate the challenges of data science, Mansinghka referenced a reporter investigating:

(a) Russian interference in the most recent presidential election.

(b) Fairness of Internet service providers for net neutrality.

(c) Changes in socioeconomic patterns of neighborhood settlements.

d [✓ (d)] Bias in evaluating the risk of criminals recommitting crimes.

(e) Reliability of human data scientists vs. BayesDB.

2. Berwick explained that an important outcome of the merge operator in humans is:

(a) The aptitude to reason with multiple sensory inputs.

(b) The capability to link memories with place.

c [✓ (c)] The ability to create hierarchical representations.

(d) The capacity to complete logical induction proofs.

(e) The potential to multitask on at most two objectives.

3. Pratt compared the Chauffeur and the Guardian systems which differ in that:

(a) Chauffeur is suitable for city driving and Guardian for highway driving.

[✓ (b)] Chauffeur is made for full autonomy and Guardian for driver assistance.

b (c) Chauffeur is designed solely for the elderly and Guardian for adolescents.

(d) Chauffeur is a lesser legal liability and Guardian a lesser financial risk.

(e) Chauffeur is capable of higher speeds and Guardian greater acceleration.

4. Winston noted that:

(a) Cross-modal coupling helps relate folk tales to modern political theatre.

(b) Self-organizing maps find optimal routes from point A to point B in minimal time.

(c) Correlation is easily determined between data sets only when noise is minimal.

e (d) Cross-modal coupling may be a mechanism used in word-sense disambiguation.

[✓ (e)] Self-organizing maps may be a mechanism used in symbol grounding.

5. Winston suggested that the best way to recognize actions, such as sitting, is to:

(a) Use deep neural nets with no more than a hundred layers.

(b) Train deep neural nets using depictions of stick figures instead of real people.

c [✓ (c)] Have a system imagine what it would be like if a model were aligned to an image.

(d) Train a self-organizing map with images of people performing actions.

(e) Use correlation with a large image library of people performing actions.

6. Winston believes that:

(a) The next triumph in deep net learning will be learning from children's stories.

[✓ (b)] The role of stories in human intelligence explains why merge matters.

b (c) Constraint propagation is a key mechanism in story understanding.

(d) Story understanding is a special case of deductive reasoning.

(e) Stories are important in the humanities but not, for example, in explaining circuits.

(No test material on this page.)

(Tear-off sheet)

Data and table of misclassifications for Problem 2, Part A:

	Training points: six (6) individuals					
	Will	**Max**	**Nancy**	**Steve**	**Mews**	**Billy**
$\boldsymbol{h_1}$	H	H	H	DG	DG	DG
$\boldsymbol{h_2}$	DG	DG	H	H	H	H
$\boldsymbol{h_3}$	H	H	DG	DG	DG	DG
$\boldsymbol{h_4}$	H	DG	DG	DG	DG	DG
$\boldsymbol{h_5}$	DG	DG	H	DG	H	DG
Actual Classification	**DG**	**H**	**H**	**H**	**DG**	**DG**

Weak Classifier	**Misclassified Training Points**
$\boldsymbol{h_1}$	Will, Steve
$\boldsymbol{h_2}$	Max, Mews, Billy
$\boldsymbol{h_3}$	Will, Nancy, Steve
$\boldsymbol{h_4}$	Will, Max, Nancy, Steve
$\boldsymbol{h_5}$	

Common Adaboost equations:

1. $$\alpha = \frac{1}{2} \ln\left(\frac{1-\epsilon}{\epsilon}\right)$$

2. $$w_{new} = \left\{ \begin{array}{l} \frac{1}{2}\frac{1}{1-\epsilon} w_{old} \text{ , if the point was correctly classified} \\ \frac{1}{2}\frac{1}{\epsilon} w_{old} \text{ , if the point was incorrectly classified} \end{array} \right\}$$