Practice EXAM: SPRING 2012
CS 6375
INSTRUCTOR: VIBHAV GOGATE

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, 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: ______________
- Problem 7: ______________

- TOTAL: ______________

1 SHORT ANSWERS

1. (2 points) Describe two ways in which you can use a Boolean classifier to perform multi-category (or multi-class) classification. What are the pros and cons of each.

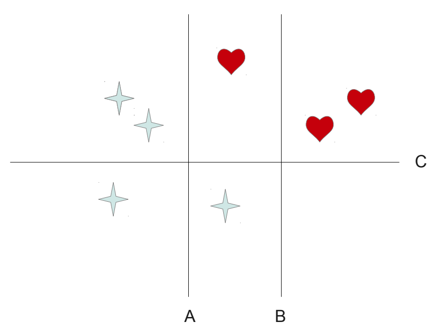
2. (2 points) The VC-dimension of k-nearest neighbor classifier is k. True or False. Explain.

3. (2 points) k-means clustering is a special case of hard EM. True of False. Explain.

4. (2 points) You are given a primal graph in which the maximum degree of a node is 50. You conclude that variable elimination is infeasible on this graph. Is your conclusion correct? Explain.

5. (2 points) We saw in class that Logistic regression (LR) and Gaussian Naive Bayes (GNB) have the same functional form. Therefore, the bias of LR equals that of GNB. True or False. Explain.

6. (2 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 LEARNING THEORY

1. (4 points) Assume that the data is 2-dimensional and each attribute is continuous (i.e., its domain is the set of reals $\mathbb{R}$). Let the hypothesis space H be a set of axis parallel SQUARES in 2-dimensions. What is the VC dimension of H? Explain.

2. (4 points) H is a finite hypothesis class, i.e. $|H| < \infty$. Show that the VC dimension of H is upper bounded by $log_2|H|$.

3. (2 points) You develop two classifiers, classifier "A" and classifier "B", and test them on approximately hundred large datasets available on the UCI machine learning repository. You find that classifier "A" has better accuracy than classifier "B" on all of them. Therefore, you conclude that no matter what dataset you use, classifier "A" will always be better than "B." True or False. Explain.

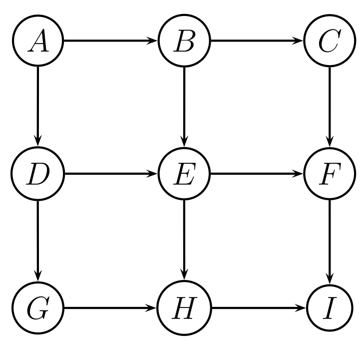
3 BAYESIAN NETWORKS

1. (2 points) Suppose we know that A is conditionally independent of B given C. Which of the following is sufficient to compute $P(A|B,C)$. Circle all those that apply.

   - $P(C)$, $P(C|A)$ and $P(A)$
   - $P(A)$, $P(A,C)$
   - $P(A|C)$

2. (2 points) Given a Bayesian network B, there exists a distribution Pr such that Pr is a perfect-map of B. True or False. Explain.

3. (4 points) Consider the Bayesian network given below:



   Show the steps in the variable elimination algorithm for computing $P(I = i)$ along the ordering (A, B, C, D, E, F, G, H). Can you say something about the treewidth of the primal graph.

4. (6 points) Suppose that you will learn the parameters of the grid Bayesian network given in the previous question using the EM algorithm.

   (a) (3 points) Assume that you are given a partially observed dataset in which variable "I" is always missing while the remaining variables are always completely observed. Let n be the number of examples in the dataset. What is the time and space complexity of the EM algorithm for this case.

   (b) (3 points) Assume that you are given a partially observed dataset in which variables $\{A, B, C, D, E, F\}$ are always missing while the remaining variables are always completely observed (i.e., you have complete data on G, H and I). Let n be the number of examples in the dataset. What is the time and space complexity of the EM algorithm for this case.

4 CLASSIFICATION ALGORITHMS

Consider the following dataset over 4 attributes. Ds values are in the range [0,100]. For the other three attributes, all of their possible values appear in this dataset.

	A	B	C	D	Class
Ex1	T	X	F	75	true
Ex2	F	Y	T	20	false
Ex3	T	Z	T	10	true
Ex4	F	Y	T	35	false
Ex5	F	X	T	90	false
Ex6	T	Z	F	50	false

1. (4 points) How much information about the Class variable is gained by knowing the value of feature B ?

2. (4 points) What are the three nearest-neighbors to Example 6? Explain. If this example was in the test set instead of the training set, would k-NN predict it correctly (using k=3)?

5 CLASSIFICATION ALGORITHMS CONTINUED

1. (2 points) Which of the following classifiers will have zero training error on the XOR function $(X_1 \wedge \neg X_2) \vee (\neg X_1 \wedge X_2)$. Circle or mark all those that apply.
   - 3-nearest neighbor
   - Logistic Regression
   - SVMs with quadratic kernel
   - Neural network with one hidden node.

2. (4 points) Can you use a perceptron to represent the following Boolean function $Y = (X_1 \wedge X_2) \vee (X_1 \wedge \neg X_2)$ perfectly, where X_1, X_2 and X_3 are Boolean variables. If so, draw the perceptron. If not, explain why not.

3. (2 points) A neural network with 3 hidden nodes has smaller bias than a perceptron. True or False. Explain.

6 SUPPORT VECTOR MACHINES

Suppose you are given 6 one-dimensional points: 3 with negative labels at -1, 0, and 1 and 3 with positive labels at -2, 2, and 3.

1. (1 point) Draw the points on a 1-D line.

2. (2 points) What kernel i.e., a function $f(x)$ will you use so that the transformed points are linearly separable.

3. (2 points) Draw the transformed points on a line and show the decision boundary and support vectors.

4. (2 points) The linear separator will be of the form $w_0 + w_1 f(x)$. Compute the value of w_0 and w_1.

5. (2 points) Now suppose we transform the 6 points to the feature space $(x, f(x))$, where $f(x)$ is your feature transformation from part 2. In other words, you now have 6 2-dimensional points. Draw the six points on a two-dimensional plane, along with the decision boundary for hard-margin linear SVM. Finally, indicate which points are support vectors.

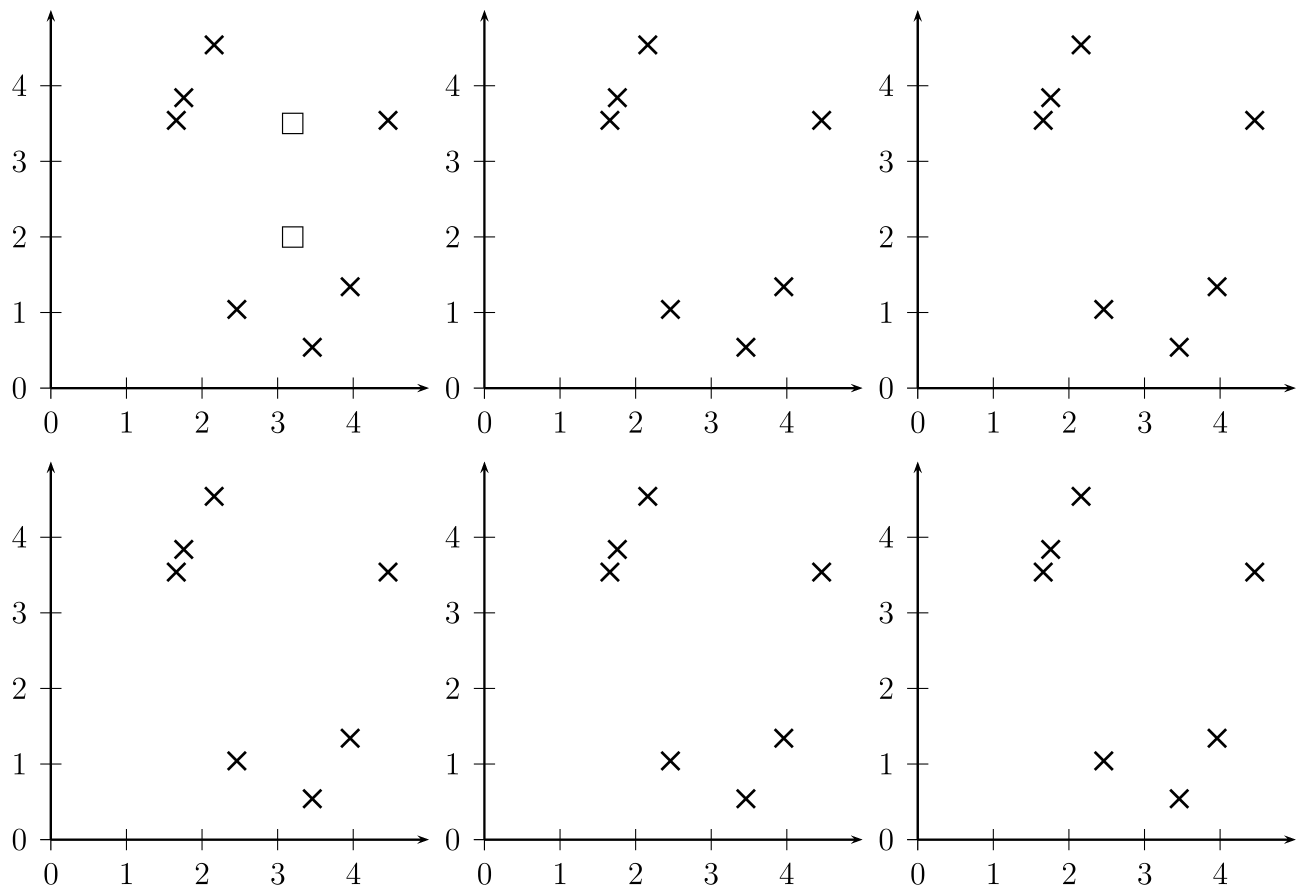
6. (2 points) Your decision boundary from part 5 has the form $w_0 + w_1 x + w_2 f(x)$. Give the values of w_0 , w_1 , w_2.

7. (2 points) The feature mapping $x \to (x, f(x))$ in parts 5 and 6 is associated with a kernel $K(x, x')$ where x, x' are points in the original one-dimensional feature space. Write down this kernel.

8. (2 points) What is the VC dimension of the linear SVM in the feature space $(x, f(x))$?

7 CLUSTERING

1. (3 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. (2 points) Write the formula for the cost function optimized by k-means. Explain the notation used clearly.

3. (2 points) Given an advantage of agglomerative clustering over k-means and an advantage of k-means over agglomerative clustering.