FINAL EXAM: FALL 2013
CS 6375
INSTRUCTOR: VIBHAV GOGATE

You are allowed a two-page cheat sheet. You are also allowed to use a calculator. 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.

NAME _____

UTD-ID if known _____

- Problem 1: _____
- Problem 2: _____
- Problem 3: _____
- Problem 4: _____
- Problem 5: _____
- Problem 6: _____
- Problem 7: _____
- Problem 8: _____
- TOTAL: _____

1 Decision Trees [13 points]

You are given the following Dataset S:

X	Y	Z	Class
a	1	0	Yes
b	1	0	Yes
c	1	0	No
a	1	0	Yes
c	0	1	No
b	0	1	No
a	0	0	No
a	1	0	Yes

1. (3 points) What is the Entropy of the Dataset S?

Solution: The entropy of the data is 1.

2. (3 points) What is the Entropy of the attribute X?

Solution: Entropy[X] $= -1/2 \times \log(1/2) - 1/4 \times \log(1/4) - 1/4 \times \log(1/4) = 3/2$

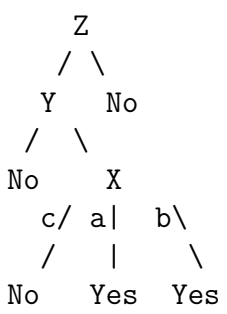
Data replicated here for convenience:

X	Y	Z	Class
a	1	0	Yes
b	1	0	Yes
c	1	0	No
a	1	0	Yes
c	0	1	No
b	0	1	No
a	0	0	No
a	1	0	Yes

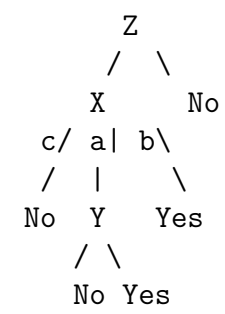
3. (3 points) You are given that: Gain(S,X)= 0.344; Gain(S,Y) = 0.2049; and Gain(S,Z) = 0.5171. Use it to choose the appropriate root node. Then, build the rest of the tree in a way that is consistent with the given data. Please note that beyond the root node, you don't need to use any specific algorithm; just make sure your tree is consistent with the data. Draw the tree below.

Solution:

First Tree



Second Tree



4. (4 points) For the above dataset S, does there exist a conjunctive feature such that if the feature is true then the class is Yes. Otherwise the class is No. Explain your answer (either find such a feature or prove that none exists)? Recall that a conjunctive feature looks like this: $(A = a) \wedge (B = b) \wedge \neg(C = c)$.

Solution: Feature is $(\neg Z \wedge Y \wedge \neg(X = c))$.

2 SHORT ANSWERS [20 points]

1. (3 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. (3 points) The VC-dimension is always less than the size of the hypothesis space. True or False. Explain your answer. No credit if you don't explain your answer.

Solution: True. $VC(H) = \log_2 |H|$

3. (3 points) EM algorithm is always guaranteed to converge to a local minima. True or False. Explain your answer. No credit if you don't explain your answer.

Solution: True. At each iteration, EM improves its cost function and is guaranteed to converge to a local optima.

SHORT ANSWERS CONTINUED [20 points]

4. (3 points) For every (Discrete) distribution, there exist a Bayesian network such that the Bayesian network is an I-map of the distribution. True or False. Explain your answer. No credit if you don't explain your answer.

Solution: True. Trivially, a complete Bayesian network is an I-map of the distribution.

5. (3 points) Gaussian Naive Bayes is a linear classifier. True or False. Explain your answer. No credit if you don't explain your answer.

Solution: Depends. only when the variances are assumed to be independent of the class.

SHORT ANSWERS CONTINUED [20 points]

6. (5 points) As a consultant to a local plant that manufactures usb drives you need to estimate the failure probability of their usb drives. In order to do that, a series of experiments is performed; in each experiment several usb drives are tried until a good one is found. The number of failures, k, is recorded.

   (a) (2 points) Given that p is the failure probability, what is the probability of k failures before a good usb drive is found?

   Solution:

   $$p^k(1-p)$$

   (b) (3 points) You have performed m independent experiments of this form, recording $k_1, k_2, \ldots, k_m$. Estimate the most likely value of p as a function of m and $k_1, k_2, \ldots, k_m$.

   Solution:

   $$\begin{aligned} LL &= \sum_{i=1}^{m} \ln(p_i^k * (1-p)) \\ &= \sum_{i=1}^{m} k_i \ln(p) + \ln(1-p) \\ &= m \ln(1-p) + \sum_{i=1}^{m} k_i \ln(p) \end{aligned}$$

   Now find the p that maximizes the likelihood (take derivative and set to 0:

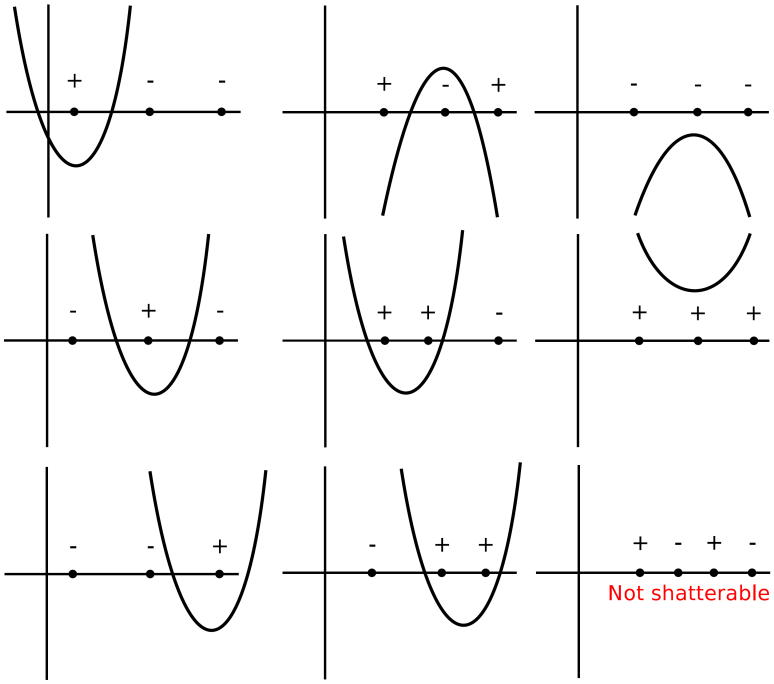
   $$\begin{aligned} \frac{dLL}{dp} &= -\frac{m}{1-p} + \sum_{i=1}^{m} \frac{k_i}{p} = 0 \\ p &= \frac{\sum_{i=1}^{m} k_i}{m + \sum_{i=1}^{m} k_i} \end{aligned}$$

3 LEARNING THEORY [10 points]

1. (8 points) Let us define a set of concepts $H = sgn(ax^2+bx+c)$ where a, b, and c are real numbers and $sgn(z)$ is 1 when z is positive, and 0 otherwise. What is the VC dimension of H? Prove your claim.

Solution:

The VC dimension is 3. To prove this, we need only show that there is one configuration such that all labelings are shatterable, and that no set of 4 points are shatterable. Note that we are only dealing with points on the x axis (although the VC dimension is still 3 in two dimensions). Any alternating labeling of four points will result in a configuration that is not shatterable because quadratic functions can change signs at most twice.



2. (2 points) A neural network is always going to perform better in terms of accuracy than random guessing. True or False. Explain your answer. No credit will be given if you don't explain your answer.

Solution: No. From the no free lunch theorem.

4 AdaBoosting [10 points]

For your convenience, the AdaBoost algorithm that we discussed in class is printed on the last page. Suppose you have two weak learners, l_0 and l_1, and a set of 1700 points and you start with weight $= 1/1700$ for each data point.

1. (3 points) You find that l_0 makes 100 mistakes while l_1 makes 400 mistakes. Choose which learner l_0 or l_1 will AdaBoost use in the first round. Justify your answer.

Solution: We will choose classifier l_1 because its error is smaller.

2. (3 points) Compute the weights α_0 and ϵ_0.

Solution:

$$\epsilon_0 = 100/1700 = 1/17$$

$$\alpha_0 = \ln\left\{\frac{1 - 1/17}{1/17}\right\} = \ln(16)$$

3. (4 points) Compute the weights for the next iteration. Precisely specify the weights on examples on which your chosen classifier made a mistake and the weight on examples on which the classifier was correct.

Solution: On the 100 points on which the classifier made a mistake, the weighting coefficients are:

$$w = (1/1700) * 16$$

On the points on which the classifier did not make a mistake, the weighting coefficients are

$$w = (1/1700)$$

Typically, we normalize the weights to yield a distribution over the points.

5 Perceptron [10 points]

In this problem, you will use the Perceptron learning algorithm to learn a separating line for the following six data points:

x_1	x_2	class
1	6	−
3	3	−
-3	2	+
3	-5	−
-4	-2	+
0	-5	+

Assume that we have two features x_1 and x_2 and we initialize the weights to $w_1 = w_2 = 0$. Let $f(x)$ be given by the following function

$$f(x) = sgn(w_1x_1 + w_2x_2 + w_0x_0)$$

where x_0 is the bias term.

1. (5 points) Assume that the bias term is always equal to 0. Apply the perceptron algorithm with a learning rate of 1 to the data in the order it is given in the table. Give the final weight vector you arrive at and state whether or not it is consistent with the data.

Solution:

Trivial: Skipped: Iterations.

At the end of second iteration, we get $(-12, -6)$, which classifies the data perfectly.

2. (5 points) Consider adding a new data point: $x_b = (0.5, 0)$. Assume that Perceptron will not be given the label of x_b until it has started training. Can the algorithm as presented in the previous part represent a hypothesis consistent with all the data (assuming that the bias term $x_0 = 0$)?

   - If so, give one $w = (w_1, w_2)$ that is consistent when x_b is positive and one w that is consistent when x_b is negative (you dont have to run the algorithm again).
   - If not, suggest a modification to the set-up in the previous part that enables the algorithm to represent a consistent hypothesis in either case.

Solution: The representation is not expressive enough. Replace x_0 by a positive number (e.g., 1). The data is still linearly separable.

6 Neural Networks [10 points]

1. (10 points) Assume ab and cd are two-bit binary numbers. efg is a 3-bit binary number obtained by adding ab and cd. Namely, $ab + cd = efg$. For example $ab = 11$, $cd = 11$ then $efg = 110$.

   Build a neural network with one hidden layer to implement e. The input nodes are a, b, c, d and the bias terms.

Solution: Many solutions possible here.

7 Support Vector Machines [20 points]

An SVM is trained with the following data:

x_1	x_2	class
-1	-1	-1
1	1	1
0	2	1

Let α_1, α_2 and α_3 be the Lagrangian multipliers associated with this data (α_i is the lagrangian associated with the i-th point).

1. (5 points) Using the polynomial kernel of degree 2, what (dual) optimization problem needs to be solved in terms of the lagrangian multipliers in order to determine their values? The polynomial kernel is given by $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)$ where $\mathbf{x}_i$ and $\mathbf{x}_j$ are vectors.

Solution: The Kernel matrix is given by

$$\begin{bmatrix} 9 & 1 & 1 \\ 1 & 9 & 9 \\ 1 & 9 & 25 \end{bmatrix}$$

Maximize:

$$\alpha_1 + \alpha_2 + \alpha_3 - \frac{1}{2}(9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 9\alpha_2^2 + 18\alpha_2\alpha_3 + 25\alpha_3^2)$$

subject to:

$$\alpha_1, \alpha_2, \alpha_3 \geq 0 \quad and \quad -\alpha_1 + \alpha_2 + \alpha_3 = 0$$

2. (5 points) Let us say that I solved the above optimization problem for you and found that $\alpha_1 = \alpha_2 = 1/8$ and $\alpha_3 = 0$. Moreover, $b = 0$. Can you tell me which of the three points are support vectors. Explain your answer. No credit given without correct explanation.

Solution: The support vector are points 1 and 2 because α_1 and α_2 are greater than zero.

3. (5 points) Assuming $\alpha_1 = \alpha_2 = 1/8$, $\alpha_3 = 0$,$b = 0$, how will the SVM classify the point $(x_1 = -1, x_2 = 0)$? Explain your answer. No credit given without correct explanation.

Solution: The dot product of the kernel with the test point is $(4, 0)$.

$$-\frac{1}{8}(4) + \frac{1}{8}(0) < 0$$

Therefore, the class is -1.

4. (5 points) Assuming $\alpha_1 = \alpha_2 = 1/8$, $\alpha_3 = 0$,$b = 0$, how will the SVM classify the point $(x_1 = 1, x_2 = 0)$? Explain your answer. No credit given without correct explanation.

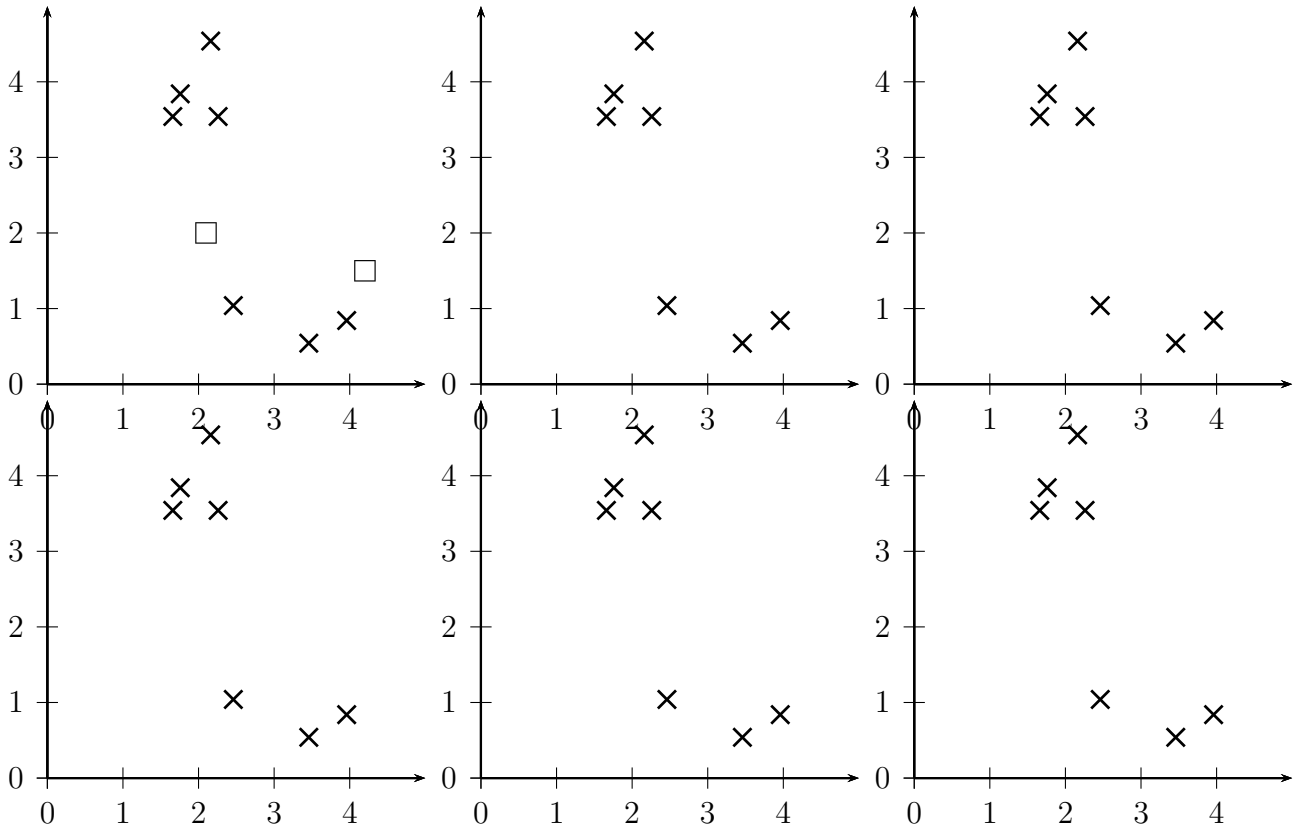
Solution: The dot product of the kernel with the test point is $(0, 4)$.

$$-\frac{1}{8}(0) + \frac{1}{8}(4) > 0$$

Therefore, the class is $+1$.

8 CLUSTERING [7 points]

1. (4 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. Use Euclidean distance as the distance measure.



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. (3 points) Do single link clustering and complete link clustering have the same computational complexity. Explain your answer. No credit will be given without a correct explanation.

Solution: Complete link clustering $O(n^2 \log(n))$ has higher complexity than single link clustering $O(n^2)$. The reason for this difference between single-link and complete-link is that distance defined as the distance of the two closest members is a local

property that is not affected by merging; distance defined as the diameter of a cluster is a non-local property that can change during merging.

AdaBoost

1. Initialize the data weighting coefficients $\{w_n\}$ by setting $w_n^{(1)} = 1/N$ for $n = 1, \ldots, N$.
2. For $m = 1, \ldots, M$:
   (a) Fit a classifier $y_m(\mathbf{x})$ to the training data by minimizing the weighted error function

$$J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n) \tag{14.15}$$

   where $I(y_m(\mathbf{x}_n) \neq t_n)$ is the indicator function and equals 1 when $y_m(\mathbf{x}_n) \neq t_n$ and 0 otherwise.
   (b) Evaluate the quantities

$$\epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(\mathbf{x}_n) \neq t_n)}{\sum_{n=1}^{N} w_n^{(m)}} \tag{14.16}$$

   and then use these to evaluate

$$\alpha_m = \ln \left\{ \frac{1 - \epsilon_m}{\epsilon_m} \right\}. \tag{14.17}$$

   (c) Update the data weighting coefficients

$$w_n^{(m+1)} = w_n^{(m)} \exp \{\alpha_m I(y_m(\mathbf{x}_n) \neq t_n)\} \tag{14.18}$$

3. Make predictions using the final model, which is given by

$$Y_M(\mathbf{x}) = \text{sign} \left(\sum_{m=1}^{M} \alpha_m y_m(\mathbf{x}) \right). \tag{14.19}$$