# 6.034 Quiz 4
# 5 December 2012

| Name | Alyssa P. Hacker |
| --- | --- |
| 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.

| | | |
| --- | --- | --- |
| **Martin Gale** | **Dylan Holmes** | **Sarah Lehmann** |
| **Igor Malioutov** | **Robert McIntyre** | **Ami Patel** |
| **Sila Sayan*** | **Mark Seifter** | **Stephen Serene** |

*** If Sila is your TA, also circle the TA whose section you've been attending.**

| Problem number | Maximum | Score | Grader |
| --- | --- | --- | --- |
| 1 | 50 | | |
| 2 | 50 | | |
| Total | 100 | | |

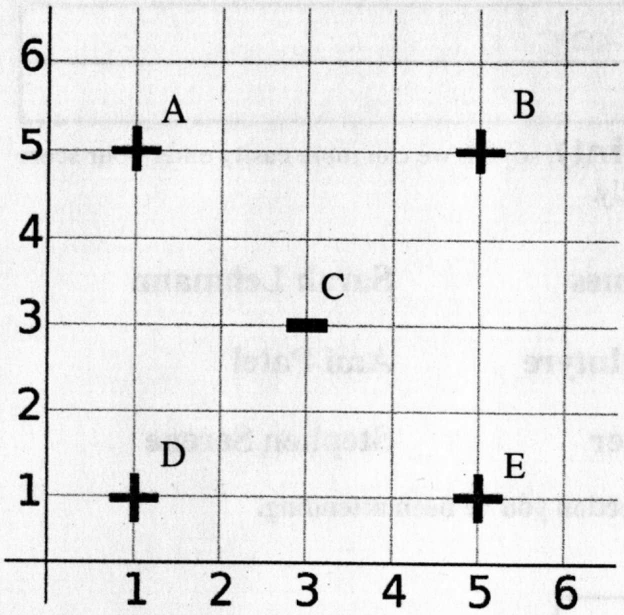
| Problem number | Maximum | Score | Grader |
| --- | --- | --- | --- |
| 3 | 7 | | |

**We recommend you reserve a few minutes for problem three.**

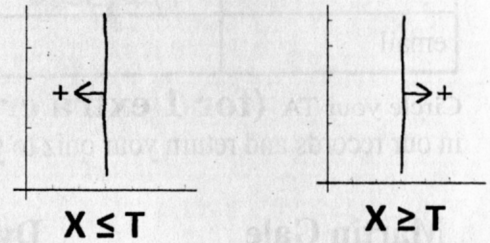
**There are 9 pages in this quiz. As always, open book, open notes, open just about everything, including a calculator, but no computers.**

# Problem 1: AdaBoost (50 pts)

In this problem, you'll use boosting to construct a classifier for the following training dataset:




For your collection of weak classifiers, you'll use vertical line tests with either of the following forms




$$h(x, y)=\begin{cases}+1 & if\ x \leq T\\ -1 & if\ x > T\end{cases} \qquad h(x, y)=\begin{cases}+1 & if\ x \geq T\\ -1 & if\ x < T\end{cases}$$

The first kind of test classifies a point as positive if it's to the LEFT of a certain vertical line.

The second kind of test classifies a point as positive if it's to the RIGHT of a certain vertical line.

### *Part A (10 pts)*

**A1 (5 pts)** For the following set of six classifiers, list all the training points (A, B, C, D, E) that each misclassifies. If a classifier misclassifies no points, write NONE instead.

| Classifier | Misclassified Training Points |
|---|---|
| X ≤ 2 | BE |
| X ≤ 4 | CBE |
| X ≤ 6 | C |

| Classifier | Misclassified Training Points |
|---|---|
| X ≥ 2 | ACD |
| X ≥ 4 | AD |
| X ≥ 6 | ABDE |

**A2 (5 pts)** Which of the six classifiers from this list, if any, will surely never have the least error? If none of the classifiers can be eliminated for sure, write NONE instead.

X ≤ 4, X ≥ 2, X ≥ 6

* Each of these classifiers makes a superset of the misclassifications that another classifier makes.

## *Part B, Three rounds (24 pts)*

**Perform three iterations of boosting using the training points (A, B, C, D, E) and the six classifiers from the table in Part A. In case of a tie, use whichever classifier comes first in this list:**

**$X \le 2$   $X \le 4$   $X \le 6$   $X \ge 2$   $X \ge 4$   $X \ge 6$**

**Fill out this table with the weights of the training points at the start of each round, the weak classifier chosen each round, the error of that weak classifier, and its coefficient.**

| | Round 1 | Round 2 | Round 3 |
|---|---|---|---|
| **weight$_A$** | 1/5 | 1/8 | 1/12 |
| **weight$_B$** | 1/5 | 1/8 | 3/12 |
| **weight$_C$** | 1/5 | 4/8 | 4/12 |
| **weight$_D$** | 1/5 | 1/8 | 1/12 |
| **weight$_E$** | 1/5 | 1/8 | 3/12 |
| **weak classifier (h)** | $X \le 6$ | $X \le 2$ | $X \ge 4$ |
| **classifier error (ε)** | 1/5 | 1/4 | 1/6 |
| **coefficient (α)** | $\frac{1}{2}\ln 4$ | $\frac{1}{2}\ln 3$ | $\frac{1}{2}\ln 5$ |

**(Show your work in the space below for partial credit.)**

AFTER ROUND 1:

C is misclassified.

NEW WEIGHT:

$\frac{1}{2} \cdot \frac{1}{1/5} \cdot \frac{1}{5} = 1/2$

A, B, D, E correctly classified.

NEW WEIGHT:

$\frac{1}{2} \frac{1}{4/5} \frac{1}{5} = 1/8$

AFTER ROUND 2:

B, E misclassified

NEW WEIGHT:

$\frac{1}{2} \frac{1}{1/4} \frac{1}{8} = \frac{1}{4}$

C correctly classified

$\frac{1}{2} \frac{1}{3/4} \frac{4}{8} = \frac{1}{3}$

A, D correctly classified

$\frac{1}{2} \frac{1}{3/4} \frac{1}{8} = \frac{1}{12}$

## *Part C, Retrospective (16 pts)*

**C1 (5 pts)** **Consider the total classifier $H$ which you produce after three rounds of boosting. List all the training points (A, B, C, D, E) which $H$ misclassifies. If there are none, write NONE instead.**

NONE

**Show your work for partial credit. You may find the following logarithms helpful:**

$$\log(1)=0 \qquad \log(2)\approx\frac{7}{10} \qquad \log(3)\approx 1 \qquad \log(4)\approx\frac{7}{5} \qquad \log(5)\approx\frac{8}{5}$$

$$\log(ab)=\log(a)+\log(b)$$

The easiest way to see this is that all three weak classifiers make disjoint mistakes.

Otherwise:

| TRAINING POINT | $X \leq 6$ | $X \leq 2$ | $X \geq 4$ | | |
|---|---|---|---|---|---|
| A | +1 | +1 | −1 | $7/10 + \frac{1}{2} - \frac{8}{10} = \frac{4}{10}$ | positive ✓ |
| B | +1 | −1 | +1 | $7/10 - \frac{1}{2} + \frac{8}{10} = \frac{10}{10}$ | positive ✓ |
| C | +1 | −1 | −1 | $7/10 - \frac{1}{2} - \frac{8}{10} = \frac{-6}{10}$ | negative ✓ |
| D | +1 | +1 | −1 | $7/10 + \frac{1}{2} - \frac{8}{10} = \frac{4}{10}$ | positive ✓ |
| E | +1 | −1 | +1 | $7/10 - \frac{1}{2} + \frac{8}{10} = \frac{10}{10}$ | positive ✓ |

**C2 (9 pts)** Now that you've finished boosting, your friends Alyssa, Ben, and Cy want to offer their advice.

Alyssa says: "You don't need boosting to find a good combination of those classifiers! Look at the table in Part A—you can see immediately that by combining **three** of those classifiers, **all with α = 1**, you could classify the training data perfectly."

Ben says: "Actually, you don't even need three classifiers—by combining just **two** of those classifiers, both **with α = 1**, you could classify the training data perfectly. you could classify the training data perfectly."

Cy says: "In fact, I see **one** classifier in the table that can classify the training data perfectly."

In the space below, concisely explain why each friend is correct or incorrect.

**Alyssa:** Correct. Three of the classifiers in the list make disjoint errors.

**Ben:** Incorrect. Although there are pairs of classifiers with disjoint errors, you must use more than two classifiers in order to avoid a tied vote.

**Cy:** Incorrect. Each of the weak classifiers in the list misclassifies at least one training point.

**C3 (2 pts)** Alyssa makes a mistake when programming Adaboost: instead of choosing the weak classifier with the <u>lowest</u> error, her program chooses the weak classifier with the <u>highest</u> error. Ben Bitdiddle claims that this algorithm will pick the same classifier in every round. Explain whether Ben is correct or incorrect.

Ben is incorrect; Alyssa's algorithm will not loop forever, picking the same classifier each round.

Instead, the algorithm will first pick the classifier with the highest error; after reweighting, that classifier will have an error of exactly ½. Therefore, as long as there are classifiers with error > ½ available, those will be preferred by Alyssa's algorithm.

TA Note: You can try out Alyssa's algorithm on this boosting problem to see how it works.

# Problem 2: Bayes Nets (50 points)

## *Part A(40 pts)*

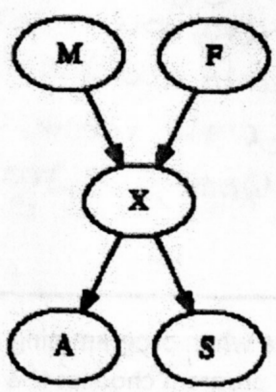
**As part of her UROP project, Alyssa Hacker decides to model the expression of the elusive Gene X. Here are the behaviors she has observed so far:**

- **A child is more likely to express Gene X if her mother or father expresses Gene X.**
- **Anyone who expresses Gene X often develops astonishing math skills.**
- **Anyone who expresses Gene X is at risk of becoming a social misfit.**

**After collecting some sample data, Alice produces the following Bayes net. Here, X denotes "Patient expresses Gene X", M denotes "Patient's mother expresses Gene X", F denotes "Patient's father expresses Gene X", A denotes "Patient develops astonishing math skills", and S denotes "Patient becomes a social misfit".**

| P(M) |
|---|
| 0.1 |

| P(F) |
|---|
| 0.1 |




| M | F | P(X\|MF) |
|---|---|---|
| T | T | 0.8 |
| T | F | 0.5 |
| F | T | 0.6 |
| F | F | 0.01 |

| X | P(A\|X) |
|---|---|
| T | 0.7 |
| F | 0.03 |

| X | P(S\|X) |
|---|---|
| T | 0.8 |
| F | 0.4 |

**A1 (10 pts)** **Use the above network to determine which of the following probability statements are TRUE or FALSE. Circle the correct answer in each case.**

| | | |
|---|---|---|
| P(F\|M) = P(F) | TRUE | FALSE |
| P(S\|A,F,M) = P(S\|F,M) | TRUE | FALSE |
| P(X\|M)=P(M\|X) | TRUE | FALSE |
| P(S\|X,A) = P (S\|X) | TRUE | FALSE |

**A2 (10 pts)** Suppose Alyssa's mother expresses Gene X. Compute the probability that Alyssa expresses Gene X given only this observation. **(Your answer should be numerical.)**

$$P(X|M) = P(F)\cdot P(X|FM) + P(\bar{F})\cdot P(X|\bar{F}M)$$
$$= \left(\frac{1}{10}\right)\left(\frac{8}{10}\right) + \left(\frac{9}{10}\right)\left(\frac{5}{10}\right)$$
$$= \frac{53}{100}$$

**A3 (10 pts)** Alyssa later learns that her father also expresses Gene X. How will the probability that Alyssa expresses Gene X, given this new information and the fact that Alyssa's mother expresses Gene X compare to the probability you computed in A2? **(Circle the best answer.)**

Smaller Equal Bigger

For credit, you **must** concisely explain the intuition behind this result.

This situation is a "noisy-OR" configuration: the probability that Alyssa expresses Gene X is increased if we know that her mother or father expresses Gene X, or especially if both parents express Gene X. We can see these probabilities using our answer to A2, above, and reading $P(X|MF)$ off of the Bayes Net.

**A4 (10 pts)** Write an **algebraic expression** for the probability that a child develops astonishing math skills and the child's father expresses Gene X. Express your answer in terms of probabilities you could read off directly from the Bayes network. **(You do not need to calculate a numerical answer)**

$$P(A,F) = \sum_{\substack{\text{possible}\\ M}} \sum_{\substack{\text{possible}\\ X}} P(A{=}T|X)\cdot P(X|M,F{=}T)\cdot P(M)\cdot P(F{=}T)$$

### *Part B (10 pts)*

After his graduation, Ben Bitdiddle decides to capitalize on his hard-earned MIT skills as part of his new start-up – Emoticon Traders. He wants to make a lot of money making trades by determining whether financial news articles are positive (BUY) or negative (SELL).

Ben decides to use Naive Bayes classification for the task. He has chosen six words to use as features for detecting BUY/SELL outcomes. The words are:

**bullish, hot, junk, stock, the, company**

And the probabilities associated with each are:

| | | | |
|---|---|---|---|
| P(BUY) = 0.4 | P( bullish \| BUY) = 0.7 | P( hot \| BUY) = 0.8 | P( junk \| BUY) = 0.4 |
| P(SELL) = 0.6 | P( bullish \| SELL) = 0.2 | P( hot \| SELL) = 0.1 | P( junk \| SELL) = 0.9 |
| | P( stock \| BUY) = 0.8 | P( the \| BUY) = 0.95 | P( company \| BUY) = 0.9 |
| | P( stock \| SELL) = 0.8 | P( the \| SELL) = 0.95 | P( company \| SELL) = 0.9 |

How will Ben's system classify an article that contains just the following four out of six keywords: "the", "hot", "junk", and "company"? **(Circle one.)**

BUY SELL

**You must** show work for credit.

Naive Bayes

$P(Y \,\&\, the \,\&\, hot \,\&\, junk \,\&\, company \,\&\, \overline{bullish} \,\&\, \overline{stock})$

$= P(Y)\,P(the|Y)\,P(hot|Y)\,P(junk|Y)\,P(company|Y)\,P(\overline{bullish}|Y)\,P(\overline{stock}|Y)$

when $Y=BUY$: $(.4)(.95)(.8)(.4)(.9)(.3)(.2)$

when $Y=SELL$: $(.6)(.95)(.1)(.9)(.9)(.8)(.2)$ ← This probability is larger, so SELL

(prior) the hot junk company $\overline{bullish}$ $\overline{stock}$

## 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)** Brooks's subsumption architecture features:

3

1. Speed, by way of direct calls to the Unix kernel.
2. Power, by way of sophisticated world models .
3. Robust action, by way of functional abstraction.
4. Flexibility, by way of implementing a universal Turing machine.
5. Error recovery, by way of fault-tolerant hardware.

**2)** Katz's START system suggested the feasibility of:

3

1. Military robots.
2. Minsky's layered theory of problem solving.
3. Watson and Siri.
4. Applications of AI in resource allocation.
5. Applications of AI in medicine.

**3)** Mansinghka's probabilistic programming talk suggested application in:

3

1. National reconciliation following civil wars.
2. Tracing neural pathways in the nematode *Caenorhabditis Elegans*.
3. Determining compliance with the nuclear test ban treaty.
4. Simulating protein folding patterns by running programs backwards.
5. Prevention of identity theft.

**4)** Borchardt's transition-space representation is based on the idea that in human thinking:

4

1. State leads to state.
2. State leads to state change.
3. State change leads to state.
4. State change leads to state change.

5) General Problem Solver (GPS) is an architecture based on

1

1. Means-ends analysis.
2. Flow of information through short term memory.
3. Developing common sense rules through visual observation.
4. Learning to prevent mistakes by way of negative examples.
5. Sussman's box and wire metaphor.

**6)** Reification occurs in a semantic net when:

2

1. Meaning emerges from human interpretation.
2. A relation is treated as a node.
3. A subject is connected to an object via a relation.
4. Trajectories and transitions are expressed in subject-relation-object triples.
5. A prince becomes a king.

**7)** The Church programming language features a sophisticated, built-in mechanism for:

4

1. Constraint propagation.
2. Rule chaining.
3. Building goal trees from program executions.
4. Searching program execution spaces.
5. Boosting classifiers.