PRACTICE FINAL EXAM: SPRING 2012
CS 6375
INSTRUCTOR: VIBHAV GOGATE

May 14, 2012

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: _____

Solution:
EXAM SOLUTIONS

1 SHORT ANSWERS

1. (2 points) The training error of 1-nearest neighbor is always zero (assuming that the data is not noisy). True or False. Explain.

Solution: TRUE. The class of each example is the class of its nearest neighbor, which is itself.

2. (2 points) You are given a dataset with the following statistics: # examples = 500, #attributes=20, #classes=2. You are pretty sure that the data is linearly separable. You perform 5-fold cross validation on the dataset using linear support vector machines and decision trees (ID3) and obtain the following results:

Table 1

	Error Fold1	Error Fold2	Error Fold3	Error Fold4	Error Fold5	Average Error
Linear SVM	1/100	0/100	1/100	30/100	1/100	33/500
ID3	7/100	4/100	5/100	4/100	5/100	25/500

where

$$Error = \frac{\#\text{mis-classified examples}}{\#\text{test-examples}}$$

Your boss Mr. LightWeight looks at Table 1 and concludes that ID3 is the best performing scheme for datasets of this type. Can you support his conclusion?

Solution: I cannot support his conclusion (no wonder he is Mr. LightWeight). Linear SVMs are performing better than ID3 on all folds except the 4-th fold. I suspect that in the 4-th fold, support vectors or examples close to them are in the test set. As a result, linear SVMs have larger error.

3. (2 points) A Neural network with one hidden layer that contains only polynomial number of (hidden) nodes can represent any Boolean function. True or False. Explain your answer.

Solution: FALSE. Some functions, for example, the parity function requires exponential number of hidden nodes, if only one hidden layer is used.

4. (2 points) 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.

Solution: See midterm solutions.

5. (2 points) Assuming Boolean attributes, the depth of a decision tree classifier can never be larger than the number of training examples. True or False. Explain.

Solution: TRUE. Each node in the decision tree divides the examples into two non-empty sets. Since there are "n" training examples, the number of nodes from the leaf to the root (i.e., the height) is bounded by "n."

2 LEARNING THEORY

1. (4 points) Assume that the data is 4-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 rectangles in 4-dimensions. What is the VC dimension of H? Explain.

Solution: We saw in class that the VC-dimension of axis-parallel rectangles is 2d where d is the dimensionality of the data. Therefore, the VC-dimension of 4-d rectangles=$2 \times 4 = 8$.

2. (4 points) Assume that the data is 2-dimensional and each attribute can take an integer value from the range $[0, 199]$. Let the hypothesis space H be a set of axis parallel rectangles in 2-dimensions. Given an upper bound on the number of examples sufficient to assure that any consistent learner will, with probability 99%, output a hypothesis with error at most 0.05.

   Hint1: Given a region in a plane bounded by the points (0,0) and (n-1,n-1), the number of axis parallel rectangles with integer-valued boundaries is $(\frac{n(n+1)}{2})^2$.

   Hint2: You can use one of the following two formulas:

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

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

   Solution: The hint is wrong. The number of rectangles is $(n(n-1)/2)^2$. Here, $n = 200$ and therefore $|H| = ((200)(200+1)/2)^2 = (20100)^2$. Since the learner is consistent, we will use the second formula and substitute $|H| = 20100^2$, $\epsilon = 0.05$ and $\delta = 0.01$ to get a bound on m.

3. (2 points) Within the setting of PAC learning, it is impossible to assure with probability 1 that the concept will be learned perfectly (i.e., with true error = 0), regardless of how many training examples are provided. True or False. Explain.

   Solution: True. PAC learning assures that the error is less than ϵ. Therefore, ϵ cannot be zero.

3 BAYESIAN NETWORKS

1. (2 points) An empty graph is a trivial D-map. True or False. Explain.

Solution: True. A DAG G is a D-map of Pr iff $I_{\Pr} \subseteq I_G$. Since an empty graph represents all possible conditional independence statements about the variables, all CI statements in Pr must be included in I_G. Therefore, it is a trivial D-map.

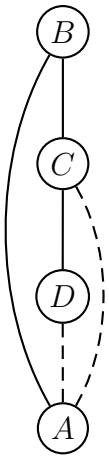
2. (2 points) What is the complexity of computing $P(E = e)$ using variable elimination in the following Bayesian network along the ordering (A, B, C, D). The edges in the Bayesian network are $A \rightarrow B$, $A \rightarrow C$, $B \rightarrow C$, $C \rightarrow D$, and $D \rightarrow E$.

Solution: The functions after instantiating the evidence variable are $P(B|A)$, $P(C|A)$, $P(C|B)$, $P(A)$, $P(D|C)$ and $\phi(D)$. The induced graph along the ordering is shown below.

The number of children of A, B, C and D are 2, 1, 1 and 0 respectively. Therefore, the width is 2. The complexity is $O(n \exp(w + 1))$ where n is the number of non-evidence variables and w is the width of the ordering. Therefore, the complexity is $O(4 \exp(3))$.

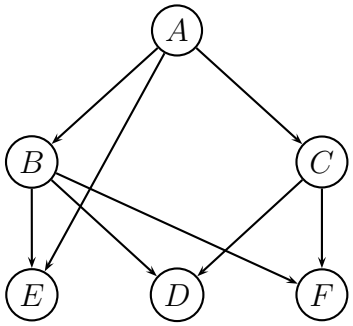
3. (2 points) What is the complexity of computing $P(E = e)$ using variable elimination in the following Bayesian network along the ordering (B, C, D, A). The edges in the Bayesian network are $A \rightarrow B$, $B \rightarrow C$, $C \rightarrow D$, and $D \rightarrow E$.

Solution: The functions after instantiating the evidence variable are $P(B|A)$, $P(C|B)$, $P(D|C)$, $P(A)$ and $\phi(D)$. The induced graph along the ordering is shown below.



The number of children of A, B, C and D are 2, 2, 1 and 0 respectively. Therefore, the width is 2. The complexity is $O(n \exp(w+1))$ where n is the number of non-evidence variables and w is the width of the ordering. Therefore, the complexity is $O(4 \exp(3))$.

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



Are the following statements true? Explain your answer.

- (1 point) A is conditionally independent of D given $\{B, C\}$.
- (1 point) E is (marginally) independent of F.
- (2 points) Which edges will you delete to make A marginally independent of C.

Solution:

- True. The only edges in new graph after running the d-sep algorithm given in class notes are B-A and A-C. Since A and D are disconnected, they are conditionally independent given {B,C}.
- False. The graph obtained by running the d-sep algorithm is same as the original graph except that it is undirected and it does not include the node D.
- Remove the edge $A \rightarrow C$. Now A and C are marginally independent.

4 DECISION TREES

The following is data from last eight Dallas Mavericks game:

Game#	Opponent	Point Guard	Fouls	Result
1	Weak	Strong	No	Win
2	Strong	Strong	Many	Loss
3	Strong	Weak	Many	Loss
4	Weak	Weak	Many	Loss
5	Strong	Weak	No	Win
6	Weak	Weak	Few	Win
7	Strong	Weak	Few	Loss
8	Strong	Strong	Few	Win

1. (4 points) What is the entropy of the data set? (Result is the class attribute)

Solution: 4 wins and 4 losses. This implies that the entropy is 1.

2. (3 points) What is the information gain if you split the dataset based on the attribute Fouls?

Solution:

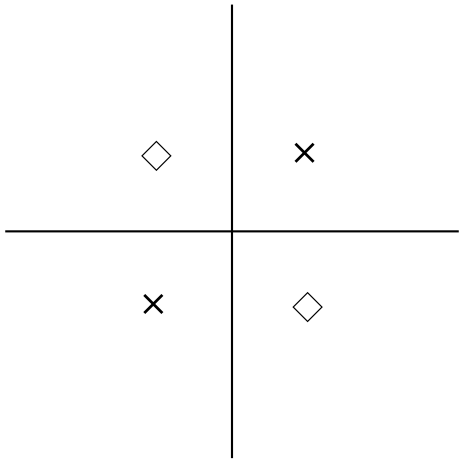
$$Gain(fouls) = 1 - (2/8)(0) - (3/8)(0) - 3/8[-(1/3)\log_2(1/3) - (2/3)\log_2(2/3)]$$

3. (3 points) We tell you that $Gain(S, Opponent) = Gain(S, PointGuard) = 0.05$. Based on your answer in (b) and this information, which attribute will you choose as the root node for the decision tree? Circle the appropriate option below.

- Opponent
- PointGuard
- Fouls

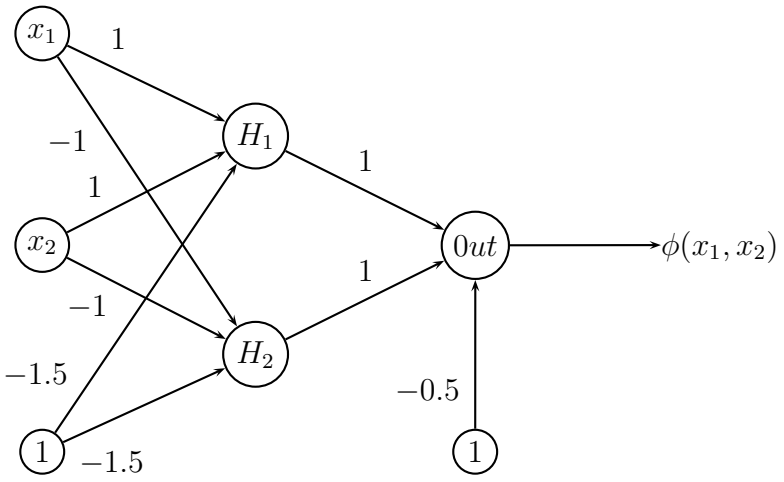
Solution: Gain of Fouls is greater than both opponent and PointGuard. Therefore, we will chose Fouls.

5 NEURAL NETWORKS and LOGISTIC REGRESSION

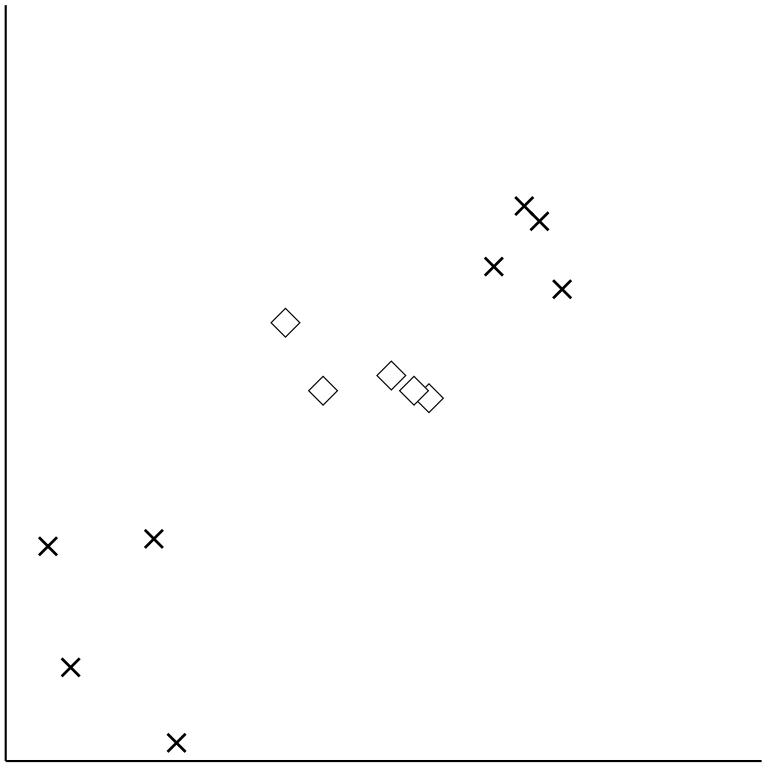


1. (5 points) Consider the data set given above. Assume that the co-ordinates of the points are (1,1), (1,-1), (-1,1) and (-1,-1). Draw a neural network that will have zero training error on this dataset. (Hint: you will need exactly one hidden layer and two hidden nodes).

Solution: There are many possible solutions to this problem. I describe one way below. Notice that the dataset is not linearly separable. Therefore, we will need at least two hidden units. Intuitively, each hidden unit will represent a line that classifies one of the squares (or crosses) correctly but mis-classifies the other. The output unit will resolve the disagreement between the two hidden units. I am assuming that the symbol $\times$ is positive and the other symbol implies negative class.



All hidden and output units are simple threshold units (aka sign units). Recall that each sign unit will output a $+1$ if $w_0x_0 + w_1x_1 + \ldots + w_nx_n > 0$ and -1 otherwise.

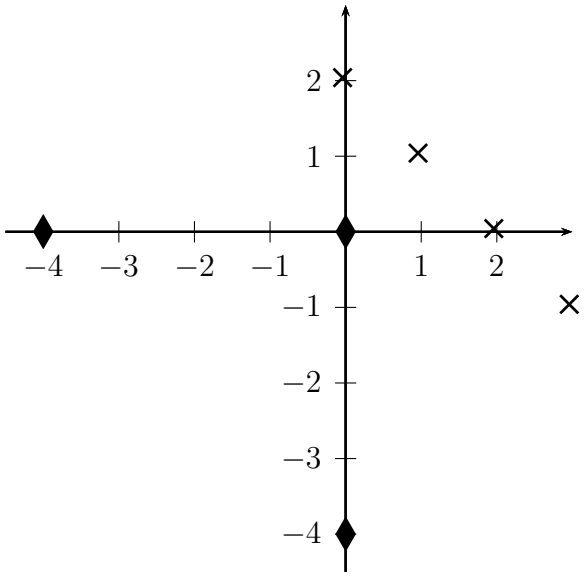


2. (5 points) Which of the following will classify the data perfectly and why? Draw the decision boundary.

   - Logistic Regression
   - Logistic Regression with a quadratic kernel.

Solution: Logistic Regression with a quadratic kernel. There will be a circle around the squares or a parabola depending on your which quadratic kernel you use.

6 SUPPORT VECTOR MACHINES



Consider the dataset given above. In this problem, we will use the data to learn a linear SVM of the form $f(x) = sign(w_1x_1 + w_2x_2 + w_3)$, with $||w|| = 1$.

1. (4 points) What is that value of w that the SVM algorithm will output on the given data?

Solution: $w_1 = w_2 = \frac{1}{\sqrt{3}}$ and $w_3 = -\frac{1}{\sqrt{3}}$.

2. (3 points) What is the training set error (expressed as the percentage training points misclassified)?

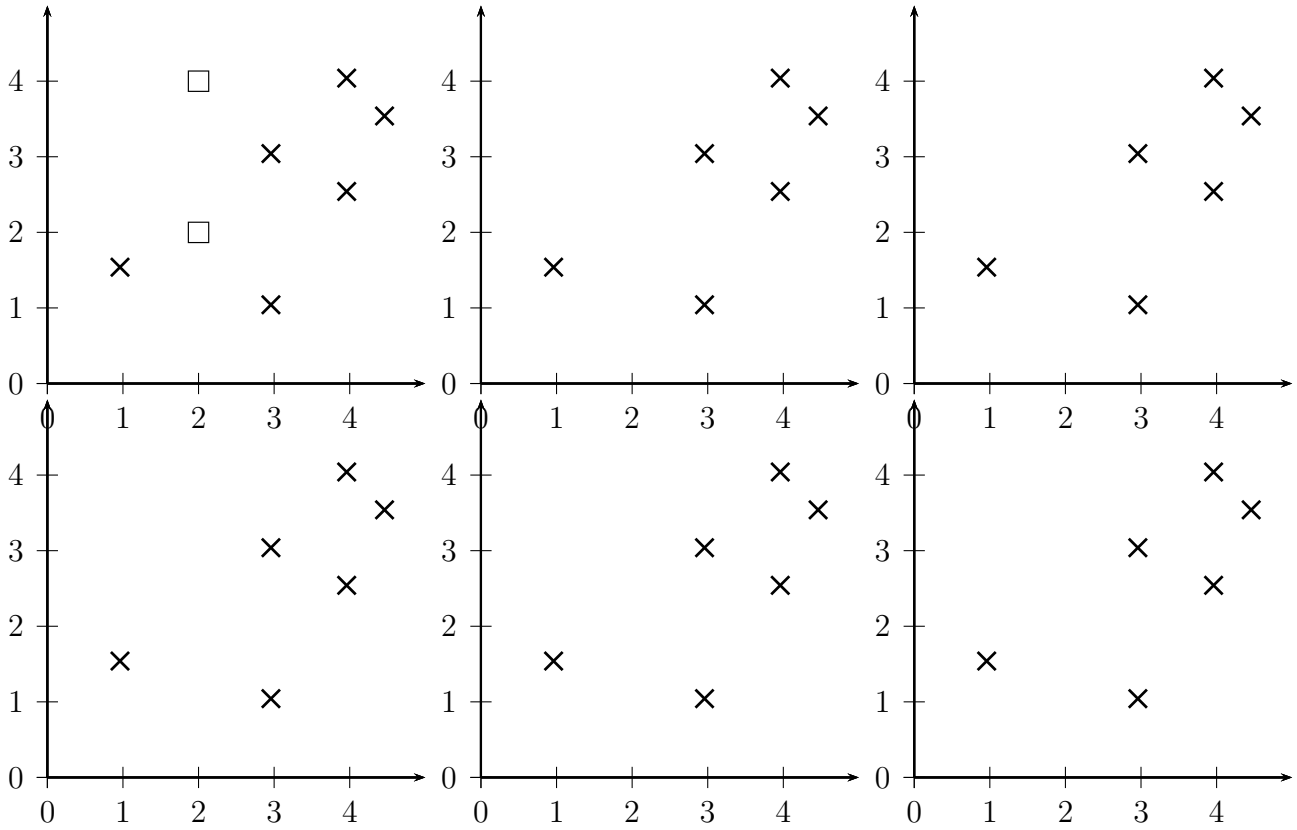
Solution: 0%. All classified correctly.

3. (3 points) What is the 7-fold cross validation error?

Solution: 1/7.

7 CLUSTERING

1. (5 points) Starting with two cluster centers indicated by squares, perform k-means clustering on the six 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.

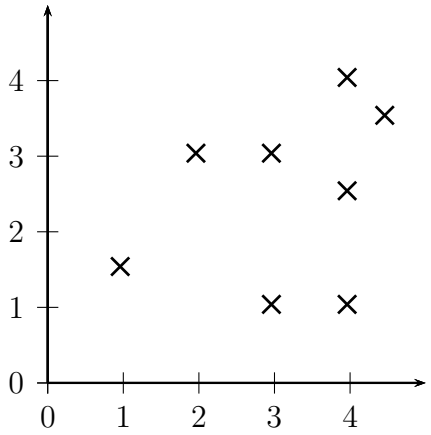


Solution: k-means will converge in two steps. In the first step, the points at the top will be assigned to the cluster center at the top and the points at the bottom will be assigned to the cluster center at the bottom. In the second step, the new cluster centers will be the mean of the points at the top and the points at the bottom respectively.

2. (2 points) Explain at a high level how K-means differs from the EM algorithm.

Solution: k-means yields hard clustering while EM yields soft clustering.

3. (3 points) Draw the agglomerative clustering tree for the following data set. You must use single link clustering.



Solution: Some judgement calls here. Many solutions possible.