Spring 2020

Quantitative Analysis
Chung-Ming Kuan

Problem Set 11

Due: 6/1

Part One: Hand-Written Exercise

1. Please make a summary of decision tree, bagging, random forest and boosting.

Part Two: Computer Exercise

1. Load the `Boston` data set in `R` and create a new variable `High`, which is a binary response and equals "*yes*" when `medv` $>$ 22 and "*no*" otherwise. Please answer the following questions:

   (a) Let `High` be our variable of interest and all the other variables in the data set, except for `medv`, be our predictors.
   Please construct a `bagging` model with 500 trees, what is the variable that, on average, decrease the Gini index the most (hence most important) according to this model?

   (b) Let `High` be our variable of interest and all the other variables in the data set, except for `medv`, be our predictors.
   Please construct a `bagging` model with 500 trees and a `random forest` with 500 trees and m (number of variables that can be considered for each split) $= 3$. Plot the OOB error across different number of trees for these two models.

   (c) Let `medv` be our variable of interest and all the other 13 variables in the data set, except for `High`, be our predictors.
   Please construct a `boosting` model with $\lambda = 0.1$, and d (interaction depth) $= 1, 2, 3, 4$. Choose the best number of trees for each model, using 10-fold CV, ranging from 1 to 1000. The optimal number of trees should be different for each d.

   (d) Following (e), among the four models from (4.a) ($\lambda = 0.1$, $d = 1, ..., 4$ with corresponding optimal number of trees), which yield the smallest 10-fold CV error?