Assignment Administered by Team 17: Bagging and Random Forest

BUAD 5082 – Spring 2019

---

1. Objectives

The purpose of this assignment is to provide you with some experience working with the randomForest() functions and several of its supporting functions.

2. What You Will Need

Access to a Windows computer with R

3. Solutions

The solutions will be posted on March 2nd

4. Tasks

Problem 1: Random Forest Classification

A. a rm(list=ls()) and set the random seed to 5082.
B. Create a data frame for the 'ISLR' Smarket dataset.
C. Split the data into training, and test sets using 70/30 split using sample().
D. Grow a 500-trees ensembled by bagging with mtry = 3 on the training data to predict Direction using all the variables. Make sure importance = TRUE, replace = TRUE.
E. Display the importance of predicting variables, with descending order of importance on 'MeanDecreaseAccuracy'.
F. Plot a graph of variable importance in the graphics window. Be sure to include a Title.
G. Plot error rates for the Random Forest model using plot(rf).
H. Find the least grown tree with minimum OOB error rate.
I. Set seed = 5082 and rebuild the model with number of trees in H.
J. What is the OOB error rate of this most grown model?
K. Look at the confusion matrix of this model, what are the Type I and Type II error rates?
L. Using roc.area, find the AUC score of this model.
M. Plot the ROC curve, with legends and title.
N. What's special about this ROC curve? Why?
O. Do you want to change the cutoff for this problem? Why?
P. Predict the test set using this model and construct a confusion matrix.
Q. How does the model perform on the test set?
R. Any thoughts on this problem?

Problem 2: Random Forest Regression

A. rm(list=ls()) and set the random seed to 5082.
B. Create a data frame for the 'ISLR' Boston dataset.
C. Split the data into training, and test sets using 70/30 split using sample().
D. Grow a 500-trees ensembled by bagging with mtry = 'number of predictors' on the training data to predict medv using all the variables. Make sure importance = TRUE, replace = TRUE.
E. Display the importance of predicting variables, with descending order of importance on '%IncMSE'.
F. Plot a graph of variable importance in the graphics window. Be sure to include a Title.
G. Repeat D to F, instead, use mtry = round(sqrt(number of predictors)) to grow a 500-tree forest.
H. What are the differences between the two models?
   Hint: look at the variable importance.
I. Plot MSEs for the trees using plot(rf).
J. Find the least grown tree with minimum OOB MSE.
K. Set seed = 5082 and rebuild the model with number of trees in J.
L. What is the OOB MSE of this most grown model?
M. Predict the test set using this model.
N. What is the train MSE? What is the test MSE?
O. Which one is larger, OOB MSE, train MSE or test MSE? Why?