Midterm: CS 6375
Spring 2018

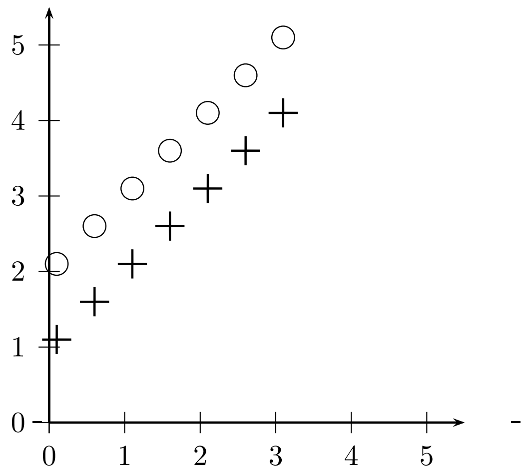
The exam is closed book (1 cheat sheet allowed). Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, use an additional sheet (available from the instructor) and staple it to your exam.

- **NAME** ____________________
- **UTD-ID if known** ____________________

Question	Points	Score
Decision Trees	10	
Neural Networks	10	
Support Vector Machines (SVMs)	10	
Short Questions	20	
Total:	50	

Question 1: Decision Trees (10 points)

Consider a large dataset D having n examples in which the positive (denoted by $+$) and negative examples (denoted by $\circ$) follow the pattern given below. (Notice that the data is clearly linearly separable).



(a) (5 points) Which among the following is the "best upper bound" (namely the smallest one that is a valid upper bound) on the number of leaves in an optimal decision tree for D (n is the number of examples in D)? By optimal, I mean a decision tree having the smallest number of nodes. Circle the answer and explain why it is the best upper bound. No credit without a correct explanation.

1. $O(n)$
2. $O(\log n)$
3. $O(\log \log n)$
4. $O((\log n)^2)$

Consider the dataset given below. X_1, X_2, X_3 and X_4 are the attributes (or features) and Y is the class variable.

X_1	X_2	X_3	X_4	Y
3	0	0	1	+
1	1	0	0	−
2	0	1	1	−
5	1	1	0	+
4	1	0	1	+
6	0	1	0	−

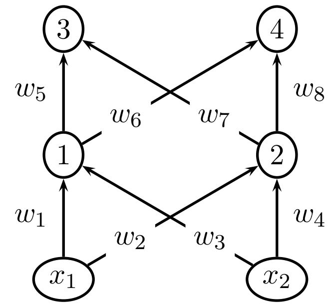
(b) (2 points) Which attribute (among X_1, X_2, X_3 and X_4) has the highest information gain?

(c) (3 points) In the above dataset, is the attribute having the highest information gain useful (namely will it help improve generalization)? Answer YES/NO and then

- Explain why the attribute is useful if your answer is "YES."
- If your answer is "NO", explain how will you change the information gain criteria so that such useless attributes are not selected.

Question 2: Neural Networks (10 points)

Consider the Neural network given below.



Assume that all internal nodes and output nodes compute the sigmoid $\sigma(t)$ function. In this question, we will derive an explicit expression that shows how back propagation (applied to minimize the least squares error function) changes the values of w_1, w_2, w_3, w_4, w_5, w_6, w_7 and w_8 when the algorithm is given the example (x_1, x_2, y_1, y_2) with y_1 and y_2 being outputs at 3 and 4 respectively (there are no bias terms). Assume that the learning rate is η. Let o_1 and o_2 be the output of the hidden units 1 and 2 respectively. Let o_3 and o_4 be the output of the output units 3 and 4 respectively.

Hint: Derivative: $\frac{d}{dt}\sigma(t) = \sigma(t)(1 - \sigma(t))$.

(a) (2 points) Forward propagation. Write equations for o_1, o_2, o_3 and o_4.

(b) (4 points) Backward propagation. Write equations for δ_1, δ_2, δ_3 and δ_4 where δ_1, δ_2 , δ_3 and δ_4 are the values propagated backwards by the units denoted by 1, 2, 3 and 4 respectively in the neural network.

(c) (4 points) Give an explicit expression for the new (updated) weights w_1, w_2, w_3, w_4, w_5, w_6, w_7 and w_8 after backward propagation.

Question 3: Support Vector Machines (SVMs) (10 points)

Consider the following 2-D dataset (x_1 and x_2 are the attributes and y is the class variable).

Dataset:

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

(a) (5 points) Precisely write the expression for the dual problem (assuming Linear SVMs). Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ be the lagrangian multipliers associated with the four data points.

(b) (5 points) Identify the support vectors and compute the value of α_1, α_2, α_3 and α_4. (Hint: You don't have to solve the dual optimization problem to compute α_1, α_2, α_3 and α_4.)

Question 4: Short Questions (20 points)

Consider a linear regression problem $y = w_1 x + w_2 z + w_0$, with a training set having m examples $(x_1, z_1, y_1), \ldots, (x_m, z_m, y_m)$. Suppose that we wish to minimize squared error (loss function) given by:

$$Loss = \sum_{i=1}^{m} (y_i - w_1 x_i - w_2 z_i - w_0)^2$$

under the assumption $w_1 = w_2$.

(a) (5 points) Derive a batch gradient descent algorithm that minimizes the loss function. (Note the assumption $w_1 = w_2$; also known as parameter tying).

(b) (5 points) When do you expect the learning algorithm for the logistic regression classifier to produce the same parameters as the ones produced by the learning algorithm for the Gaussian Naive Bayes model with class independent variances.

(c) (4 points) Describe a reasonable strategy to choose k in k-nearest neighbor.

(d) (6 points) You are given a coin and a thumbtack and you put Beta priors $Beta(5, 5)$ and $Beta(20, 20)$ 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 20 heads and 80 tails for both the coin and the thumbtack. Are each of the following two statements true or false.

- The MLE estimate of both the coin and the thumbtack is the same.
- 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.

Explain your answer mathematically.