Department of Computer and Information Science and Engineering

UNIVERSITY OF FLORIDA

CAP4770/CAP5771 Fall 2015

Exam II

Instructor: Prof. Daisy Zhe Wang

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

This exam contains 4 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:	**4**	**4**	**4**	**12**
Score:				

I. [4 points] Clustering.

suppose we have a 6 points in 2D space: A1(0, 0), A2(0, 4), A3(2, 1), A4(2, 3), A5(4, 0), A6(4.4) and would like to run Kmeans clustering on the 6 points with k(cluster number) = 2, using **Euclidean distance**. Now if we choose the A1 & A2 as the initial cluster centroids:

1. After 1st iteration of calculation, what are the new centroids and assignment of the 6 points? Show your calculations.
2. What is the final clustering result(final centroids and assignment of points to each cluster)? Visualize the final clustering result(draw a circle around the points in the same cluster in a 2D grid, and mark the centroids).

Note: The Euclidean distance of two points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ is : $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$.

Solution: 1. First we need to calculate each point's distance to the intial centers(detailed calculation omitted here), The assigments are {A1, A3, A5}, {A2, A4, A6} new centroids (2, 1/3), (2, 3 + 2/3).

2. at the second iteration, calculate the distance of each point to the new centroids, (not shown here) and the assigments are {A1, A3, A5}, {A2, A4, A6}
Reached the stopping criteria because assignment does not change now. (visulization not shown here)

II. [4 points] Naive Bayes.

Consider the following data set with three Boolean features, W, X, Y , and Boolean classification C.

W	X	Y	C
T	T	T	T
T	F	T	F
T	F	F	F
F	T	T	F
F	F	F	T

We now encounter a new example: W = F, X = T, Y = F. How should this example be classified using the Naive Bayes method? Show your computations.

Solution: $P(W=F|C=T)P(X=T|C=T)P(Y=F|C=T)P(C=T)=1/2*1/2*1/2*2/5=1/20$
$P(W=F|C=F)P(X=T|C=F)P(Y=F|C=F)P(C=F)=1/3*1/3*1/3*3/5=1/45$

Label the example as T

III. [4 points] Classification.

Suppose you are training a machine learning model for T steps, observing its performance at various values of T. You observe:

T	Training Set Accuracy	Validation Set Accuracy
5	0.85	0.84
10	0.89	0.79

Explain the difference in scores at T=10. Which of the two different steps for this model would you use? Explain your choice.

Solution: Score much lower on validation set at T=10 $->$ overfitting. The model has been trained too long and it has ceased finding true trends in this data, rather its fitting this particular sample of data. You should use the model at T=5. It has the highest validation set accuracy, and doesnt show signs of overftting.

(*This page is intentionally left blank*)