Midterm: CS 6375
Fall 2019

The exam is closed book (1 page cheat sheet allowed). Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use an additional sheet (available from the instructor), write your name on the sheet and staple it to your exam.

- **NAME** _____
- **UTD-ID if known** _____

Question	Points	Score
Decision Trees	10	
Neural Networks	10	
Support Vector Machines	10	
Short Questions	10	
AdaBoost	10	
Total:	50	

Question 1: Decision Trees (10 points)

(a) (5 points) Is the following statement True/False. Justify your answer. No credit if the justification/explanation is incorrect.

The decision tree algorithm we discussed in class (the one using the information gain heuristic without pruning) yields a globally optimal decision tree. (By optimal, we mean that it yields a decision tree having the smallest number of nodes and the highest accuracy on the training set.)

(b) (5 points) Is the following statement True/False. Justify your answer. No credit if the justification/explanation is incorrect.

You are given a d-dimensional linearly separable training data having n examples. The size of the decision tree for this data is guaranteed to be polynomial in d.

Question 2: Neural Networks (10 points)

(a) (6 points) Draw a neural network having minimum number of nodes that represents the following function. Please provide a precise structure as well as a setting of weights. You can only use **simple threshold units** (namely, $o = +1$ if $\sum_i w_i x_i > 0$ and $o = -1$ otherwise) as hidden units and output units. X_1, X_2 and X_3 are 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

(b) (2 points) True/False. Explain your answer in 1-2 sentences. Given a neural network having one hidden layer and at least n hidden nodes where n is the number of features, the back-propagation algorithm is susceptible to initialization. Namely, the parameters returned by the algorithm will be different for different initialization strategies.

(c) (2 points) Describe one approach to prevent over-fitting in neural networks.

Question 3: Support Vector Machines (10 points)

Consider the dataset given below (x_1, x_2 are the attributes and y is the class variable):

x_1	x_2	y
0	0	-1
$-1/3$	1	$+1$
$1/3$	1	$+1$
-2	-2	$+1$

(a) (3 points) Write the expression for the primal problem for this dataset under the assumption that you are using the linear SVM.

(b) (2 points) Is the data linearly separable? Circle one: YES NO.

(c) (5 points) If your answer to the previous question is NO, please suggest a suitable Kernel which will ensure separability (namely, the function used separates the positive examples from the negative examples). If your answer is YES, please use the linear Kernel for this part. Write the expression for the SVM dual problem w.r.t. the Kernel and the data given above.

Question 4: Short Questions (10 points)

(a) (3 points) Assume that you are given a Naive Bayes model defined over n binary features and one class variable having two values. Assume that at test time, k out of the n features are missing. Then, is the following statement True or False. *The Naive Bayes model will be impractical because determining the posterior marginal probability distribution over the class variable given the observed features will require time that scales exponentially in k in the worst case*. Explain your answer.

(b) (3 points) *The bias of a k nearest neighbor classifier increases with k*. True/False. Explain your answer.

(c) (4 points) Let us assume that the data (y) was generated from the following distribution:

$$\Pr(y) = \frac{\theta^y e^{3.5\theta}}{y!}$$

Let us assume that you are given n data points $y_1, \ldots, y_n$ drawn independently from $\Pr(y)$.

Write down the expression for the log-likelihood of data.

Derive an expression for θ such that the log-likelihood of data is maximize (namely find the maximum likelihood estimate of θ).

Question 5: AdaBoost (10 points)

Consider the following dataset. $(X_1, X_2) \in \mathbb{R}^2$ and Y is the class variable.

X_1	X_2	Y
0	8	$-$
1	4	$-$
3	7	$+$
-2	1	$-$
-1	13	$-$
9	11	$-$
12	7	$+$
-7	-1	$-$
-3	12	$+$
5	9	$+$

We will use two rounds of AdaBoost to learn a hypothesis for this data set. Consult the AdaBoost algorithm given on the last page of this exam. In each round m, AdaBoost chooses a weak learner that minimizes the error ϵ_m. As weak learners, use hypotheses of the form (a) $h_1 \equiv [X_1 > \theta_1]$ or (b) $h_2 \equiv [X_2 > \theta_2]$, for some integers θ_1, θ_2 (either one of the two forms, not a disjunction of the two). There should be no need to try many values of θ_1, θ_2; appropriate values should be clear from the data.

(a) (3 points) Which weak learner will AdaBoost choose in the first iteration ($m = 1$)? Be sure to provide a precise value for θ_1 or θ_2 for this learner.

(I have copied the data from the previous page to this page for your convenience.)

X_1	X_2	Y
0	8	$-$
1	4	$-$
3	7	$+$
-2	1	$-$
-1	13	$-$
9	11	$-$
12	7	$+$
-7	-1	$-$
-3	12	$+$
5	9	$+$

(b) (7 points) Which weak learner will AdaBoost choose in the second iteration ($m = 2$)? Again, be sure to provide a precise value for θ_1 or θ_2 for this learner.