Department of Computer and Information Science and Engineering

UNIVERSITY OF FLORIDA

CAP4770/CAP5771 Fall 2016

Midterm Exam

Instructor: Prof. Daisy Zhe Wang

This is a in-class, closed-book exam.

This exam contains 5 single-sided sheets of paper (excluding this one).

Write all answers on these pages, preferably on the white space in the problem statement. Continue on the draft pages if running out of space but **clearly number** your answers if doing so.

Make sure you attack every problem; partial credit will be awarded for incomplete or partially correct results.

THIS IS A CLOSED BOOK, CLOSED NOTES EXAM.

Name:

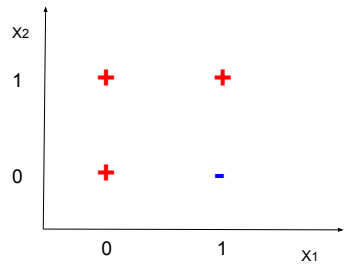
UFID:

For grading use only:

Question:	**I**	**II**	**III**	**Total**
Points:	**40**	**30**	**30**	**100**
Score:				

I. [40 points] Machine Learning.

A set of reasonably clean sample records was extracted by Barry Becker from the 1994 Census database. We are interested in predicting whether a person makes over 50K a year. For simplicity suppose we model the two features with two boolean variables $X_1, X_2 \in \{0,1\}$ and label $Y \in \{0,1\}$ where $Y = 1$ indicates a person makes over 50K. In Figure 1 we show three positive samples ("+" for $Y = 1$) and one negative samples ("-" for $Y = 0$). Please complete the following questions.



1 (6 points). If we train a KNN classifier (K=1) based on data in Figure 1, and then try to classify the same data. Which sample(s) must be misclassified by this classifier?

Solution: The $(1,0)$ sample must be misclassified.

2 (6 points). For predicting samples in Figure 1, which model is better: Logistic Regression or Linear Regression. Please explain why.

Solution: Logistic Regression. Because Logistic Regression predicts values between 0 and 1, which is consistent with the target space Y, but Linear Regression predicts any values.

3 (7 points). Is there any logistic regression classifier using X_1 and X_2 that can perfectly classify the examples in Figure 1? How about if we change label of point $(0,1)$ from "+" to "-"?

Solution: Logistic regression forms linear decision surface. Because data points in Figure 1 is linear separable, they can be perfectly classified. But if we make the change, then data points are not linearly separable so **no** logistic regression classifier can perfectly classify the examples.

4 (7 points). Suppose we have trained a linear regression model $y = ax + b$ where $a = 0.5$ and $b = 1.0$, on a set of training data points $D = \{(1.0, 1.6), (1.5, 1.5), (3.0, 2.4)\}$. Please calculate the mean squared errors of this model on D.

Solution: $MSE = \frac{0.1^2+0.25^2+0.1^2}{3} = 0.0275$

5 (7 points). Suppose we learn a Naive Bayes classifier from the examples in Figure 1, using MLE (maximum likelihood estimation) as the training rule. Write down the prior probabilities $P(Y)$, and conditional probabilities $P(X_i|Y = 1)$. Note: both $P(Y)$ and $P(X_i|Y)$ should be Bernoulli distributions.

Solution: $P(Y = 0) = 0.25, P(Y = 1) = 0.75$
$P(X_1 = 0|Y = 1) = \frac{2}{3}$
$P(X_1 = 1|Y = 1) = \frac{1}{3}$
$P(X_2 = 0|Y = 1) = \frac{1}{3}$
$P(X_2 = 1|Y = 1) = \frac{2}{3}$

6 (7 points). If we train a classifier based on data in Figure 1, and then we apply that classifier on a testing dataset. The testing confusion matrix is given by:

	+	-
+	9	9
-	1	5

What is the precision and recall of that classifier?

Solution: precision: 9/(9+1) = 0.9
recall: 9/(9+9) = 0.5

II. [30 points] Map-Reduce.

For each of the following problems describe how you would solve it using map-reduce. You should explain how the input is mapped into (key, value) pairs by the map stage, i.e., specify what is the key and what is the associated value in each pair, and, if needed, how the key(s) and value(s) are computed. Then you should explain how the (key, value) pairs produced by the map stage are processed by the reduce stage to get the final answer(s). If the job cannot be done in a single map-reduce pass, describe how it would be structured into two or more map-reduce jobs with the output of the first job becoming input to the next one(s).
You should just describe your solution algorithm. You should not translate the solution into detailed programming language. For each of the following question please describe: (1) The input and output of MapReduce program; (2) How the Map step works; (3) how you reduce the key-value pairs in the Reduce step.

1 (15 points). The input is a list of housing data where each input record contains information about a single house: (address, city, state, zip, value). The output should be the average house value in each zip code.

Solution: (1) Input: (address, city, state, zip, value), Output: (zip, value).
(2) Map Stage: Read and input and generate key-value pairs as output: key - "zip", value - "value"
(3) Reduce Stage: merge pairs of the same key by adding the value.

2 (15 points). The input contains two lists. One list gives voter information for every registered voter: (voter-id, name, age, zip). The other list gives disease information: (zip, age, disease). For each unique pair of age and zip values, the output should give a list of names and a list of diseases for people in that zip code with that age. If a particular age/zip pair appears in one input list but not the other, then that age/zip pair can appear in the output with an empty list of names or diseases, or you can omit it from the output entirely, depending on which is easier. (Hint: the keys in a map/reduce step do not need to be single atomic values.)

Solution: (1) Input: (voter-id, name, age, zip) and (zip, age, disease), Output: ((age, zip), [[values],[diseases]]).
(2) Read and input and generate key-value pairs as output: key - "(age, zip)", value - [[name],[]] for the first list, key - "(age, zip)", value - [[],[disease]] for the second list.
(3) Reduce Stage: merge pairs of the same key by merging two sublists of values.

III. [30 points] Hypothesis Testing.

A research group made a poll of 1029 US residents to estimate the satisfaction level $(0-10)$ to the US President. The average satisfaction level of that 1029 samples was $Y = 7.1$. Now we randomly select a group of elder people of $n = 25$ from that samples. Please establish a t-test to show if the average satisfaction level of 8.1 with standard deviation $s = 2.0$ from the selected group of elder people is sinificantly ($\alpha = 0.05$) different from the samples.

Hint:

$t_{critical} = 1.711$ for one-tails t-test with $df = 24$ and $\alpha = 0.05$
$t_{critical} = 2.064$ for two-tails t-test with $df = 24$ and $\alpha = 0.05$
$t_{critical} = 1.708$ for one-tails t-test with $df = 25$ and $\alpha = 0.05$
$t_{critical} = 2.060$ for two-tails t-test with $df = 25$ and $\alpha = 0.05$

1. (6 points) Write down H_0/H_1.

Solution: $H_0 : Y = 7.1$
$H_1 : Y \neq 7.1$

2. (6 points) Estimate standard error of the mean.

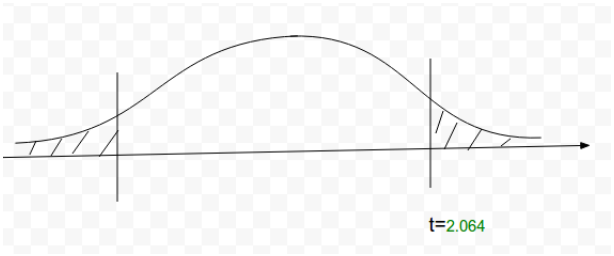
Solution: $S_M = \frac{s}{\sqrt{n}} = \frac{2.0}{\sqrt{25}} = 0.4$

3. (6 points) Calculate the t-statistic.

Solution: $t = \frac{M-Y}{s_M} = \frac{8.1-7.1}{0.4} = 2.5$

4. (6 points) Sketch the critical regions.

Solution:



5. (6 points) Explain if the difference is significant.

Solution: Because $t_{critical} = 2.064$ for two-tails with $df = 24$ and $\alpha = 0.05$. Thus we should reject H_0. The difference is significant under $\alpha = 0.05$.

(*This page is intentionally left blank*)