Homework Assignment 3

Due date: Wednesday 10/19 (hardcopy in class)

1. Recall that given a confusion matrix with TP true positives, TN true negatives, FP false positives (predicted positive but are true negatives) and FN false negatives, we defined alternative measures of success other than accuracy as follows: Precision = $\frac{TP}{TP + FP}$, Recall = $\frac{TP}{TP + FN}$, and TPR (true positive rate), FPR = $\frac{FP}{TN + FP}$ (false positive rate). A ROC curve plots TPR on the y axis vs. FPR on the x-axis.

A learning algorithm produces a hypothesis that ranks examples in the test set so that their labels are ordered as follows, where the list from left to right denotes ranking from highest to lowest score:

+ - + + + - + + - - - - - + - - - -

By this notation we mean that the top ranking example has a + label, the second ranked has a - label and so on. If the algorithm uses a threshold just below the second example then it has a true positive rate of 1/8 and a false positive rate of 1/12.

Draw a ROC curve, and a Precision-Recall curve for this ranking using a threshold after each “useful” boundary between example to generate the points on the curve.

2. Suppose that you are given a hypothesis for a particular domain, you test it on 100 independent examples and you observe 93% accuracy.

Calculate a $N = 0.90$ confidence interval for the accuracy of the hypothesis. Use the two solutions given in class to the problem of unknown variance to compare the effect on the size of the intervals. Solution (1) uses the fact that $p(1 - p) \leq \frac{1}{4}$ and solution (2) uses $\hat{p}$ instead of $p$ in the term for the variance.

3. You run your favorite algorithm on a new dataset using a 10 fold cross validation scheme and get the following accuracies in the folds: 0.88 0.75 0.79 0.80 0.91 0.83 0.77 0.79 0.82 0.85. Use the formulation of the $T$ confidence interval to calculate two intervals for the average accuracy, using confidence of 0.90 and 0.95 respectively.

4. In class we discussed filter methods and the wrapper method for feature selection. Please explain the wrapper method, in what ways it differs from filter methods, and whether these differences are advantages or disadvantages of the wrapper method.