6.034 Quiz 4

3 December 2014

Name	Saul Ushin
Email	

Circle your TA **(for 1 extra credit point)**, so that we can more easily enter your score in our records and return your quiz to you promptly.

Josh Blum	**Elisa Castañer**	**Pedro Cattori**
Martin Gale	**Malcom Gilbert**	**Dylan Holmes**
Jessica Noss	**Duncan Townsend**	**Siyao Xu**

Problem number	Maximum	Score	Grader
1	50		
2	50		
Total	100		

SRN	6		

There are 9 pages in this quiz, including this one. As always, open book, open notes, open just about everything, including a calculator, but no computers.

Problem 1: Adaboost (50 points)

Part A: Two rounds of boosting (32 points)

A1 (22 points)

You have six training points (A, B, C, D, E, F) and five classifiers (h_1, h_2, h_3, h_4, h_5) which make the following misclassifications:

Classifier	Misclassified training points (A, B, C, D, E, F)					
h_1	A			D		F
h_2				D		
h_3		B	C			
h_4	A	B				F
h_5		B	C	D		

Perform two rounds of boosting with these classifiers and training data. In each round, pick the classifier with the **lowest error rate**. Break ties by picking the classifier that comes first in this list: h_1, h_2, h_3, h_4, h_5. Space for scratch work is provided on the following page.

	Round 1		Round 2	
weight_A	1/6	✓	1/10	✓
weight_B	1/6	✓	1/10	✗
weight_C	1/6	✓	1/10	✗
weight_D	1/6	✗	1/2 = 5/10	✓
weight_E	1/6	✓	1/10	✓
weight_F	1/6	✓	1/10	✓
Error rate of h_1	3/6		7/10	
Error rate of h_2	1/6 lowest error rate		5/10	
Error rate of h_3	2/6		2/10	
Error rate of h_4	3/6		3/10	
Error rate of h_5	3/6		7/10	
weak classifier (h)	h_2		h_3	
classifier error (ε)	1/6		2/10 = 1/5	
voting power (α)	$\frac{1}{2}\ln(5)$		$\frac{1}{2}\ln(4)$	

$\frac{1}{2}\ln\left(\frac{1-1/6}{1/6}\right)$ $\frac{1}{2}\ln\left(\frac{1-1/5}{1/5}\right)$

Space provided for scratch work

Round 2

Classified correctly in Round 1: A, B, C, E, F

$w_{new} = \frac{1}{2} \cdot \frac{1}{1-\varepsilon} \cdot w_{old} = \frac{1}{2} \cdot \frac{1}{1-1/6} \cdot \frac{1}{6} = \frac{1}{10}$ (Also, $\sum_{A,B,C,E,F} w_i = \frac{1}{2}$)

Misclassified in Round 1: D

$w_{new} = \frac{1}{2} \cdot \frac{1}{\varepsilon} \cdot w_{old} = \frac{1}{2} \cdot \frac{1}{1/6} \cdot \frac{1}{6} = \frac{1}{2}$ (Also, $\sum_{D} w_i = \frac{1}{2}$)

A2 (6 points)

Three of the training points (B, D, F) have been selected below. For each one, decide whether the ensemble classifier $H(\vec{x})$ produced after two rounds of boosting misclassifies or correctly classifies that point. Circle the best answer in each case. If the answer can't be determined from the available information, circle "Can't tell" instead.

because h_2 has more voting power

Training point	Classification by ensemble classifier		
B	Correctly classified	Misclassified	Can't tell
D	Correctly classified	Misclassified	Can't tell
F	Correctly classified	Misclassified	Can't tell

because both h_2 and h_3 classify F correctly

A3 (4 points)

Suppose you continue the Adaboost procedure from Part A for a **total** of 2014 rounds. (You may assume it doesn't terminate before then.) If you always pick the classifier with the **lowest** error rate, which training data point will have the **smallest** weight at the end of the 2014th round? (Circle one.)

A **B** **C** **D** **E** **F**

Because it is never misclassified

Part B (18 points)

This section consists of questions about Adaboost *in general*—they do not rely on the preceding section. Decide whether each of the statements below is true or false. Circle the one best answer in each case.

True False 1. Adaboost accounts for outliers by lowering the weights of training points that are repeatedly misclassified.

Adaboost increases weights of misclassified points.

True False 2. When you update weights, the training point with the smallest weight in the previous round will always increase in weight.

If $\varepsilon < \frac{1}{2}$ and the point with the smallest weight is classified correctly, its weight will decrease.

True False 3. Four weak classifiers that make disjoint errors can always be used to construct a perfect classifier by picking $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$.

True False 4. In each round, any weak classifier with an error rate less than ½ must misclassify fewer than half of the training points.

could misclassify lots of points with small weights

True False 5. After two rounds of boosting, the ensemble classifier $H(\vec{x})$ always behaves like the conjunction (AND) of two weak classifiers.

$H(\vec{x})$ takes a vote between two weak classifiers.
If they disagree, whichever one has a higher voting power wins.

True False 6. If you have 100 training points, the updated weight of a training point can never be greater than ¾.

Updated weight can never be greater than 1/2.

True False 7. A weak classifier will never have an error rate greater than ½.

Error rate can be 1 if it misclassifies everything.

True False 8. Once a weak classifier is picked in a particular round, it will never be chosen in any subsequent round.

True False 9. Weak classifiers with error rates (ε) close to 100% are assigned negative voting powers (α).

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

Problem 2: Bayesian Inference (50 points)

★ Note: This problem consists of four independent subproblems. ★

Part A: Bayes rule and model selection (10 points)

Suppose you have three coins in a bag: the first coin is fair, with HEADS on one side and TAILS on the other. The second has HEADS on both sides. The third has TAILS on both sides.

Your friend takes a coin out of the bag at random and flips it. When it lands, you observe that the side of the coin facing up is HEADS. What is the probability that this is the fair coin?

$\frac{2}{3}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$
(A)	(B)	(C)	(D)

Check your intuition: You should treat each coin as a different model M_1, M_2, M_3 for explaining the evidence E of having HEADS on one side. Thus, the probability that the coin is fair is $P(M_1|E)$, which you can compute using Bayes' Law.

Space for scratch work.

Priors: $P(M_1) = P(M_2) = P(M_3) = \frac{1}{3}$

$$P(M_1|E) = \frac{P(E|M_1)\,P(M_1)}{P(E)} = \frac{(1/2)(1/3)}{(1/2)} = \boxed{\frac{1}{3}}$$

$P(E|M_1) = P(\text{HEADS} \mid \text{fair coin}) = \frac{1}{2}$

$P(E) = P(\text{HEADS}) = \frac{1}{2}$ because the bag contains 3 HEADS and 3 TAILS

Part B: Sensitivity and specificity (20 points)

Senioritis is a rare but treatable condition that occurs in about 20 out of every 10,000 individuals in the general population. A recently developed test for senioritis—the PUNT scan—is 99.5% sensitive and 80% specific. (This means that 995 out of 1000 people with senioritis correctly test positive, and 800 out of every 1000 people without senioritis correctly test negative. The PUNT scan always reports either "positive" or "negative".)

For notation, we can let D be the variable "You have senioritis" and let T be the variable "You test positive for senioritis". Then the information above is:

$P(D)$	20 out of 10,000
$P(T \mid D)$	995 out of 1,000
$P(\bar{T} \mid \bar{D})$	800 out of 1,000

Part B1 (10 points) What is the probability of obtaining a positive test result, regardless of whether you have senioritis?

The marginal probability of a positive test result is *approximately* (circle one):

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

You must show your work to receive credit. Write down the equations you intend to solve, if any, and indicate what values you're plugging in. You probably won't need a calculator, since you only need an approximate final answer.

$$P(T) = P(T|D) \cdot P(D) + P(T|\bar{D}) \cdot P(\bar{D})$$

$$= \frac{995}{1000} \cdot \frac{20}{10{,}000} + \left(1 - \frac{800}{1000}\right) \cdot \left(1 - \frac{20}{10{,}000}\right)$$

$$\approx 1 \cdot 0 + \frac{2}{10} \cdot 1$$

$$= 20\%$$

Part B2 (10 points) Suppose your PUNT scan returns a positive result. In this case, the probability that you indeed have senioritis is *most nearly* (circle one):

0%   10%   20%   30%   40%   50%   60%   70%   80%   90%   100%

<u>You must show your work to receive credit.</u> Write down the equations you intend to solve, if any, and indicate what values you're plugging in. You probably won't need a calculator, since you only need an approximate final answer.

$$P(D|T) = \frac{P(T|D)\,P(D)}{P(T)}$$

$$= \frac{\frac{995}{1000} \cdot \frac{20}{10{,}000}}{20\%}$$

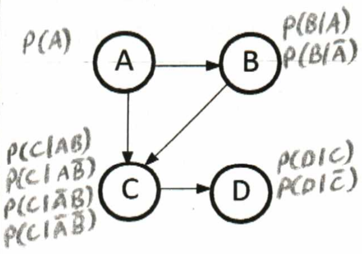
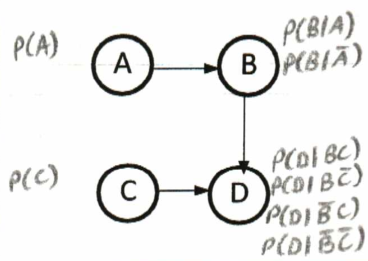
$$\approx \frac{1 \cdot 0}{0.2}$$

$$= 0\%$$

Part C: Applying d-separation (16 points)

In the figure below, there are two Bayes nets and some independence statements. For each of the statements below and each Bayes net, circle TRUE if the statement is true for the net, and FALSE if the statement is false for the net.

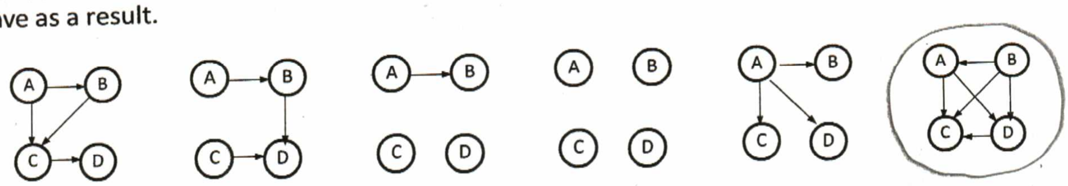
Note: Assume that the only independence statements that are true are the ones enforced by the shape of the network.

		
A is independent of C.	TRUE FALSE	TRUE FALSE
Given C, A is independent of D.	TRUE FALSE	TRUE FALSE
"Given A and C, B is indep. of D" $P(B\|DAC) = P(B\|AC)$	TRUE FALSE	TRUE FALSE
Assuming all of the variables are boolean, how many parameters does each Bayes net have? (The number of parameters is the total number of entries in all probability tables.)		
# of parameters	9	8

Part D: Make no assumptions (4 points)

In exactly **one** of the Bayes nets below, **none of the variables are independent**. Circle it.

Hint: A Bayes net that makes no independence assumptions is a model that is consistent with every joint probability table. You may find it useful to consider how many parameters it must have as a result.



Problem 3: Spiritual and Right-Now

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

1 Boris Katz's START system was an inspiration leading to:

5

1. The Deep Blue chess-playing program.
2. The introduction of subliminal advertising.
3. The Heritage system for genetic counseling.
4. The Google and Stanford systems for describing actions in pictures.
5. The Siri and Watson systems for question answering.

2 Probabilistic programing has been used with success to:

3

1. Analyze tweets.
2. Analyze voting data.
3. Analyze satellite data.
4. Improve athletic training.
5. Improve use of PowerPoint presentations in education.

3 A key idea behind the Genesis story-understanding system is the use of:

1

1. Rules and search for reasoning and conceptual analysis.
2. Constraint propagation to anticipate story outcomes.
3. Support vector machines to find common features in story collections.
4. Boosting to determine whether stories are interesting.
5. A small collection of standard descriptive frames to summarize stories.

4 A key idea behind the success of deep neural nets is:

1

1. Unsupervised, layer by layer training, aimed at producing outputs that are the same as inputs.
2. Use of dot-product kernel functions to transform difficult problems into easier problems.
3. Coupling small learning rates to massive computing to avoid local maxima.
4. Linear dependance of the net depth on the number of classes to be recognized.
5. Use of back propagation, accelerated by using a larger rate multiplier in layers closer to inputs.

5 Developing a most probable story-generation model from a collection of stories involves:

5

1. Recognizing similar stories using a nearest neighbor classifier.
2. Recognizing similar stories using a boosting classifier.
3. Assuming all possible story models are equally likely.
4. Averaging over a large number of stories using crowd sourcing to decide what to merge.
5. Using a carefully tuned prior that is a function of model size to guide Bayesian structure search.

6 When a rat stop moving on a track, activity in the hippocampus:

4

1. Stops completely.
2. Depends on whether the rat is hungry or recently fed.
3. Can be decoded to determine whether the rat is happy or stressed.
4. Sometimes replays activity as if the rat is moving forward, sometimes as if moving backward.
5. Bounces around, suggesting the rat has random thoughts about where it is and what it is doing.