Assignment 4 - Ensemble Methods

Extended Due Date: December 5, 11:30 pm

If you submit the assignment after the deadline, the following penalty is applied:

◇ 10% penalty if the submission is before December 8, 11:30 pm (if the mark before applying the penalty is 78 out of 100, after applying the penalty it is 78 - 7.8 = 70.2 out of 100).

Description:

In this assignment you are required to experiment with ensemble methods, in particular bagging, random forest and boosting. You will use the “spambase” data set from the UCI machine learning repository (available at https://archive.ics.uci.edu/ml/datasets/spambase). The data base contains spam and nonspam emails and obviously the problem is spam detection. The two classes are ”spam” (with label 1) and “nonspam” (with label 0). The number of features is 57.

First you have to separate the test and training data. Please use one third of the data for testing (as in the experiments performed by the authors of ESL for Fig 15.1 [1]). Next you will use the training set to fit (in other words train) the classifiers that are specified shortly. To train all these classifiers and to perform cross-validation (where required) you may use the utilities provided by scikit-learn. After that you will compute the test error for all the models.

The classifiers to train are:

◇ One decision tree classifier. You will use cross validation to select the best maximum number of leaves between 2 and at least 400. You can use the utilities provided by scikit-learn.

◇ 50 bagging classifiers. The base classifiers are decision trees with no restriction on the depth or number of leaves. You have to consider 50 values for the number of predictors in the ensemble, namely all integers between 50 and 2500, in increments of 50.

◇ 50 random forest classifiers. The base classifiers are decision trees with no restriction on the depth or number of leaves. You have to consider the same 50 values for the number of predictors as in the ensemble above.

◇ 50 Adaboost classifiers with decision stumps as the base classifiers (i.e., with max_depth = 1). You have to consider the same 50 values for the number of predictors as above.

◇ 50 Adaboost classifiers with decision trees with at most 10 leaves as the base classifiers (i.e., with max_leaf_nodes = 10). You have to consider the same 50 values for the number of predictors as above.

◇ 50 Adaboost classifiers with decision trees with no restriction on the depth or node number as base classifiers. You have to consider the same 50 values for the number of predictors as above.

Whenever you use randomization in your code, use the number formed of the last 4 digits of your student ID, as the seed for the pseudo number generator.

You have to write a report to present your results and their discussion. The report must include a plot with the test errors for the 5 ensemble methods versus the number of predictors used. Also include in the plot the test error for the decision tree (use the same value for each number of predictors; thus you will have a horizontal line). Discuss the results, in other words discuss how do the methods compare. Is one method better than another method for all numbers of predictors, or as the number of predictors becomes larger, or when it is smaller, etc.? How close or far are the methods in performance? ... etc. It is important to have a thorough comparison.

In the report you also have to include a plot of the cross-validation error obtained with the decision tree model for all values of the maximum number of leaves considered.

Besides the report, you have to submit your numpy code. The code has to be modular. Write a function for each of the main tasks. Also, write a function for each task that is executed multiple times. The code should include instructive comments.

To load the data set, first save the file 'spambase.data' in a subdirectory. Then you can use pandas.read_csv() to load the file using the following code (assuming that the file was stored in the subdirectory named 'Data'):

```
import pandas as pd
dataset = pd.read_csv('Data/spambase.data', header=None)
X = dataset.iloc[:, :-1].values
t = dataset.iloc[:, -1].values
```

Submission Instructions:

- Submit the report in pdf format, the python file (with extension ".py") containing your code, and a short demo video. The video should be 1 min or less. In the video, you should scroll down your code, show that it runs and that it outputs the results for each part of the assignment. Submit the files in the Assignments Box on Avenue.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd Ed., Springer, 2009 (ISBN 9780387848570), available for free download at https://web.stanford.edu/ hastie/ElemStatLearn/