Assignment 6

Due: 3/20

Note: Show all your work.

Problem 1 (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 in Section 6 of the online lecture Module 4 (also 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	0.21	0.13
2	0.12	0.1
3	0.09	0.20
4	0.15	0.2
5	0.03	0.15
6	0.07	0.05
7	0.13	0.14
8	0.14	0.21
9	0.05	0.23
10	0.14	0.17

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

Problem 2 (10 points). The following table shows a test result of a classifier on a dataset.

Tuple_id	Actual Class	Probability
1	P	0.82
2	N	0.75
3	N	0.94
4	P	0.85
5	P	0.81
6	P	0.90
7	N	0.74
8	P	0.73
9	N	0.91
10	P	0.76

Problem 2-1. For each row, compute *TP*, *FP*, *TN*, *FN*, *TPR*, and *FPR*.
Problem 2-2. Plot the ROC curve for the dataset.

Problem 3 (10 points). For this problem, you will run bagging and boosting algorithms that are implemented on Weka on the *german-bank.arff* dataset

Problem 3-1 (5 points). Run Bagging twice first with Naïve Bayes as a base classifier and next with J48 as a base classifier. For each result, capture the screenshot of a part of Classifier Output window that shows "Correctly Classified Instances" and "Confusion Matrix" and include them in your submission. Compare and discuss the performance of the two models with the result from homework 5.

Problem 3-2 (5 points) Run AdaBoostM1 twice first with Naïve Bayes as a base classifier and next with J48 as a base classifier. For each result, capture the screenshot of a part of Classifier Output window that shows "Correctly Classified Instances" and "Confusion Matrix" and include them in your submission. Compare and discuss the performance of the two models with the result from homework 5. Also compare the result with that of Problem 3-1 (Bagging result).

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 Section 7. Following the instruction in the manual (especially Section 7.4.2), build and test SimpleLogistic and RandomForest classifiers on *german-bank.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.