CS 5565, LAB7 (Tree-Based Methods) 60 Points

Name _____

1. View the videos at the following URLs
   https://www.youtube.com/watch?v=0wZUXtvAtDc
   https://www.youtube.com/watch?v=IY7oWGXb77o

   You may download the R Code for Labs and the Data Sets to use from the textbook website.
   http://www-bcf.usc.edu/~gareth/ISL/

2. (25 points total) In the lab, a classification tree was applied to the `Carseats` data set after converting `Sales` into a qualitative response variable. Now we will seek to predict `Sales` using regression trees and related approaches, treating the response as a quantitative variable.

   (a) (5 points) Split the data set into a training set and a test set.

   (b) (5 points) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

   (c) (5 points) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

   (d) (5 points) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important.

   (e) (5 points) Use random forests to analyze this data. What test MSE do you obtain? Use the `importance()` function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.

3. (35 points total) We now use boosting to predict `Salary` in the `Hitters` data set.

   (a) (5 points) Remove the observations for whom the salary information is unknown, and then log-transform the salaries.

   (b) (5 points) Create a training set consisting of the first 200 observations, and a test set consisting of the remaining observations.

   (c) (5 points) Perform boosting on the training set with 1,000 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding training set MSE on the y-axis.

   (d) (5 points) Produce a plot with different shrinkage values on the x-axis and the corresponding test set MSE on the y-axis.

   (e) (5 points) Compare the test MSE of boosting to the test MSE that results from applying two of the regression approaches seen in Chapters 3 and 6.

   (f) (5 points) Which variables appear to be the most important predictors in the boosted model?

   (g) (5 points) Now apply bagging to the training set. What is the test set MSE for this approach?