CSCI567 Machine Learning (Fall 2008) Assignment #5

Instructor: Dr. Sofus A. Macskassy
TA: Cheol Han

Due time: 5:00pm, Nov 25, 2008

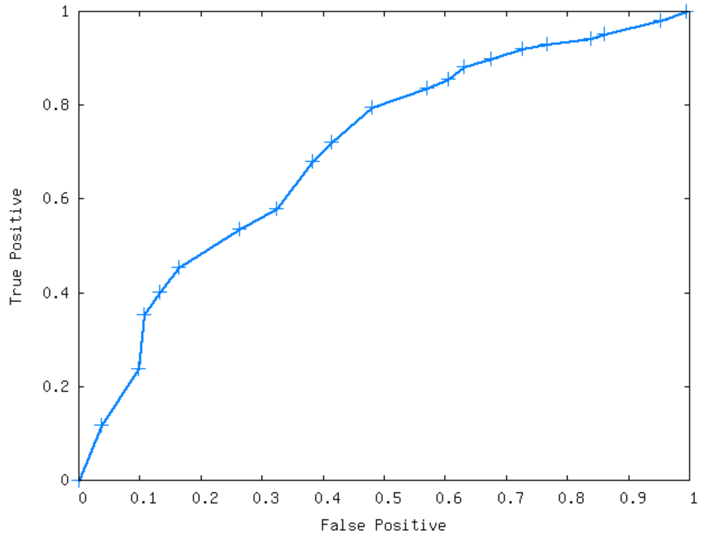
Student Name: ______________________

Student ID: ______________________

1. (**Cost-Sensitive Learning**, 20 points)
Suppose you run a learner on a data set, and it comes back with 21 distinct thresholds. You rank all instances by these thresholds and for each threshold you calculate the true-positive rate and false-positive rate and get the following table:

FP	TP	Treshold
0	0	100
0.0370943183851526	0.118230593753712	95
0.0968031236781045	0.238230593753712	93
0.107443653905474	0.352533774326488	90
0.132230725946249	0.402533774326488	83
0.164140599771191	0.454572062224442	76
0.263709739192864	0.534923219475437	71
0.32281131283499	0.57860503948066	60
0.382049216107127	0.681091064190152	45
0.414339785178702	0.721091064190152	41
0.479355121381361	0.79588320792435	38
0.569489556290122	0.835630358834613	33
0.604857471240753	0.855630358834613	29
0.629625969142352	0.88287327921599	27
0.67453458226317	0.900015074992161	23
0.725507232671498	0.920015074992161	20
0.766334495213198	0.93118655127536	16
0.838523833441139	0.942478390342781	13
0.859502764379531	0.952478390342781	12
0.952413400915812	0.979718100089046	4
0.993844820740749	1.00	0

Plotting the values, you get the following ROC curve:



Find the best threshold to use if:

a. FP’s are 5 times as costly as FN’s.
b. FP’s are as costly as FN’s.
c. FN’s are 3 times as costly as FP’s.
d. FN’s are 2 times as costly as FP’s.

Report the threshold and plot the lines for each of these constraints.

2. **(Bias-Variance,** 15 points) – based on Chapter 4, Question 9 in book

Let us say, given the samples $X = \{x_i , y_i \}$, we define $g(x) = y_1$, namely our estimate for any *x* is the *y* value of the first instance in the (unordered) dataset X.

a) What can you say about its bias and variance as compared with $g(x) = 3$?

b) What about its bias and variance as compared with $g(x) = \sum_i y_i / N$?

c) What about its bias and variance as compared with an ordered set (on *y*), such that $g(x) = \min_i y_i$?

3. (**Evaluation,** 20 points)

a. Consider a learned hypothesis, *h*, for some Boolean concept. When *h* is tested on a set of 200 examples, it classifies 91 correctly. What is the standard deviation and the 95% confidence interval for the true error rate for $error_D(h)$?

b. Suppose hypothesis *h* commits *r* = 17 errors over a sample of *n* = 93 samples drawn iid. What is the 90% confidence interval (two-sided) for the true error rate? What is the 95% one-sided interval (i.e., what is the upper bound *U* such that $error_D(h) <= U$ with 95% confidence)? What is the 90% one-sided interval?

c. You are about to test a hypothesis *h* whose $error_D(h)$ is known to be in the range between 0.15 and 0.45. What is the minimum number of examples you must collect to assure that the width of the two-sides 95% confidence interval will be smaller than 0.1?

d. Let us say we have three classification algorithms. How can we order these three from best to worst?

4. (**Weka Experiments with Bagging and Boosting**, 45 points)
In this part of the homework, you will experiment with Bagging and Boosting.

Learning Algorithms. Bagging and AdaboostM1 are available under the "Meta" category in Weka. Please use the following settings:

- Bagging: set numIterations to 30. You will run experiments with the classifier set to Trees.J48, Functions.logistic, and Bayes.naiveBayesSimple.
- AdaboostM1: set maxIterations to 30. Set weightThreshold to 100000. You will run experiments with the classifier set to the same three algorithms as for Bagging.

For J48, set the "unpruned" option to True (this is done in the meta-classifier dialogue box). You can use the default settings for all other parameters of J48, NaiveBayesSimple, and Logistic Regression. Optional: Rerun the experiments with pruning turned on and see if it makes any difference.

In addition to running Bagging and AdaBoostM1, you should rerun a single decision tree, a single Naive Bayes, and a single logistic regression.

Data Sets. You will apply these three algorithms to the same data sets that you have been using before: `hw_gmm`, `hw_step`, and `statlog`. You will not construct learning curves this time. Instead, you should just train and test on the following files:

o	Domain	Training Data File	Test Data File
o	statlog	statlog.arff	statlog_test.arff
o	hw_gmm	hw_gmm-250.arff	hw_gmm-test.arff
o	hw_step	hw_step-250.arff	hw_step-test.arff

NOTE: You should use the train and test files for `statlog` that were provided at the same location os this homework rather than the files you created in homework 3.

Results. You should turn in three tables in the following format (**10 points**):

hw_gmm:

Base learner	Single	Bagging	Boosting
J48	xxx	yyy	zzz
Logistic	xxx	yyy	zzz
NaiveBayes	xxx	yyy	zzz

hw_step:

Base learner	Single	Bagging	Boosting
J48	xxx	yyy	zzz
Logistic	xxx	yyy	zzz
NaiveBayes	xxx	yyy	zzz

statlog:

Base learner	Single	Bagging	Boosting
J48	xxx	yyy	zzz
Logistic	xxx	yyy	zzz
NaiveBayes	xxx	yyy	zzz

Where xxx gives the error rate of a single classifier of the indicated Base Learning, yyy gives the error rate of a bagging (30 iterations), and zzz gives the error rate of AdaboostM1 (maximum 30 iterations).

Answer the following questions (**5 points each**):

a) Which algorithms+data sets are improved by Bagging?
b) Which algorithms+data sets are improved by Boosting?
c) Can you explain these results in terms of the bias and variance of the learning algorithms applied to these domains? Are some of the learning algorithms unbiased for some of the domains? Which ones?

Now, set the number of iterations to 3,5,10,20, and 50 (for Bagging and Boosting both) and run J48, Logistic and naive Bayes on the `hw_gmm`, `hw_step` and `statlog` data sets.

Provide six tables in the following format (**10 points**):

```
DATASET (LEARNER)   Bagging                     Boosting
Iteration  TrainError  TestError   TrainError  TestError
           ActualIterations
3          xxx         yyy         zzz         www         kkk
5
10
20
50
```

Where DATASET is `hw_gmm`, `hw_step` or `statlog` and LEARNER is J48, Logistic or Naïve Bayes. You can get the training error by selecting 'test on train set' in the test set options. The number of actual iterations for boosting is reported when you run it.

Answer the following (**5 points each**):

d) Do the training and test error follow the pattern that you would expect? If yes, why is this what you would expect and if no, why not?
e) Explain why the number of actual iterations for Boosting is not always the same as the number of iterations that you requested.