QUALIFIER: SPRING 2012
MACHINE LEARNING
CS 6375

The exam is closed book. You are allowed four pages of double sided cheat sheets. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use a separate page and attach it. Also, attach your cheat sheets with your exam.

TIME: 2 hours and 30 minutes

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

- TOTAL: _____

1 SHORT ANSWERS [10 points]

1. (2 points) For linearly separable data, can a small slack penalty hurt the training accuracy when using a linear SVM (no kernel)? If so, explain how. If not, why not?

2. (2 points) The Leave-one-out cross validation error of 1-nearest neighbor classifier is always zero. True or False. Explain.

3. (3 points) In boosting, would you stop the iteration if the following happens? Justify your answer.

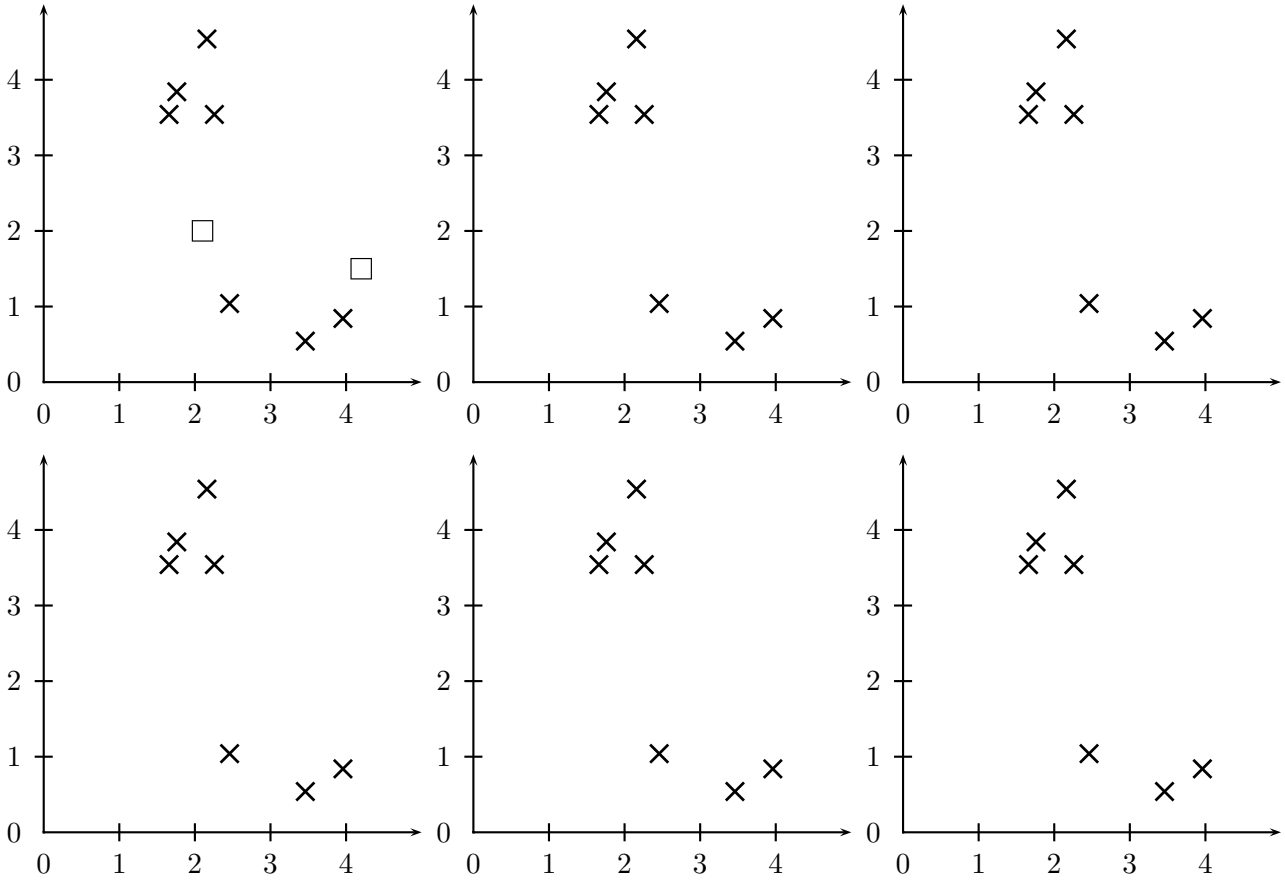
   (a) (1 point) The error rate of the combined classifier on the original training data is 0.

   (b) (2 points) The error rate of the current weak classifier on the weighted training data is 0.

4. (3 points) Gaussian Naive Bayes is a linear classifier. True or False. Explain.

2 CLUSTERING and THE EM ALGORITHM [20 points]

1. (5 points) Starting with two cluster centers indicated by squares, perform k-means clustering on the following data points (denoted by $\times$). In each panel, indicate the data assignment and in the next panel show the new cluster means. Stop when converged or after 6 steps whichever comes first.



2. (5 points) Draw the agglomerative clustering tree for the dataset given above. You must use single link clustering.

3. (5 points) Consider the dataset given below. All variables are Boolean. "?" denotes a missing value.

A	B	C	D
0	1	1	?
1	0	0	?
0	1	1	?
1	1	0	?
0	0	1	?
1	1	1	?

Assume that D is a class variable and you are learning a Naive Bayes model. Starting with probabilities that are initialized uniformly (i.e., all probabilities are initialized to 0.5), calculate the parameters of this Naive Bayes model using the EM algorithm. Stop at convergence or after 3 iterations, whichever is earlier. Does the EM algorithm converge and after how many iterations?

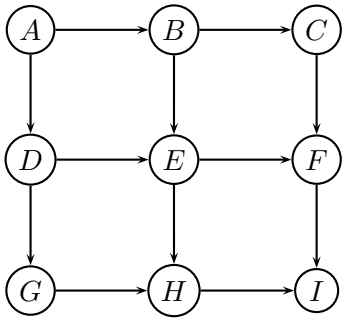
4. (5 points) Let us generalize our experience with such datasets, Naive Bayes and EM with uniform initialization. Assume that you have complete data with no missing values over all the input attributes but the class variable is always missing. Assume that you will learn the parameters of a Naive Bayes model using the EM algorithm with uniform initialization. Answer the following questions based on these assumptions.

   - At convergence, what will be the parameters of the Naive Bayes model?
   - After how many iterations will EM converge?

3 BAYESIAN NETWORKS and HIDDEN MARKOV MODELS [15 points]

1. (2 points) Suppose we are running the variable elimination algorithm for computing the probability of evidence. We find that after eliminating the first variable, we get a function ϕ that has all zeros, namely $\phi(\mathbf{x}) = 0$ for all assignments $\mathbf{X} = \mathbf{x}$ where $\mathbf{X}$ is the scope of ϕ. In this case, what can we say about the probability of evidence? Explain your answer.

2. (2 points) Consider the Bayesian network given below. Is A conditionally independent of F given $\{B, E\}$? Explain your answer.



3. (5 points) Recall that a Bayesian network is a generative model. In other words, it can be used to generate data. Provide an algorithm for generating data from a Bayesian network. What is its time complexity? Assume that the domain size of each variable is constant.

   Hint 1: To generate data from a Naive Bayes model, we sample the class variable first. Then given the sampled value of the class variable, say $C = j$, we sample a value for each attribute X_i from the conditional distribution $P(X_i|C = j)$.

   Hint 2: Before sampling a variable, we have to sample all of its parents.

4. (6 points) Recall that a HMM makes a 1-Markov assumption, i.e., the state variable X_t is conditionally independent of $X_{1:t-2}$ given X_{t-1}. Consider a HMM that makes a 2-Markov assumption instead, i.e., the state variable X_t is conditionally independent of $X_{1:t-3}$ given $\{X_{t-1}, X_{t-2}\}$. Describe in your own words how the filtering algorithm for this 2-Markov HMM will be different from the filtering algorithm for HMMs (you don't have to provide a pseudo code). Also, compare the time and space complexity of the 2-Markov HMM filtering algorithm with 1-Markov HMM filtering algorithm. What will be the computational complexity of filtering if we make a k-Markov assumption instead, where $k > 2$?

4 CLASSIFICATION [45 points]

4.1 PART A

1. (3 points) Circle the correct answer from the two choices in bold:

   - A generative model usually has **smaller larger** bias than a discriminative model.
   - Generative training (e.g., Naive Bayes) often works better than discriminative training (e.g., Logistic Regression) when we have **limited large** amount of data.
   - A polynomial-sized decision tree **can cannot** always be converted to a polynomial sized DNF. (A DNF is a disjunction of conjunctive formulas. For example, $(A \wedge B) \vee (\neg A \wedge C)$ is in DNF but $(A \vee B) \wedge (\neg A \vee C)$ is not in DNF).

2. (5 points) Draw a neural network that represents the following function.

$$\phi(x_1, x_2) = \begin{cases} -1 & \text{if } x_1 + x_2 = 0 \\ +1 & \text{otherwise} \end{cases}$$

   Here x_1 and x_2 are bi-valued discrete variables that take values from the domain $\{-1, +1\}$.

   Each hidden and output unit that you use must be sign() units. Recall that given inputs $(x_0, \ldots, x_n)$ and weights $(w_0, \ldots, w_n)$, each sign() unit will output a $+1$ if $\sum_{i=0}^{n} w_i x_i \geq 0$ and -1 otherwise. x_0 is the bias input which always equals 1.

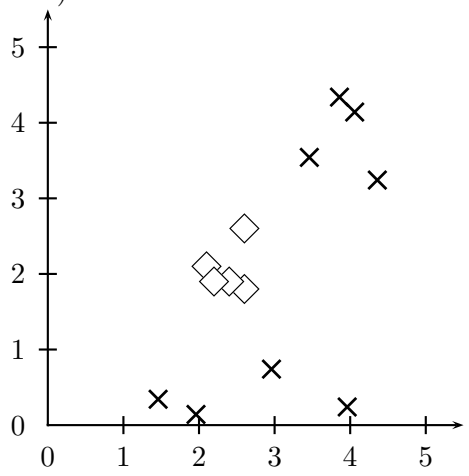
3. (2 points) Assume that all attributes in the data are Boolean and we have limited amount of training data but large amount of test data. Is the following statement True or False? Explain.

   Neural networks with $k > 2$ hidden layers will always be better in terms of accuracy on the unseen (test) examples than random guessing.

4.2 PART B

For each of the following datasets, write alongside each classifier whether it will have zero training error on the dataset. Also, explain why in one sentence or by drawing a decision boundary. No credit if the explanation is incorrect.

1. (3 points)

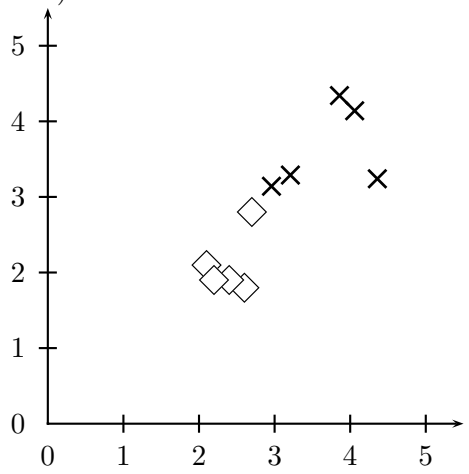


Logistic Regression:

3-nearest neighbors:

SVMs with a quadratic kernel:

2. (3 points)

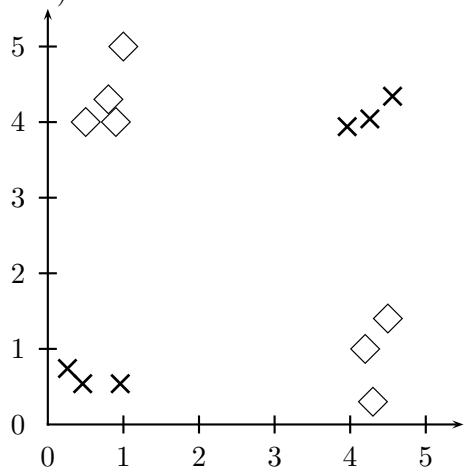


Logistic Regression:

3-nearest neighbors:

SVMs with a quadratic kernel:

3. (3 points)



Logistic Regression:

3-nearest neighbors:

SVMs with a quadratic kernel:

4.3 PART C

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 3^{rd} degree error (loss function) given by:

$$Loss = \frac{1}{m} \sum_{i=1}^{m} (y_i - w_1 x_i - w_0)^3$$

4. (3 points) Calculate the gradient with respect to the parameters w_1 and w_0. Hint: $\frac{dx^k}{dx} = kx^{k-1}$.

5. (3 points) Write down pseudo-code for online gradient descent for this problem.

6. (2 points) Consider a d-bit parity problem in which the input is a d-dimensional Boolean vector $\mathbf{x} = (x_1, \ldots, x_d)$ and the output is 1 or $true$ if the number of 1's in $\mathbf{x}$ is odd and 0 or $false$ otherwise. Is the problem linearly separable for $d > 1$? Explain.

7. (4 points) Estimate the number of nodes that a decision tree will require to model the d-bit parity function perfectly. Assume that your training data contains all possible, 2^d combinations and each split is on a value of a SINGLE Boolean variable. (**Hint**: Construct trees for $d = 3$ and $d = 4$ and compare the number of nodes in them to the maximum number of nodes that a decision tree can have).

8. (2 points) Consider a non-uniform prior which assigns positive probability mass to each possible hypothesis. As the number of data points grows to infinity, the MLE estimate of a parameter approaches the MAP estimate to arbitrary precision. True of False. Explain.

9. (2 points) There is no training data set for which a decision tree learner and logistic regression will output the same decision boundary. True or False. Explain.

5 COMPUTATIONAL LEARNING THEORY [10 points]

Useful formulas for this section:

$$m \geq \frac{1}{\epsilon}\left(\ln(1/\delta) + \ln(|H|)\right)$$

$$m \geq \frac{1}{2\epsilon^2}\left(\ln(1/\delta) + \ln(|H|)\right)$$

$$m \geq \frac{1}{\epsilon}\left(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon)\right)$$

1. (5 points) Consider the class C of concepts of the form $(a \leq x_1 \leq b) \wedge (c \leq x_2 \leq d)$. Note that each concept in this class corresponds to a rectangle in 2-dimensions. Let a, b be integers in the range $[0, 199]$ and c, d be integers in the range $[0, 99]$. Give an upper bound on the number of training examples sufficient to assure that for any target concept $c \in C$, any **consistent learner** using $H = C$ will, with probability 0.99, output a hypothesis with error at most 0.05.

2. (5 points) Consider the class C of concepts of the form $(a \leq x_1 \leq b) \wedge (c \leq x_2 \leq d) \wedge (e \leq x_3 \leq f)$. Note that each concept in this class corresponds to a hyper-rectangle in 3-d. Now suppose that a, b, c, d, e, f take on real values instead of integers. Give an upper bound on the number of training examples sufficient to assure that for any target concept $c \in C$, a learner will, with probability 0.95, output a hypothesis with error at most 0.01.