CS 5350/6350: Machine Learning, Fall 2015

Sample Midterm Questions

Here are a set of questions to give a flavor of the midterm exam. (The actual midterm will not be as long as this.) Feel free to discuss these questions with the instructor, the TAs and other students.

1. How would you train a decision tree using the ID3 algorithm if some attributes are missing? (You might get asked to step through this procedure for a small dataset like the Tennis data in the lecture.)

2. Show that the following dataset is linearly separable by providing a linear threshold unit that correctly classifies the examples.

x_1	x_2	x_3	y
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0

3. How would you avoid overfitting when you use the decision tree algorithm? Why might shorter decision trees be more robust to noise in the training data?

4. (An exercise question from the class lectures) Suppose you want to build a nearest neighbors classifier to predict whether a beverage is a coffee or a tea using two features: the volume of the liquid (in milliliters) and the caffeine content (in grams). You collect the following data:

Volume (ml)	**Caffeine (g)**	Label
238	0.026	Tea
100	0.011	Tea
120	0.040	Coffee
237	0.095	Coffee

   What is the label for a test point with Volume = 120, Caffeine = 0.013? Why might this be incorrect? How would you fix the problem?

5. (An exercise question from the class lectures) What will happen when you choose K to the number of training examples for a K-nearest neighbor classifier?

6. For each function below, state whether it can be written as a linear threshold unit in terms of the variables specified. If it can be written as one, write the linear threshold unit that is equivalent to the function. If not, suggest a transformation of the underlying space so that the function is linear in the new space.

   (a) $\neg x_1$

(b) $x_1 \vee \neg x_2$

(c) $(x_1 \vee \neg x_2) \wedge (\neg x_1 \vee x_3)$

7. Show that the Halving algorithm for a finite concept space C will not make more than $\log |C|$ mistakes. Apply this to get a limit on the number of mistakes the algorithm will make for the class of k-conjunctions of n Boolean variables.

8. State with an explanation whether the following are true or false.

   (a) The mistake bound model assumes that training and test examples are drawn from the same fixed, but unknown distribution.

   (b) The Perceptron mistake bound theorem guarantees that the algorithm will find a linear separator for *any* dataset.

   (c) Unlike online learning, batch learning does not seek to minimize the number of mistakes that the learner makes.

9. Prove the Perceptron mistake bound.

10. How many mistakes will the Perceptron algorithm make for disjunctions with n attributes? To answer this, you will first have to identify what R and γ are for this concept class.

11. Prove the Winnow mistake bound.

12. You are given a binary classification dataset where the examples are 100000 dimensional Boolean vectors. You suspect that the true classifier could not be a function of more than 100 features. Given this information would you prefer using the Perceptron or the Winnow algorithm for learning? Why?

13. You wish to learn a hidden concept f using m training examples that are drawn from a distribution D. If the training set is called S and the hypothesis that your learning generates is h, write expressions for the training and generalization errors.

14. Suppose our learning problem has n binary features. What is the size of the hypothesis space consisting of all decision trees over this space?