ISyE 7406: Data Mining & Statistical Learning
HW#5

Ensemble Method. Apply **random forests** and **boosting** to a data set of your choice. If you want, you can choose a data set from the course (e.g., past lectures or homeworks including simulated data sets) or from R (e.g., the *ISLR* package) or other sources. The only exception is the spam email data set, since we have used it extensively in our lectures. It might be okay if you want to use the dataset from your proposal, esp. if it is a large, complicated dataset, but you can do so only if each group member works independently on the homework without collaboration and if all group members agree.

Please write a report to summarize your analysis, subject to the following requirements:

(a) Be sure to fit both **random forests** and **boosting** on a training set and to evaluate their performance on a test set.

(b) How accurate are the results compared to simple **baseline methods**? For instance, some candidate baseline methods can be KNN, linear regression, LDA, logistic regression, local smoothing, tree, etc., whichever are appropriate.

(c) Which of these approaches yields the best performance in term of smallest testing error?

(d) You need to explain how or why you choose certain tuning parameters in these approaches, based only on the training set. This can be done through either cross-validation of the training set, or variable selection such as AIC or BIC from the training set, or any other reasonable approaches.

(e) In your writeup, please follow the guideline of final course project. In particular, please provide necessary background on the data set of your choice, so that readers can understand your data set and analysis.

Remarks: the purpose of this homework is to prepare you for the course project and the final exam. If feasible, please use the final report format, and it is okay without the title page or reference or other non-essential materials. Also in the final exam, you are given a training set, and then are asked to predict on a testing set – your grade in the final exam will mainly be based on how small the testing error is.

Note that the use of cross-validation in this homework will be slightly different from those in HW#1 and HW#2, in the sense that here you should use the cross-validation only to the training set itself with the aim of helping you find the best set of tuning parameters.

There will be no universal solutions to this homework, as we expect that students will choose different kinds of data sets. As a result, the peer review comments and grading will depend heavily on your writeup/presentation and explanations including (1) what is your data set, (2) how you tune parameters in each approach, (3) whether your conclusions are appropriate based on your numerical comparisons of different approaches; and (4) whether your presentation is clear, e.g., whether it is easy to read to your report.

ISyE 7406: Data Mining & Statistical Learning
Optional HW (No credits, and Not Graded!)

This is an optional HW. **No credits and not graded.** It might help you better understand the tree-based method.

Tree-based Method. Consider the *Orange Juice (OJ)* dataset, which is part of the *ISLR* package in R. The data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice, and a number of characteristics of the customer and product are recorded.

(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

(b) Fit a classification tree with "gini" criterion to the training set, with the binary variable "*Purchase*" as the response and the other variables as predictors. Use the "summary()" function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?

(c) Create a plot of the tree and interpret the results.

(d) Predict the response on the test data, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
Hints: The confusion matrix between two vectors, say, Y and $Yhat$, can be obtained in R by the "table()" function, i.e., "table($Y, Yhat$)".

(e) Use the training set to determine the optimal tree size that corresponds to the lowest cross-validation classification error rate.

(f) Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. Note that cross-validation does not necessarily lead to selection of a pruned tree, and if so, then create a pruned tree with fewer number of terminal nodes.

(g) Compare the pruned and unpruned trees, in terms of both training and testing error rates. Which is better, and does it match your intuition?

Remarks: The following R code helps you to get the *OJ* data set:

```
## You need to first install the R package ISLR
library(ISLR)
data(OJ)
head(OJ)
?OJ
```