ST635 Homework 5

Spring 2016

Due: April 4 (Monday), 2016

An Exercise using Classification Tree

This problem involves the *OJ* data set which is part of the "ISLR" package. The data contains 1070 purchases where the customer either purchased Citrus Hill or Minute Maid Orange Juice. A number of characteristics of the customer and product are recorded.

1. Check out the **R** function *sample()*. Use the *sample()* function to create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.

2. Use *tree()* function to fit a tree on the training data set, with *Purchase* as the response and the other variables as predictors. Use the *summary()* function to produce summary statistics about the tree, and describe the results obtained. What is the training misclassification error rate? How many terminal nodes does the tree have?

3. Type in the name of the tree object used in part (2) in **R** to get a detailed text output of the fitted tree from part (2). Pick one of the terminal nodes, and interpret the information displayed in the output.

4. Create a plot of the tree obtained from part (2), and interpret the results.

5. Using the fitted tree in part (2), predict the response on the test data set. What is the test miscalssification error rate?

6. Apply *cv.tree()* function to the training set in order to determine the optimal tree size using ten-fold cross-validation. Produce a plot with tree size on the horizontal axis and cross-validated error score on the vertical axis. Which tree size corresponds to the lowest cross-validated error rate?

7. Produce a pruned tree corresponding to the optimal tree size obtained using cross-validation. If cross-validation does not lead to selection of a pruned tree, then create a pruned tree with five terminal nodes.

8. Compare the training error rates between the pruned and un-pruned trees. Which one is higher?

9. Compare the test error rates between the pruned and un-pruned trees. Which one is higher?