Assignment 6

Due: 3/18

Note: Show all your work.

Problem 1 (10 points) Consider the following confusion matrix.

		predicted class	
actual class		C1	C2
	C1	428	72
	C2	328	781

Note: C1 is positive and C2 is negative.

Compute *sensitivity*, *specificity*, *precision*, *accuracy*, *F-meassure*, F_2, and MCC measures.

Problem 2 (10 points) Suppose you built two classifier models *M*1 and *M*2 from the same training dataset and tested them on the same test dataset using 10-fold cross-validation. The error rates obtained over 10 iterations (in each iteration the same training and test partitions were used for both *M*1 and *M*2) are given in the table below. Determine whether there is a significant difference between the two models using the statistical method discussed that we discussed in the class (this method is also discussed in Section 8.5.5, pp 372-373 of the textbook). Use a significance level of 1%. If there is a significant difference, which one is better?

Iteration	M1	M2
1	0.13	0.19
2	0.12	0.1
3	0.09	0.12
4	0.15	0.1
5	0.03	0.07
6	0.07	0.05
7	0.2	0.1
8	0.14	0.11
9	0.12	0.07
10	0.14	0.11

Note: When you calculate *var(M1 – M2)*, calculate a sample variance (not a population variance).

Problem 3 (10 points). The following table shows a test result of a classifier on a dataset.

Tuple_id	Actual Class	Probability
1	P	0.72
2	N	0.70
3	N	0.87
4	P	0.92
5	P	0.75
6	P	0.89
7	N	0.82
8	P	0.73
9	N	0.91
10	P	0.96

Problem 2-1. For each row, compute *TP*, *FP*, *TN*, *FN*, *TPR*, and *FPR*.
Problem 2-2. Plot the ROC curve for the dataset. You must draw the curve yourself (i.e., don't use Weka, R, or other software to generate the curve).

Problem 4 (10 points). This is a practice of comparing performance of classifier models using ROC curves. You can plot ROC curves using Weka Knowledge Flow. On the Blackboard course web site, I posted a Weka Manual under Course Documents. How to use Knowledge Flow is described in Chapter 7. Following the instruction in the manual (especially Section 7.4.2), build and test Logistic and RandomForest classifiers on *crx-data.arff* dataset, and capture the screenshot which shows two ROC curves. Include this screenshot in your submission. Compare and discuss the performance of the two models using the ROC curves.

Problem 5 (Extra Credit 10 points). This problem is a practice of using Weka to perform t-tests to compare performance of classifier models. There is an instruction in the *Experimenter* chapter (Chapter 6) of Weka 3.8 Manual. It is your responsibility to read the manual and learn how to use Weka's Experimenter to perform t-tests.

For this problem, build three classifier models, Naïve Bayes, Multilayer Perceptron (neural network), and J48 from the *crx-data.arff* dataset, which you used in Problem 4. Then, perform t-tests and determine the ranks of the classifier models based on the test result. You must show, step by step, all screenshots of Weka Experimenter that you have gone through and also you must explain how you determined their ranks.

Submission:

Include all answers in a single file and name it *lastName_firstName*_HW6.EXT. Here, "EXT" is an appropriate file extension (e.g., docx or pdf). If you have multiple files, then combine all files into a single archive file. Name the archive file as

*lastName_firstName*_HW6.EXT. Here, “EXT” is an appropriate archive file extension (e.g., zip or rar). Upload the file to Blackboard.