Final: CS 6375
Spring 2015

The exam is closed book (2 cheat sheets allowed). If you run out of room for an answer, use an additional sheet (available from the instructor) and staple it to your exam.

- **NAME** _____
- **UTD-ID if known** _____
- **Time:** 2 hours 40 minutes.

Question	Points	Score
Decision Trees	10	
Linear versus Non-linear functions	10	
Naive Bayes and Logisitic Regression	10	
Neural Networks	10	
VC-Dimension	10	
Bayesian networks	10	
AdaBoosting	10	
Learning Bayesian networks	10	
Clustering Techniques	10	
Support Vector Machines	10	
Total:	100	

Question 1: Decision Trees (10 points)

Consider the dataset given below. X_1 and X_2 are the attributes and Y is the class variable. 0 indicates negative class and 1 indicates positive class:

X_1	X_2	Y
0.5	2.5	0
0.5	5.5	1
1.5	4.5	0
2.5	1.5	1
2.5	4.5	0
4.5	1.5	0
4.5	4.5	1

(a) (6 points) Which feature has the highest information gain. Recall that since the attributes are continuous, you have to use features which compare the value of the attribute against a threshold. For example, $X_1 < 7$ is a feature.

Solution: The features are $X_1 < 1$, $X_1 < 2$, $X_1 < 3$, $X_2 < 2$, $X_2 < 3$ and $X_2 < 5$. Out of them $X_2 < 5$ has the highest information gain.

(b) (4 points) Draw a Decision Tree (need not be the best one) that has 100% accuracy on the dataset. Will your decision tree correctly classify the following test point: $X_1 = 0.5$, $X_2 = 7.5$, $Y = 1$. Explain your answer.

Solution: Many solutions possible here.

Question 2: Linear versus Non-linear functions (10 points)

Recall that a linear threshold function or a linear classifier is given by: If $(w_0 + \sum_i w_i x_i) > 0$ then the class is positive, otherwise it is negative. Here, $x_1, \ldots, x_n$ are the attribute values and $w_0, \ldots, w_n$ are the weights (w_0 is the bias term). Assume that 1 is true and 0 is false.

(a) (5 points) Consider a function over n Binary attributes where $n \geq 2$ and n is even that is defined as follows. If exactly $n/2$ attributes are false, then the class is **positive**, otherwise the class is **negative**. Can you always represent this function using a linear threshold function. If your answer is **YES**, then give a precise numerical setting of the weights. Otherwise, clearly explain, why this function cannot be represented using a linear threshold function. No credit will be given if the explanation is incorrect.

Solution: Counter example: For $n = 2$, it represents the XOR function. Therefore, not possible.

(b) (5 points) Consider the Boolean function given below. x_1, x_2 and x_3 are the attributes and y is the class variable.

x_1	x_2	x_3	y
0	0	0	1
0	0	1	-1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	-1
1	1	0	1
1	1	1	1

Can you represent the function using a linear threshold function. If your answer is **YES**, then give a precise numerical setting of the weights. Otherwise, clearly explain, why this function cannot be represented using a linear threshold function. No credit will be given if the explanation is incorrect.

Solution: Yes. The function can be written as: If $x_2 \vee \neg x_3$ then output is $+1$, else output is -1. We can easily represent it using a Perceptron with weights: $w_0 = 0.5$, $w_1 = 0$, $w_2 = 1$ and $w_3 = -1$.

Question 3: Naive Bayes and Logisitic Regression (10 points)

(a) (10 points) Under each statement given below, write **True** if the statement is true and **False** if it is not. You don't have to explain your answer.

1. Both Logistic Regression and Discrete Naive Bayes are linear classifiers.

   Solution: True.

2. Both Naive Bayes and Logistic Regression are generative classifiers.

   Solution: False. LR is discriminative.

3. In Logistic regression, we learn the parameters by maximizing the log-likelihood of the data.

   Solution: False. We maximize the conditional log-likelihood, not the log-likelihood.

4. By using L2 regularization, we decrease the bias but increase the variance of the logistic regression classifier.

   Solution: False. We increase the bias and reduce the variance.

5. Logistic regression is more scalable than Naive Bayes because its time complexity is much smaller than Naive Bayes.

   Solution: False. LR solves a concave optimization problem using an iterative algorithm. It is much slower than Naive Bayes in which the learning algorithm just makes one pass through the data.

Question 4: Neural Networks (10 points)

(a) (10 points) Draw a neural network that represents the parity function over three variables. Formally, the parity function is given by: $f(x_1, x_2, x_3) = +1$ iff exactly 1 or 3 variables are assigned to 1 (or True). Otherwise, $f(x_1, x_2, x_3) = -1$. Note that x_1, x_2, x_3 are Boolean variables and take values from the set $\{0, 1\}$.

Solution: Several solutions possible here.

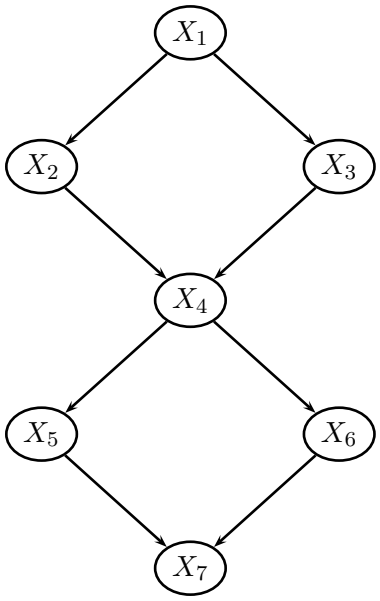
Question 5: VC-Dimension (10 points)

(a) (5 points) Show that the VC-dimension of circles centered at the origin is greater than or equal to 2.

(b) (5 points) Show that the VC-dimension of arbitrary circles (namely they can be centered at any point in the 2-D space) is greater than or equal to 3.

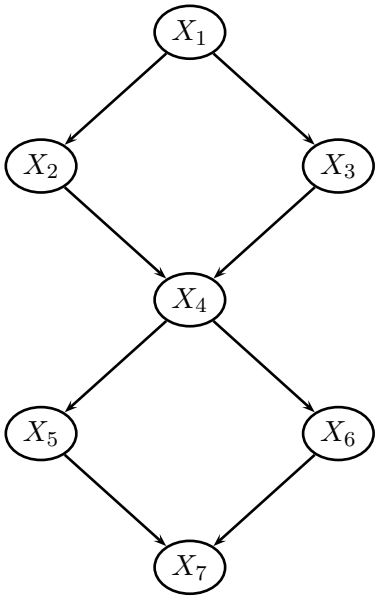
Question 6: Bayesian networks (10 points)

Consider the Bayesian network given below:



(a) (5 points) What is the time and space complexity of variable elimination along the ordering: X_4, X_1, X_3, X_2, X_5, X_6 for computing $P(X_7 = 1)$. Justify your answer.

The Bayesian network is redrawn for convenience.



(b) (5 points) Give an ordering of variables that has the best time complexity for computing $P(X_7 = 1)$ using variable elimination. Prove that (or at least provide an informal argument) no other ordering is better than your proposed ordering.

Question 7: AdaBoosting (10 points)

Consider the following labeled data (x_1, x_2, y) where x_1 and x_2 are the attributes and y is the class variable (i is the example index and not an attribute. Therefore, you cannot use it to make classification decisions):

i	x_1	x_2	y
1	11	3	$-$
2	10	1	$-$
3	4	4	$-$
4	12	10	$+$
5	2	4	$-$
6	10	5	$+$
7	8	8	$-$
8	6	5	$+$
9	7	7	$+$
10	7	8	$+$

In this problem, you will use Adaboost to learn a hidden function from this set of training examples. We will use two rounds of AdaBoost to learn a hypothesis for this data set. Recall that in round number, AdaBoost chooses a weak learner that minimizes the weighted error ϵ. As weak learners, you will use **axis parallel lines** of the form

- if $x_1 > a$, then $+$ else $-$ or
- if $x_2 > b$, then $+$ else $-$, for some integers a, b

(either one of the two forms, not a disjunction of the two).

(a) (2 points) The first step of AdaBoost is to create an initial data weight distribution D_1 (also called calculating the data weighting co-efficients). What are the initial weights given to data points with index 4 and 7 by the AdaBoost algorithm, respectively?

(b) (3 points) Which of the following three hypotheses minimizes the weighted error in the first round of AdaBoost, using the distribution D_1 computed in the above question? Circle one. Justify your answer.

(1) $h_1 : x_2 > 9$ (2) $h_2 : x_2 > 4$ and (3) $h_3 : x_2 > 7$.

Solution: $y > 4$, because it will misclassify only one point.

(c) (2 points) What is the weighted error ϵ of the best classifier computed above in part (b)?

(d) (3 points) Which of the following three hypotheses minimizes the weighted error in the second round of AdaBoost. Circle one. Justify your answer.

(1) $h_1 : x_2 > 9$ (2) $h_2 : x_1 > 5$ and (3) $h_3 : x_2 > 7$.

Question 8: Learning Bayesian networks (10 points)

Consider a Bayesian network with edges $A \rightarrow B$ and $A \rightarrow C$, and parameters which are given below:

- $P(A = 1) = 0.3$
- $P(B = 1|A = 1) = 0.5$, $P(B = 1|A = 0) = 0.8$
- $P(C = 1|A = 1) = 0.1$, $P(C = 1|A = 0) = 0.5$

Consider the dataset given below:

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

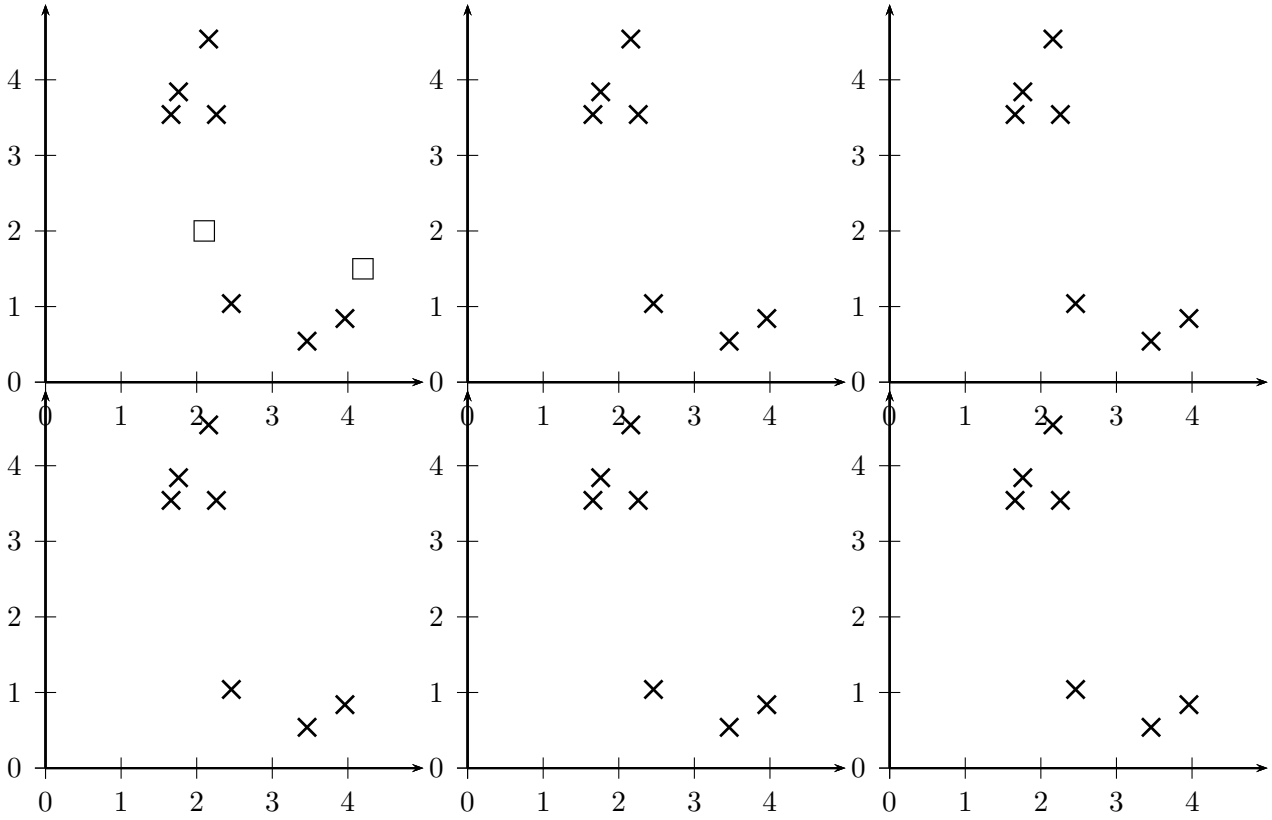
Assume that the CPTs are the CPTs at some iteration of EM. What you are going to do is derive the new set of parameters after running one iteration of EM.

(a) (5 points) Show the calculations involved in the E-step. Recall that in the E-step, you make the dataset bigger (by considering all possible completions) and weigh each new data point appropriately.

(b) (5 points) Show the calculations involved in the M-step. What are the new parameters?

Question 9: Clustering Techniques (10 points)

(a) (6 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.

(b) (4 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.

Question 10: Support Vector Machines (10 points)

(a) (10 points) Using the following 2-D dataset (x_1 and x_2 are the attributes and y is the class variable), find the linear SVM classifier. Do your optimization using the dual problem. Namely, provide an explicit expression for the dual optimization problem, solve it (compute the values of the various α_i's) and use the solution to compute the weights attached to the two attributes as well as the bias term.

Dataset:

x_1	x_2	y
1	0	$+$
-1	2	$-$
0	-1	$+$

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

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

Recall that $Entropy(x, y) = Entropy(y, x)$. Here x and y denote the number of positive and negative examples in the dataset.

```
Entropy(0,1) = 0.0
Entropy(0,2) = 0.0
Entropy(0,3) = 0.0
Entropy(0,4) = 0.0
Entropy(0,5) = 0.0
Entropy(0,6) = 0.0
Entropy(1,1) = 1.0
Entropy(1,2) = 0.9183
Entropy(1,3) = 0.8113
Entropy(1,4) = 0.7219
Entropy(1,5) = 0.6500
Entropy(1,6) = 0.5917
Entropy(2,2) = 1.0
Entropy(2,3) = 0.9709
Entropy(2,4) = 0.9183
Entropy(2,5) = 0.8631
Entropy(3,3) = 1.0
Entropy(3,4) = 0.9852
```