DSA211 Statistical Learning with R

Homework 10

Use R functions and data file **OJ** in the **ISLR** package to solve the following problem:

1. Set the random seed to (103).

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

   (b) Fit a tree to the training data, with **Purchase** as the response and the other variables as predictors. What is the training error rate? How many terminal nodes does the tree have?

   (c) Create a plot of the tree.

   (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?

   (e) Apply the cross validation to the training set in order to determine the optimal tree size. What is the optimal tree size?

   (f) Produce a pruned tree to the training data with the corresponding optimal tree size obtained using cross-validation. What is the training error rate of this pruned tree?

   (g) Predict the response on the test data based on the pruned tree obtained in part (f), and produce a confusion matrix comparing the test labels to the predicted test labels based on the pruned tree. What is the test error rate of this pruned tree?

   (h) Produce the final pruned tree to all the data with the corresponding optimal tree size obtained using cross-validation. Plot the final pruned tree.

   (i) Use the random forests approach to analyze this data. Predict the response on the test data based on the random forests approach, and produce a confusion matrix comparing the test labels to the predicted test labels. What is the test error rate of the random forests approach? Use R-function to plot and determine which variable is the most important?

-END-