CS 436/580L – Introduction to Machine Learning
Midterm Exam

Name: _____

B-Number: _____

Instructions:

- There are 50 possible points on this exam.
- Write your name on this page.

1. **[16 points] SECTION 1: SHORT QUESTIONS**

   (a) **[2 points]** Consider a data set that is linearly separable and two solutions: a perceptron and an SVM. Are both these solutions are identical? True or False. Explain. [Answers without explanations will receive no credit.]

   (b) **[2 points]** In the backpropagation algorithm, explain the need to sum the error over downstream nodes to calculate the weight updates for internal nodes.

(c) **[2 points]** For linearly separable data, can a small slack penalty hurt the training accuracy when using a linear SVM (no kernel)? If so, explain how. If not, why not?

(d) **[2 points]** Explain the difference between gradient descent and stochastic gradient descent. Clearly identify situations where each of these algorithms are more appropriate.

(e) **[4 points]** For each of the scenarios below, choose which models are appropriate: a) Naive Bayes, b) Logistic Regression, c) Decision Trees, d) Linear SVMs, e) SVMs with slack variable, f) SVM with a polynomial kernel, g) Perceptrons, h) Neural Networks. If there is a most appropriate model, indicate that as well. Briefly explain your answer.

[HINT: *There may be more than one appropriate model for each of the scenarios.*]

i. Data that is linearly separable.

ii. Data that has outliers.

iii. Continuous valued data.

iv. Data with many features.

v. Text data.

vi. Data that is not linearly separable.

vii. Ability to generate data.

viii. Data with missing values.

(C) **[4 points]** Consider the following dataset which has 3 boolean inputs (x, y, and z) and one boolean output, u. After naive Bayes learning is complete, what would be the predicted probability $p(u = 0|x = 1, y = 1, z = 0)$?

x	y	z	u
1	0	0	0
0	1	1	0
0	0	1	0
1	0	0	1
0	0	1	1
0	1	0	1
1	1	0	1

2. **[10 points] SECTION 3: POINT ESTIMATION**

(a) Given that it is virtually impossible to find a suitable "date" for boring, geeky computer scientists, you start a dating website called "www.csdating.com." Before you launch the website, you do some tests in which you are interested in estimating the failure probability of a "potential date" that your website recommends. In order to do that, you perform a series on experiments on your friends. You ask them to go on "dates" until they find a suitable match. The number of failed dates, k, is recorded.

A **[3 points]** Given that p is the failure probability, what is the probability of k failures before a suitable "match" is found by your friend. [Hint: Think geometric distribution, k failures before one success.]

A **[7 points]** You have performed m *independent* experiments of this form (namely, asked m of your friends to go out on dates until they find a suitable match), recording $k_1, ..., k_m$. Estimate the most likely value of p as a function of m and $k_1, ..., k_m$. [Hint: Use log-likelihood. Derivative of log(x) = 1/x.]

3. **SECTION 2: Decision Trees**

(a) The following dataset will be used to learn a decision tree for predicting whether a student can score high or low in an exam, high (H) or low (L) based on when the exam is scheduled (M: morning, A: afternoon, E: evening), whether they are prepared (Y: yes, N: no), and the number of pages in the answer paper.

Time	Preparation	Pages	Score
M	Y	2	L
M	N	2	L
M	N	2	L
A	N	2	L
A	N	2	H
E	N	2	H
E	N	2	H
E	N	2	H
E	Y	3	H

i. **[1 points]** What is Entropy(Score | preparation = Y)?

ii. **[2 points]** Which attribute would a decision-tree building algorithm choose to use at the root of the tree (assume no pruning)? [Hint: Don't resort to calculating information gain. Visual inspection of the dataset will suffice to get the right answer.]

iii. **[3 points]** Draw the full tree that would be learned from these data (assume no pruning). The same hint applies here.

4. **SECTION 4: PERCEPTRONS and Neural Networks**

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

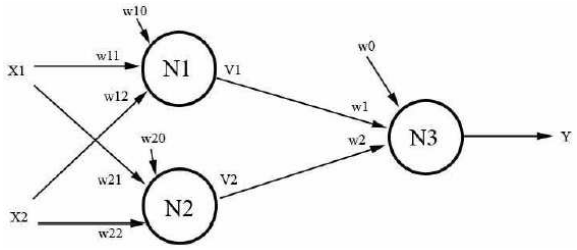
x_1	x_2	class
1	6	-1
3	3	-1
-3	2	+1
3	-5	-1
-4	-2	+1
0	5	+1

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.

- **[5 points]** Assume that the bias term (w_0x_0) 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.

- **[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 $x_0 = 0$)?
  (a) 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 don't have to run the algorithm again).
  (b) 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 ($x_b = +1$ and $x_b = -1$).

- Here is a simple 2-layer neural network with 2 hidden units and a single output unit. Consider the linear activation function $y = C \cdot \sum_i w_i x_i$ where C is a constant multiplied by a which is the weighted sum of its inputs. Also, consider the non-linear logistic activation function $y = \sigma(\text{a})$ where, $\sigma(\text{a}) = \frac{1}{1+e^{-a}}$.



(a) **[2 points]** Can this 2-layer network represent decision boundaries that a standard linear model cannot? (Assume all units use the linear activation function.) Explain your answer.

(b) **[2 points]** Now, assume the hidden units use the logistic activation function and the output unit still uses a linear activation function. Now, can this neural network represent non-linear decision boundaries? Explain your answer.

5. **[4 points] SECTION 6: Support Vector Machines**

Consider the following 1-dimensional data:

x	-3	0	1	2	3	4	5
Class	-	-	+	+	+	+	+

(a) **[2 points]** Draw the decision boundary of a linear support vector machine on this data and identify the support vectors.

(b) **[2 points]** Suppose we have another instance (x = -5,Class = +). What kernel will you use to classify the training data perfectly.

Scratch Paper

Scratch Paper