MIDTERM: Spring 2012
CS 6375
INSTRUCTOR: VIBHAV GOGATE

March 28, 2012

The exam is closed book. You are allowed a double sided one page cheat sheet. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page. Attach your cheat sheet with your exam.

NAME _____

UTD-ID if known _____

- Problem 1: _____
- Problem 2: _____
- Problem 3: _____
- Problem 4: _____
- Problem 5: _____
- Problem 6: _____

- TOTAL: _____

PROBLEM 1: TRUE/FALSE QUESTIONS (10 points)

1. (2 points) Naive Bayes is a linear classifier. True or False. Explain.

2. (2 points) Classifier A has 90% accuracy on the training set and 75% accuracy on the test set. Classifier B has 78% accuracy on both the training and test sets. Therefore, we can conclude that classifier A is better than classifier B (because it has better mean accuracy). True or False. Explain.

3. (2 points) Consider a data set that is linearly separable and two perceptrons, one trained using the gradient descent rule and the other trained using the perceptron rule. Both perceptrons will have the same accuracy on the training set and the test set. True or False. Explain.

4. (2 points) Given infinite training data over n Boolean attributes, we can always learn the target concept using a decision tree but not using Naive Bayes or Logistic Regression. True or False. Explain.

5. (2 points) A decision tree is the smallest possible representation of the target concept. In other words, there exists no other representation that is smaller than the decision tree. True of False. Explain.

PROBLEM 2: LINEAR REGRESSION (10 points)

Consider a linear regression problem $y = w_1 x + w_0$, with a training set having m examples $(x_1, y_1), \ldots, (x_m, y_m)$. Suppose that we wish to minimize the mean $5 - th$ degree error (loss function) given by:

$$Loss = \frac{1}{m} \sum_{i=1}^{m} (y_i - w_1 x_i - w_0)^5$$

1. (3 points) Calculate the gradient with respect to the parameter w_1. Hint: $\frac{dx^k}{dx} = kx^{k-1}$.

2. (4 points) Write down pseudo-code for online gradient descent on w_1 for this problem. (You do not need to include the equations for w_0)

3. (3 points) Give one reason in favor of online gradient descent compared to batch gradient descent, and one reason in favor of batch over online.

PROBLEM 3: CLASSIFICATION (10 points)

Imagine that you are given the following set of training examples. Each feature can take on one of three nominal values: a, b, or c.

F1	F2	F3	Category
a	c	a	+
c	a	c	+
a	a	c	−
b	c	a	−
c	c	b	−

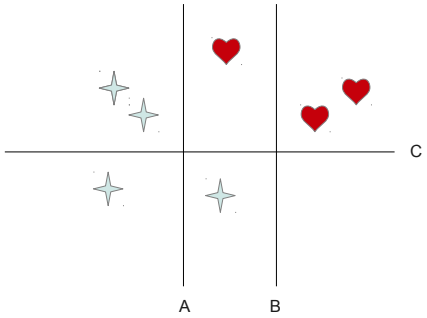
1. (5 points) How would a Naive Bayes system classify the following test example? Be sure to show your work.

   F1 = a, F2 = c , F3 = b

2. (5 points) Describe how a 3-nearest-neighbor algorithm would classify the test example given above.

PROBLEM 4: BOOSTING, LOGISTIC REGRESSION and POINT ESTIMATION(10 points)

1. (4 points) The diagram shows training data for a binary concept where positive examples are denoted by a heart. Also shown are three decision stumps (A, B and C) each of which consists of a linear decision boundary. Suppose that AdaBoost chooses A as the first stump in an ensemble and it has to decide between B and C as the next stump. Which will it choose? Explain. What will be the ϵ and α values for the first iteration?



2. (3 points) When learning a logistic regression classifier, you run gradient ascent for 50 iterations with the learning rate, $\eta = 0.3$, and compute the conditional log-likelihood $J(\theta)$ after each iteration (where θ denotes the weight vectors). You find that the value of $J(\theta)$ increases quickly then levels off. Based on this, which of the following conclusions seems most plausible? Explain your choice in a sentence or two.

   A. Rather than use the current value of η, it'd be more promising to try a larger value for the learning rate (say $\eta = 1.0$).

   B. $\eta = 0.3$ is an effective choice of learning rate.

   C. Rather than use the current value of η, it'd be more promising to try a smaller value (say $\eta = 0.1$).

3. (3 points) (**Point Estimation**) You are given a coin and a thumbtack and you put Beta priors $Beta(100, 100)$ and $Beta(1, 1)$ on the coin and thumbtack respectively. You perform the following experiment: toss both the thumbtack and the coin 100 times. To your surprise, you get 60 heads and 40 tails for both the coin and the thumbtack. Are the following two statements true or false.

   - The MLE estimate of both the coin and the thumbtack is the same but the MAP estimate is not.
   - The MAP estimate of the parameter θ (probability of landing heads) for the coin is greater than the MAP estimate of θ for the thumbtack.

   Explain your answer mathematically.

PROBLEM 5: SUPPORT VECTOR MACHINES (10 points)

Consider the following 1-dimensional data:

x	-3	0	1	2	3	4	5
Class	−	−	+	+	+	+	+

1. (3 points) Draw the decision boundary of a linear support vector machine on this data and identify the support vectors.

2. (3 points) Give the solution parameters w and b where the linear form is $wx + b$.

3. (2 points) Suppose we have another instance $(x = -5, Class = +)$. What kernel will you use to classify the training data perfectly.

4. (2 points) Calculate the leave one out cross validation error for the data set.

PROBLEM 6: Complexity Analysis (10 points)

Provide pseudo-code for an extension of the standard decision tree algorithm that does a k-move lookahead. We discussed a 1-move lookahead algorithm in class. Assume that you have N instances in the training set and each instance is defined by M binary attributes. Provide a computational analysis of the time-complexity of the algorithm with a 2-move lookahead. Hypothesize whether this algorithm will perform better or worse than the standard decision tree and give a qualitative argument why.