Assignment 4

Note: Show all your work as you have on previous assignments.

Problem 1 (20 points) Consider the following confusion matrix.

	predicted class		
actual class		C1 (positive)	C2 (negative)
	C1 (positive)	254	36
	C2 (negative)	72	324

Compute *sensitivity*, *specificity*, *precision*, *accuracy*, *F-meassure*, *F*2, and MCC measures. You have to show all your calculations.

Problem 2 (20 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	0.12	0.08
2	0.12	0.1
3	0.15	0.22
4	0.15	0.1
5	0.03	0.07
6	0.17	0.11
7	0.2	0.1
8	0.14	0.11
9	0.1	0.17
10	0.14	0.11

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

You must show all calculations, including the calculation of the test statistic.

Problem 3 (20 points). The following table shows a test result of a classifier on a dataset.

Tuple_id	Actual Class	Probability
1	P	0.92
2	N	0.70
3	N	0.76
4	P	0.92

5	P	0.83
6	P	0.89
7	N	0.79
8	P	0.73
9	N	0.82
10	P	0.96

(1). For each row, compute *TP*, *FP*, *TN*, *FN*, *TPR*, and *FPR*.
(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 (20 points). This problem is a small experiment of handling an unbalanced dataset for classification. Use *a4_p4_train.arff* and *a4_p4_test.arff* files and use J48 on Weka.

(1). Build a decision tree model from *a4_p4_train.arff* using J48 and test it on *a4_p4_test.arff*. Include the resulting confusion matrix in your submission.
(2). Create an undersampled training dataset from *a4_p4_train.arff* and name it *a4_p4_train_undersampled.arff*. Build a decision tree model from *a4_p4_train_undersampled.arff* using J48 and test it on *a4_p4_test.arff*. Include the resulting confusion matrix in your submission.
(3). Create an oversampled training dataset from *a4_p4_train.arff* and name it *a4_p4_train_oversampled.arff*. Build a decision tree model from *a4_p4_train_oversampled.arff* using J48 and test it on *a4_p4_test.arff*. Include the resulting confusion matrix in your submission.
(4). What conclusion can you draw from this experiment?

Note:

- You may use any tool(s) when creating undersampled dataset and oversampled dataset. You must describe the tool(s) you used.
- If you know how to use Python for classification, you may use Python's *DecisionTreeClassifier* instead of Weka's J48. In this case, you must submit Python script file(s) as well as confusion matrices. Use *a4_p4_train.csv* and *a4_p4_test.csv* files

Problem 5 (20 points). Use JMP Pro to build and test five classifier models – Naïve Bayes, KNN, Partition (decision tree), Boosted Tree, and Neural Network – following the instruction in *JMP-classification-assignment.pdf* file.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. Make sure the filename includes your name.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more**

numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.

- **If a problem requires you to work out formulas or calculations, you should provide this detailed work either in an Appendix in your Word/PDF or as a separate file. All separate files should include your name in the filename.**
- **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**