COMPUTER SCIENCE 4373/5473
Assignment #7

Points: 100 **Weight**: 2%
Due: Friday, Oct. 23, 2020 at 11:50 pm on BlackBoard
Note: Late assignment will not be accepted without instructor's pre–approval.

Instruction: This assignment should be completed individually. Please make sure your answer is legible (and preferably formatted using MS Words or LaTeX/LyX). Please also submit to the BlackBoard a zip file yourName_hwk07.zip which should contain your solution in a PDF, a Word document, or a Jupyter Notebook (with narrative formatted in markdown cells), the program source code, the input data, and the output of your program.

Notice: *Given the nature of on-line course, we will require you to practice using Words, Markdown, or HTML to write and format your homework solutions (**no scanned smeared image please**). It will prepare you for taking the on-line exams, where only a Words style editor (with HTML support) is available.*
Suggestion: *If the trace of an algorithm involves a lot of calculations, you may want to show the details in the first place, and then write some scripts to perform the calculation for subsequent cases. The Jupyter notebook will be very handy for such cases.*

1. [**30**] The following table contains the outcome of classifying 10 testing tuples using a probabilistic classifier. For each tuple, the actual class (P or N) is given in the second column, and the probability (of class P) returned by the classifier is in the third column.

tupleID	actual class	probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

   (a) [15] For each row of the table, assume that the threshold of probability for predicting class P is the probability in that row (see Example 8.11 on page 375 of the textbook for further description), determine the numbers of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN) the true positive rate (TPR) and false positive rate (FPR) of the entire set of tuples.

(b) [15] Plot the ROC curve for the data (by hand drawing or using a Python notebook).

2. [**30**] Suppose we have two predictive models, M_1 and M_2 and run 10 rounds of 10-fold cross-validation test using the two models. The error rates obtained from the test are given in the following table.

Models	round 1	2	3	4	5	6	7	8	9	10
M_1	30.5	32.2	20.7	20.6	31.0	41.0	27.7	26.0	21.5	26.0
M_2	22.4	14.5	22.4	19.6	20.7	20.4	22.1	19.4	16.2	35.0

There are two cases to consider.

(a) [15] In each round, a training set and a testing set are determined based on the 10-fold cross-validation method, and then the same training set is used to build the two models, and the same testing set is used to obtain errors for the two models.

(b) [15] The models are built and tested independently. A ten-round 10-fold method is first applied to M_1 to obtain the first row in the table. A second ten-round 10-fold method is then applied to $M_2$to obtain the second row in the table.

Perform a t-test at a significance level of 0.001 for each case to determine if one model is significantly better than the other. You can either write a program or use a t-distribution table (such as the one attached to this assignment) to do this test.

3. [**40**] Write a Python Jupyter notebook that performs a synchronized ten rounds 10-fold cross validation tests to obtain the classification accuracy scores for the Naive Bayesian and the AdaBoost classifiers learned from the dataset given in hwk07.csv, in which column H is the class label. Specifically, in each round, the same training set should be used to train the classifiers, and the same testing set should be used to measure the classification accuracy scores. Your program should use functions from the SciKit-Learn to create random folds, to learn classifiers, to test the classifiers and and to calculate classification accuracy scores.