Homework 4

Due on 04/13/2022

1. In this exercise, we will build tree-based models using the `College` data (see "College.csv" in Homework 2). The response variable is the out-of-state tuition (`Outstate`). Partition the dataset into two parts: training data (80%) and test data (20%).

(a) Build a regression tree on the training data to predict the response. Create a plot of the tree.

(b) Perform random forest on the training data. Report the variable importance and the test error.

(c) Perform boosting on the training data. Report the variable importance and the test error.

2. This problem involves the `OJ` data in the `ISLR` package. The data contains 1070 purchases where the customers either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of customers and products are recorded. Create a training set containing a random sample of 700 observations, and a test set containing the remaining observations.

(a) Build a classification tree using the training data, with `Purchase` as the response and the other variables as predictors. Use cross-validation to determine the tree size and create a plot of the final tree. Which tree size corresponds to the lowest cross-validation error? Is this the same as the tree size obtained using the 1 SE rule?

(b) Perform boosting on the training data and report the variable importance. What is the test error rate?