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.

Solution:

1. Classify $c_i, \neg c_i$
2. Classify Pairwise c_i, c_j

(1) is more scalable because it has smaller time complexity. However, the region where the classification is undefined is usually large.

(2) is less scalable because of the quadratic complexity. However, the region where the classification is undefined is usually small.

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

Solution: False. VC-dimension is ∞.

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

Solution: True. Hard EM E-step is equivalent to the k-means step that assigns points to clusters and the M-step is equivalent to the step that computes a new cluster center.

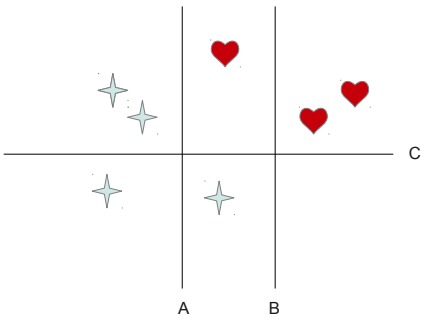
4. (2 points) You are given a primal graph in which the maximum degree of a node is 50. You conclude that variable elimination is infeasible on this graph. Is your conclusion correct? Explain.

Solution: False or can't say. Treewidth is lower bounded by the minimum degree and not by the maximum degree.

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.

Solution: False. GNB is a generative classifier and therefore it has larger bias than LR which is a discriminative classifier.

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?



Solution: It will choose B because the only example mis-classified by A is correctly classified by B (the misclassified examples are assigned higher weight in the next iteration).

In the first iteration $\epsilon = 1/7$

and $\alpha = \ln\{\frac{\epsilon}{1-\epsilon}\} = \ln(6)$.

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.

Solution: A square can shatter the 3 points that are the vertices of an equilateral triangle. Therefore, its VC-dimension is ≥ 3.

Consider the bounding square (the smallest square so that all points lie within or on the square) for any 4 points. By moving the square along the x or y axis, the label at an individual point can be changed. We will assume that the 4 points form a convex shape – if not, at least one point lies in the convex hull of the 4 points and a labeling where the interior point is labeled -1 and the external ones are labeled +1 cannot be achieved. The tightest bounding square must touch at least 3 points out of 4 (if not, we can make it smaller until this happens). Since each diagonal of the quadrilateral has 2 vertices, at least one of the two diagonals must have both points on the sides of the square. It can be shown then, for instance, that if both the points on that diagonals are labeled the same, then of the 4 possible labelings for the remaining 2 points, at least one of cannot be attained.

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

Solution: Consider a set of size $\log_2 |H| + 1$. Since each point can belong to one of two classes, the set can be labeled in $2^{(\log_2 |H|+1)} = 2|H|$ ways.

For a given set of points, a particular hypothesis provides a single labeling. Therefore $|H|$ can at most only provide $|H|$ labelings for the chosen set. So it cannot shatter a set of size $\log_2 |H| + 1$.

Therefore its VC dimension is bounded above by $\log_2 |H| + 1$.

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.

Solution: No. From the no free lunch theorem.

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.

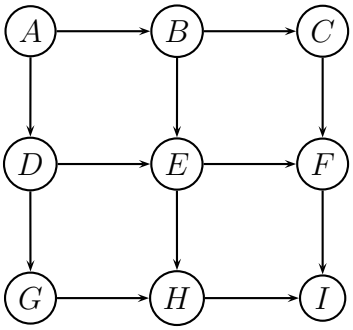
   - $P(C)$, $P(C|A)$ and $P(A)$
   - $P(A)$, $P(A,C)$
   - $P(A|C)$

   Solution: Yes: (1) and (3).

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.

   Solution: True.

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.

   Solution: The treewidth along the ordering is 3. The steps involved in VE are given below. I appears in only one CPT $P(I|F,H)$. Let the evidence instantiated CPT for it be $\psi(F,H)$.

   - Elim A: $\sum_A P(A)P(B|A)P(D|A)$. This yields a function $\phi(B,D)$
   - Elim B: $\sum_B P(E|B,D)P(C|B)\phi(B,D)$. This yields a function $\phi(C,D,E)$.
   - Elim C: $\sum_C P(F|C,E)\phi(C,D,E)$. This yields a function $\phi(D,E,F)$

- Elim D: $\sum_D P(G|D)\phi(D,E,F)$. This yields a function $\phi(E,F,G)$
- Elim E: $\sum_E P(H|E,G)\phi(E,F,G)$. This yields a function $\phi(F,G,H)$
- Elim F: $\sum_F \psi(F,H)\phi(F,G,H)$. This yields a function $\phi(G,H)$
- Elim G: $\sum_G \phi(G,H)$. This yields a function $\phi(H)$
- Elim H: $\sum_H \phi(H)$. This equals probability of evidence.

The maximum scope size of any new functions created is 3 and therefore the width of the ordering is 3. Therefore, the treewidth is at least 3. The time complexity of VE is $O(8\exp(4))$ and the space complexity is $O(8\exp(3))$.

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.

   Solution: As discussed in class, you can either update the sufficient statistics using the variable elimination algorithm for each example or complete each example. The complexity is the minimum of the two.

   The complexity of completing the dataset is $O(2n \times 9)$ because there are two possible completions for each instance and there are nine functions which we have to multiply to compute the probability or weight for each example. Since the maximum CPT size is $O(\exp(3))$, we will need $O(9\exp(3))$ to normalize the CPTs at the end. Thus the overall time complexity is $O(18n + 9\exp(3))$.

   (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 ?

Solution:

$$Gain(B) = H(Data) - \sum_{i=1}^{3} \frac{\#Data}{\#(Data|B=i)} H(Data|B=i)$$

$$H(Data) = H(2positive, 4negative) = -1/3 \log_2(1/3) - 2/3 \log_2(2/3)$$

and so on. Gain is 0.2517.

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)?

Solution: No. The two nearest examples are Ex1 and Ex3 (the third does not matter once we know the class of two).

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.

   Solution: Yes: SVMs with quadratic kernel.

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.

   Solution: Yes, draw the function and you can see that it is linearly separable.

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

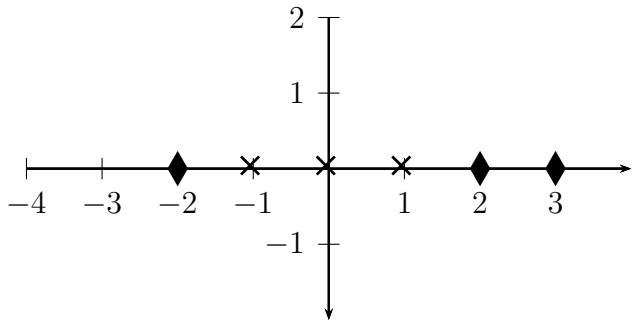
   Solution: True. As we saw in class, in a parametric model, smaller number of parameters imply smaller bias.

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.

Solution:

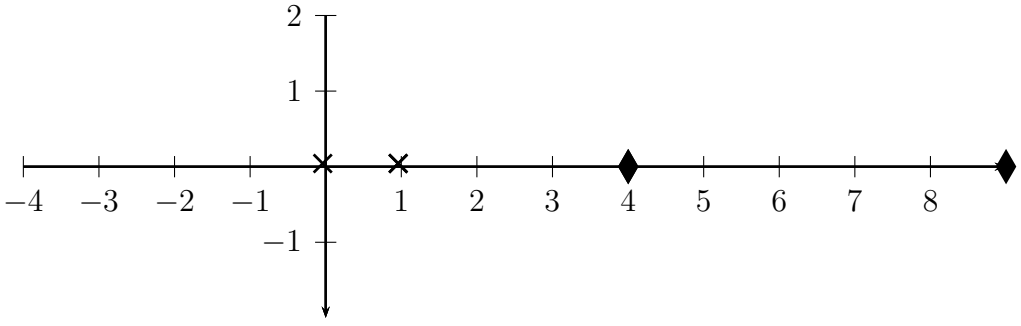


2. (2 points) What kernel i.e., a function $f(x)$ will you use so that the transformed points are linearly separable.

Solution: $f(x) = x^2$ will do.

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

Solution:

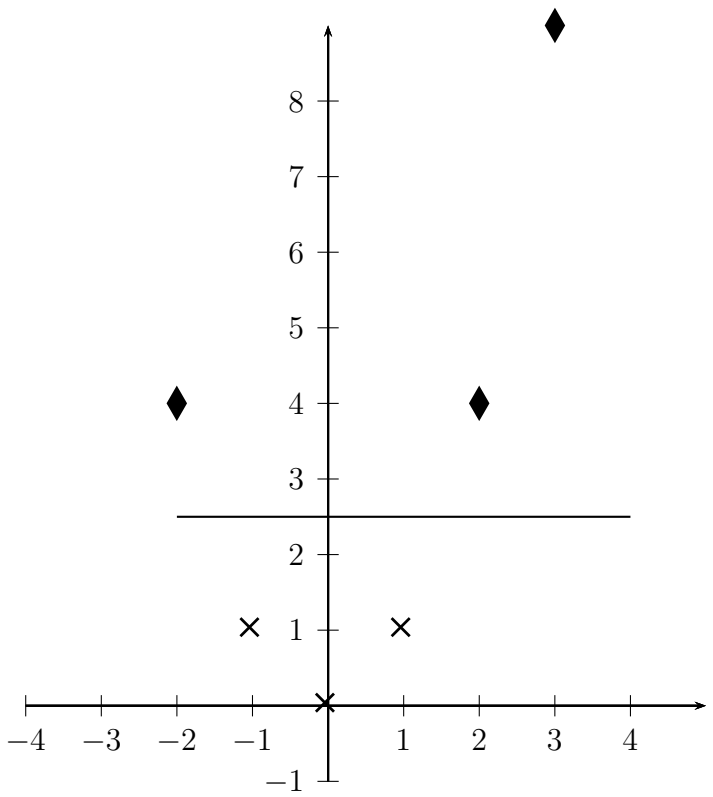


4. (2 points) The linear separator will be of the form $w_0 + w_1 f(x)$. Compute the value of w_0 and w_1.

Solution: $w_0 = -5/3$ and $w_1 = 2/3$

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.

Solution:



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.

Solution: Answer: $w_0 = -5/3, w_1 = 0, w_2 = 2/3$

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.

Solution: Kernel is

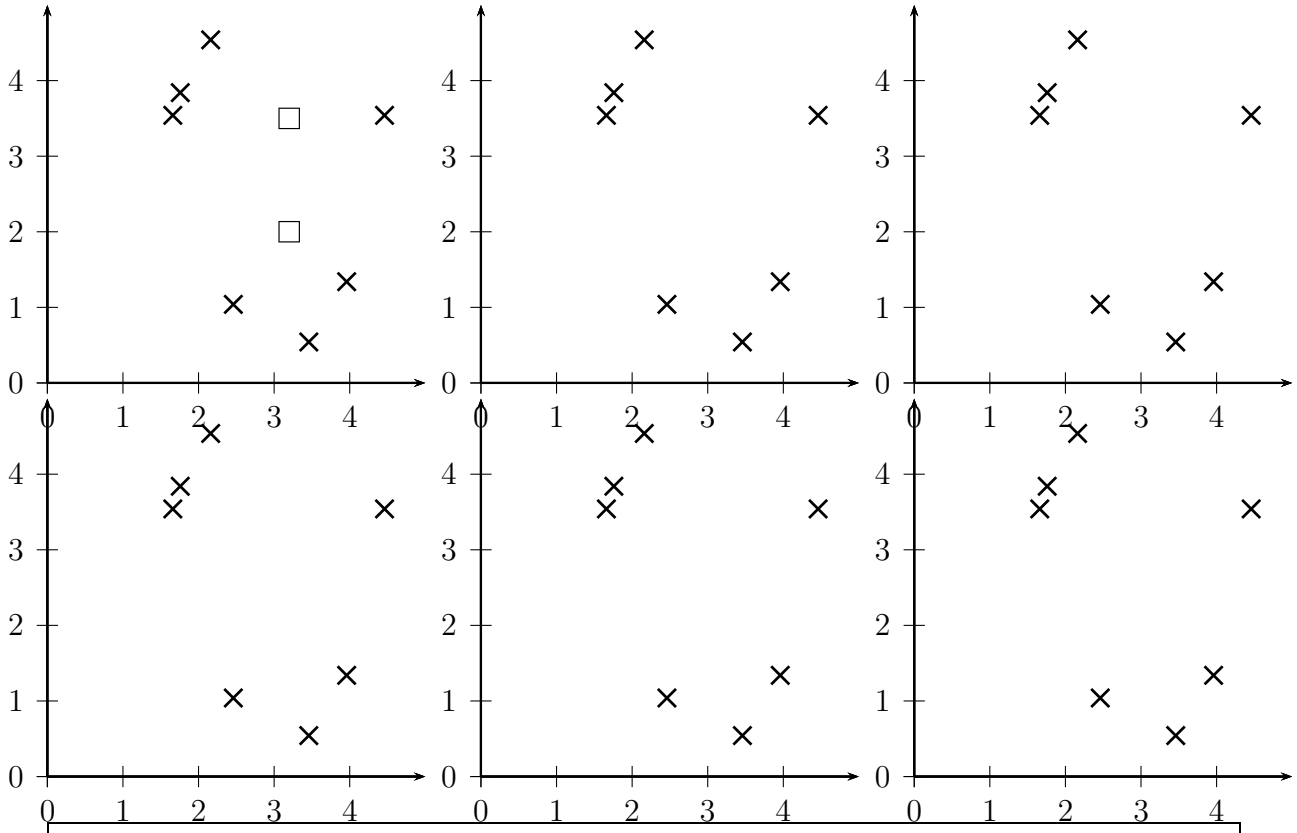
$$\begin{aligned} K(x_1, x_2) &= (x_1, x_2^2) \text{ dot product } (x_2, x_2^2) && (1) \\ &= x_1 x_2 + x_1^2 x_2^2 && (2) \end{aligned}$$

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

Solution: The given SVM is a linear classifier in 2-dimensions. VC dimension of a line is 3.

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.



Solution: K-means will converge in two steps.

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

Solution: The cost function optimized is:

$$\phi(\{x_i\}, \{a_i\}, \{c_k\}) = \sum_i dist(x_i, c_{a_i})$$

where $\{x_i\}$ is the set of points, $\{a_i\}$ are the set of assignments and $\{c_k\}$ are the k cluster means or cluster centers.

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

Solution: k-means: Often fairer, clear objective function and interpretable as a mixture model with hard assignments.

agglomerative clustering: produces many clusterings and as a result no need to specify k. Can learn long skinny clusters that k-means cannot.