CSC529: Advanced Data Mining
Assignment #2
Total: 60 points + 6 extra credit points
Due: Monday, February 12th, 2018, by midnight

The goal of this assignment is to: a) compare different classifiers by means of their accuracy performance (Problem 1), b) analyze the performance of different classifiers as a function of the dimensions of the training sets (Problem 2), c) evaluate probabilistic classifiers (Problem 3), and d) understand the basics of ensemble of classifiers (Problem 4).

Problem 1: (15 points) One of the shortcomings of the results in Problem 2 from Assignment#1 is due to the fact of considering a single sample for training set as well as for test set (a single trial of 66% for training and 34% for testing was used to build the model). The objective of this exercise is to repeat the same experiment, but now with different (same size) samples as training and test sets (in other words, repeating the holdout procedure).

a. (10 points) Repeat Problem 2.a&b from Assignment#1 on the Wine Recognition Dataset at least 30 times and report the means, variances, and Confidence Intervals (CI) for the accuracy results on the training and testing sets.
b. (5 points) Using a pair t-test, compare the mean accuracy of the Naïve Bayes and the mean accuracy of the Decision tree and discuss the results.

Note: To use the paired t-test, remember that you have to build and test the two classifiers (for our problem the decision tree and Naïve Bayes) on the same data. In other words, you do a holdout data partition and create and test both of your classifiers. Then you repeat the process for as many times/trials you want to repeat the procedure.

Problem 2: (15 points) The objective of this exercise is to analyze the performance of the previously trained Naïve Bayes and decision trees classifiers (using the approach in Problem 1) as a function of the sample size.

Let us vary the training data (and the corresponding values of the class attribute) such that the size of the training data is 25%, 35%, 45%, 55%, 65%, 75%, and 85% of the original data. Create a decision tree and a Naive Bayes classifier for each one of the seven different data splits, record the resubstitution error and the generalization error, and present the results in a plot with 4 curves (each corresponding to one type of error and one type of classifier) as a function of the size of the data. Perform an analysis of the results obtained.

Extra credit (3 points): Repeat Problem 2 on the following additional datasets and interpret the results:

winered.data: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

banknote.data: https://archive.ics.uci.edu/ml/datasets/banknote+authentication

Problem 3: (15 points) The objective of this exercise is to get familiar with the evaluation of probabilistic classifiers using ROC and Lift curves.

a. Repeat Problem 2.b from Assignment#1 on the Wine Recognition Dataset but this time considering only two classes (let us say, class 1 (positive class) versus class 2 and class 3 (negative class) since the ROC and lift curves can only be drawn for binary classification problems).
b. Draw the ROC curves for the Naïve Bayes performance on both the training and testing data. Interpret the graphs. If you would have to choose a certain probability threshold to maximize both sensitivity and specificity on the testing data, which threshold value would you select?
c. Draw the lift curves for the Naïve Bayes performance on both the training and testing data. Interpret the results. If the requirement is to get at least 80% accuracy on the data with a minimum cost of data acquisition, what size for the data would you recommend to reach that accuracy performance?

Problem 4: (15 points) The objective of this exercise is to get familiar with ensemble of classifiers and understand how the number of classifiers/learners affect the accuracy of the classifier.

The breast cancer Wisconsin dataset consists of 569 instances and 32 features. The class variable represents diagnosis (M=malignant, B=benign) and the features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. For the entire description of the dataset and download, use the following link: http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.names

a. (10 points) Model the diagnosis using ensemble learning based on bagging (with decision trees as learners) and plot the error rate as a function of the number of trees in the ensemble. Perform an analysis of your results.
b. (2.5 points) Explain if bagging is an appropriate choice for the proposed ensemble for this particular data.
c. (2.5 points) Briefly describe the differences between bagging and boosting.

Extra credit (3 points): Repeat Problem 4 on the following additional datasets and interpret the results:

d. winered.data: https://archive.ics.uci.edu/ml/datasets/Wine+Quality
e. banknote.data: https://archive.ics.uci.edu/ml/datasets/banknote+authentication

New Matlab functions for assignment #2:

"TreeBagger" = bagging of decision trees
"perfcurve"= ROC curves