CS 5350/6350: Machine Learning, Fall 2015

Midterm Exam

October 8, 2015

Instructions

- Make sure that your exam has 11 pages (excluding this cover page).
- This is a closed book, closed notes exam. You will not need a calculator. Everything you need in order to solve the problems is supplied here.
- There is an appendix with possibly useful formulas and computational short cuts at the end. If there is any complex calculation involved, you may leave the result as a fraction.
- The exam ends at 3:20 PM. You have 80 minutes to answer them to earn 100 points.
- Answer each question in the space provided. If you need more room, write on the back side of the paper and indicate that you have done so.
- **Besides having the correct answer, being concise and clear is very important. For full credit, you must show your work and explain your answers.**
- It is important that we should be able to understand your answer. **If you have made a mess, clearly mark the answer.**

Good luck!

Name: (2 points)

Problem 1 (8%):	
Problem 2 (30%):	
Problem 3 (20%):	
Problem 4 (25%):	
Problem 5 (15%):	
Total:	

1 True/False Questions

State whether the following are true or false. No explanation is needed. Each question is worth one point.

1. During the training phase of batch learning, you can use the testing examples to further improve your classifier.

   $False$

2. Smaller hypothesis spaces are better for generalization.

   $True$

3. Decision trees can only represent conjunctions, disjunctions and m-of-n Boolean functions.

   $False$

4. For any given dataset, there can be no more than one decision tree that is consistent with it.

   $False$

5. Training a decision tree can be more computationally expensive than training a nearest neighbors classifier.

   $True$

6. The Halving algorithm is not always practically viable from the computational perspective.

   $True$

7. The balanced variant of the Winnow algorithm can learn any linear classifier while the regular Winnow can only learn a subset of linear classifiers.

   $True$

8. The mistake bound model of learning assumes that training and test examples are sampled from the same probability distribution.

   $False$

2 Multiple Choice Questions

[30 points] In the questions below, **more than one answer *may* be correct.** Each question is worth three points. Circle *all* the correct answers (and only the correct answers) for the points. No partial credit will be awarded.

1. Which of the following statements about k-nearest neighbors classifiers are correct?
   (a) **As the number of training examples goes to infinity, the nearest neighbors classifiers can get arbitrarily good accuracy.**
   (b) **If you increase the value of k, after it crosses the number of training examples, the predictions of the classifier on test examples will never change.**
   (c) When k is decreased, the performance of the classifier will always decrease.
   (d) When k is increased, the performance of the classifier will always increase.

2. Use the following two data sets, each with three two-dimensional examples and binary labels and select the assertions that are valid.

Data set A		**Data set B**	
(x_1, x_2)	y	(x_1, x_2)	y
$(1, 1)$	$+1$	$(1, 1)$	-1
$(2, -1)$	-1	$(2, -1)$	$+1$
$(3, 1)$	$+1$	$(3, 1)$	-1

   (a) **The entropies of the labels for the two data sets are the same.**
   (b) **For both data sets, it is possible to find a linear classifier that correctly classifies all points.**
   (c) Exactly one of the two data sets is linearly separable.
   (d) **It is possible to add additional labeled points to both datasets that render them linearly inseparable.**

3. You have a labeled training set with 800 examples that are 100-dimensional and train a classifier using the Perceptron algorithm for twenty epochs. You find that after training, the classifier perfectly labels all training examples. Which of the following inferences are valid?
   (a) You can be sure that the classifier will be correct on all future examples.
   (b) **The training set is linearly separable.**
   (c) The Winnow algorithm would have been an unambiguously better choice for this setting.
   (d) If you had run the Perceptron algorithm for more epochs after the initial twenty, the classifier would improve to a better one.

4. You have a learning problem with 10000 features. You suspect that no more than 40 of them are relevant and want to choose a learning algorithm based on this. Then,
   (a) A nearest neighbors method is preferable to using decision trees.
   (b) **Decision trees are preferable to using nearest neighbors.**
   (c) An additive update algorithm is a better choice than a multiplicative update algorithm
   (d) **A multiplicative update algorithm is a better choice than an additive update algorithm.**

5. Select the valid statements from below.

   (a) **k-NN does not require an explicit training step.**

   (b) The decision boundary of the k-NN classifier is always linear.

   (c) The decision boundary of the 1-NN classifier is always linear.

   (d) **For the nearest neighbors method, using the Euclidean and the squared Euclidean distance are equivalent.**

6. You are stranded on a desert island and you must assess the edibility of the papayas you find on it. You find that they are characterized by ten binary features: good smell, squishy, red color, green color, etc. You want to learn a classifier to decide whether the papayas are tasty or not and you figure that the tastiness can be determined by a monotone conjunction of the ten features. How many classifiers are in your hypothesis space?

   (a) $\mathbf{2}^{10}$

   (b) $2^{2^{10}}$

   (c) 3^{10}

   (d) This is a trick question that can not be answered using the above information.

7. Which of the following statements about the Winnow algorithm are correct?

   (a) **It makes $O(k \log n)$ mistakes for k-disjunctions with n variables.**

   (b) **It is a multiplicative update algorithm.**

   (c) It is an additive update algorithm.

   (d) For a dataset representing k-disjunctions, the number of mistakes it makes is inversely proportional to the square of the margin of the data.

8. Which of the following statements about the Perceptron algorithm are correct?

   (a) After a Perceptron update on an example, the resulting weight vector will correctly classify that example.

   (b) It is a multiplicative update algorithm.

   (c) **It is an additive update algorithm.**

   (d) **For any linearly separable data, the number of mistakes it makes is inversely proportional to the square of the margin of the data.**

9. Which of the following functions can be represented by a linear classifier **without** any transformations? (Assume that the x_i's are Boolean here)

   (a) $\mathbf{x}_1 \vee \mathbf{x}_2$

   (b) $\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3$

   (c) $(x_1 \vee \neg x_2) \wedge (\neg x_1 \vee x_2)$

   (d) None of the above

10. Which of the following functions can be represented by a linear classifier **after** feature transformations? (Assume that the x_i's are Boolean here)

    (a) $\mathbf{x}_1 \vee \mathbf{x}_2$

    (b) $\mathbf{x}_1 \wedge \mathbf{x}_2 \wedge \mathbf{x}_3$

    (c) $(\mathbf{x}_1 \vee \neg \mathbf{x}_2) \wedge (\neg \mathbf{x}_1 \vee \mathbf{x}_2)$

    (d) None of the above

3 Decision Trees

Table 1 shows a modified version of the Shuttle Landing Control data set from the UCI machine learning repository. There are three features called **Err.**, **Wind** and **Magnitude**. The target concept is the column titled **Label**, which indicates whether an autolanding would be preferable to manual control of a spacecraft.

#	**Err.**	**Wind**	**Magnitude**	**Label**
1	L	head	Low	Auto
2	X	tail	Low	Auto
3	M	head	Low	Auto
4	M	head	Medium	NoAuto
5	M	tail	Low	NoAuto
6	M	tail	Medium	NoAuto
7	M	head	Strong	Auto
8	M	tail	Strong	NoAuto

Table 1: A subset of the Shuttle Landing Control Data Set

1. [3 points] What is the entropy of the collection of examples with respect to the target label?

   Solution. #(Auto) = 4, #(NoAuto) = 4 Total= 8

$$\begin{aligned} Entropy(S) &= -p^{+}\log(p^{+}) - p^{-}log(p^{-}) \\ &= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) \\ &= -\frac{1}{2}[\log(1) - \log(2)] - \frac{1}{2}[\log(1) - \log(2)] \\ &= -\frac{1}{2}[0-1] - \frac{1}{2}[0-1] \\ &= \frac{1}{2} + \frac{1}{2} \\ &= 1 \end{aligned}$$

2. [3 points] What is the entropy of the attribute **Magnitude**?

   Solution:

   Low $= \frac{4}{8}$, **Medium** $= \frac{2}{8}$, **Strong** $= \frac{2}{8}$. So we can calculate entropy of this distribution as

$$\begin{aligned} E &= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) \\ &= -\frac{1}{2}\log(\frac{1}{2}) - \frac{1}{2}\log(1/4) \\ &= \frac{1}{2} + 1 = \frac{3}{2} \end{aligned}$$

   OR (we also accepted):

   Low $= \frac{4}{8}$ $p = \frac{3}{4}$ $n = \frac{1}{4}$, **Medium** $= \frac{2}{8}$ $p = 0$ $n = \frac{2}{2}$, and **Strong**$= \frac{2}{8}$ $p = \frac{1}{2}$ $n = \frac{1}{2}$. Now calculate

entropy of each:

$$\begin{aligned} E(\text{Low}) &= -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) \\ &= -\frac{3}{4}\frac{3}{2} + 2\frac{3}{4} + \frac{1}{2} \\ &= \frac{7}{8} = .825 \\ E(\text{Medium}) &= 0 \\ E(\text{Strong}) &= 1 \\ \mathbb{E}[\text{Entropy}] &= \frac{1}{2} \times \frac{7}{8} + \frac{1}{4} \times 0 + \frac{1}{4} \times 1 \\ &= 11/16 = .6875 \end{aligned}$$

3. [6 points] Compute the information gain of the attribute **Wind?**
   Solution head= $\frac{4}{8}$ $p = \frac{3}{4}$ $n = \frac{1}{4}$, **tail** = $\frac{4}{8}$ $p = \frac{1}{4}$ $n = \frac{3}{4}$

$$\begin{aligned} \text{head} &= -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) \\ &= \frac{7}{8} = .825 \\ \text{tail} &= -\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) \\ &= \frac{7}{8} = .825 \\ \mathbb{E}[\text{wind}] &= \frac{1}{2}\frac{7}{8} \times + \frac{1}{2} \times \frac{7}{8} \\ &= \frac{7}{8} = .825 \\ IG(\text{wind}) &= 1 - .825 \\ &= .125 \end{aligned}$$

4. [8 points] Choose **Wind** as the root node for the decision tree. With this root node, write down a decision tree that is consistent with the data. You **do not** need to show how the decision tree is learned; just make sure it is consistent with the data.

   Solution: Answers will vary

4 Mistake Bound Model of Learning

Consider an instance space consisting of **integer points** on the two dimensional plane (x_1, x_2) with $-128 \leq x_1, x_2 \leq 128$. Let $\mathcal{C}$ be a concept class defined on this instance space. Each function in the concept class is defined using an integer l (with $0 < l \leq 128$) as follows

$$f_l(x_1, x_2) = \begin{cases} +1 & |x_1| \leq l \text{ and } |x_2| \leq l; \\ -1 & \text{otherwise} \end{cases} \tag{1}$$

That is, each function f_l labels points within the origin-centered square of side $2l$ as positive and points outside as negative. Our goal is to come up with a mistake-driven algorithm that can learn this concept class. This question guides you through the process of constructing and analyzing such a learning algorithm

1. [3 points] Determine the number of functions in this concept class $|\mathcal{C}|$.

   Solution: The number of functions is 128 because there is one function for each value of l. (We accepted minor variations caused by misinterpretations of the boundary conditions.)

2. [4 points] To design an error driven learning algorithm, we should be able to first write down what it means to make a mistake. Suppose our current guess for the function is f_l defined as in Equation 1 above. Say we get an input point (x_1, x_2) along with its label y. Write down expression in terms of x_1, x_2, y and l that checks whether the current hypothesis f_l has made a mistake. (Hint: Two inequalities.)

   Solution: $y(|x_1| - l) \geq 0$ and $y(|x_2| - l) \geq 0$.

   We accepted many variations of this answer. For example,

   $(|x_1| < l \text{ and } |x_2| < l \text{ and } y = -1)$ or $(|x_1| \geq l \text{ and } |x_2| \geq l \text{ and } y = +1)$

3. The next step in designing the mistake driven algorithm is to define how the algorithm updates the hypothesis on an error. Since the function is completely defined in terms of l, you need to specify how to update l on a mistake.

   (a) [4 points] How will you update l if the current hypothesis makes a mistake on a positive example?
   Solution: The value of l should be increased additively. Valid answers include $l \leftarrow l + 1$ and $l = \max(|x_1|, |x_2|)$. The first answer simply increases l, while the second one increases l to be big enough to include the current example inside the rectangle.

   (b) [4 points] How will you update l if the current hypothesis makes a mistake on a negative example?
   Solution: The value of l should be decreased. Valid answers include $l \leftarrow l - 1$ and $l = \max(|x_1|, |x_2|) - 1$.

4. [4 points] Using the answers from the previous two steps, write down a mistake-bound algorithm for the concept class. Please write this algorithm concisely.

   Solution: Several algorithms are possible, based on the update and the initialization for l. Here is one possible answer:

   (a) $l \leftarrow 1$

   (b) For each example $((x_1, x_2), y)$:

      i. $y' = f_l(x_1, x_2)$. That is, use the current value of l to predict.

      ii. if $y' \neq y$, then $l = \max(|x_1|, |x_2|)$.

   This algorithm will never make mistakes on negative examples because of the initialization, so we only have one kind of update. Alternative algorithms are possible.

5. [5 points] What is the maximum number of mistakes that this algorithm can make on *any* data set? Why?

   Solution: The above algorithm can make no more than 128 mistakes because there are only 128 possible functions. (We accepted minor variations of this answer.)

5 Linear Threshold Units

We have a learning problem whose inputs are 1000 dimensional Boolean inputs $x_1, x_2, \cdots x_{1000}$. Consider a concept class $\mathcal{C}$ defined on this instance space as follows:

> Every function f in the concept class $\mathcal{C}$ is specified using a set of 20 features $x_{f_1}, x_{f_2}, \cdots, x_{f_{20}}$ as $f(\mathbf{x}) = \neg x_{f_1} \vee \neg x_{f_2} \vee \cdots \vee \neg x_{f_{20}}$.

Here, $f_1, f_2, \cdots, f_{20}$ are feature indices each of which can range from 1 to 1000 and $\mathbf{x}$ denotes the 1000-dimensional input vector. (These indices indicate which features are relevant to the function f and thus, each function in $\mathcal{C}$ is uniquely defined by its choice of these indices.)

1. [5 points] What is the size of the concept class?

   Solution: It will be n choose k where $n = 1000$ and $k = 20$ i.e. $\frac{1000!}{980!20!}$. In words we can choose 20 different functions out of 1000! but order doesn't matter.

2. [5 points] Can the function $f(\mathbf{x}) = \neg x_1 \vee \neg x_2 \vee \cdots \vee \neg x_{20}$ in the concept class be represented by a linear threshold unit? If so, write down a linear threshold unit for this function. (Remember from class that a linear threshold function is defined by a weight vector and a bias. You need to specify both here.) If not, show a counter example that shows that for any linear threshold unit, you can find a point that will be incorrectly classified.

   Solution: Yes this can be represented by a linear function:

$$\begin{aligned} 1 &\geq (1 - x_1) + (1 - x_2) + ... + (1 - x_{20}) \\ &\geq 20 - x_1 - x_2 - ... - x_{20} \\ -19 &\geq -x_1 - x_2 - ... - x_{20} \\ 19 &\leq x_1 + x_2 + ... + x_{20} \end{aligned}$$

3. [5 points] Would you prefer to use the Perceptron, Winnow or the balanced Winnow algorithm to learn this set of functions? Why? (Please be concise with your argument here.)

   Solution: Balanced Winnow because we are dealing with a small number of pertinant features and it is not a monotone conjunction.

Some formulas you *may* find useful

- $P(A, B) = P(A|B)P(B)$
- $Entropy(S) = -p^+ \log(p^+) - p^- log(p^-)$, where S is a labeled set and p^+ and p^- are the fraction of positive and negative points in it.
- $Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$
- $\log(\frac{a}{b}) = \log(a) - \log(b)$
- $\log_2(3) \approx \frac{3}{2}$, $\log_2(10) \approx 3\frac{1}{3}$
- $\sin \frac{\pi}{e} \approx 0.915$

Blank page for scratch work

Blank page for scratch work