IOM 530: Applied Modern Statistical Learning Methods
Assignment 6 (Due 11/12/2013)

Guidelines for assignment submission:

1. Type each question before you answer it, and provide a clear separation between each part.
2. All relevant computer output should be provided unless noted otherwise.
3. Print your homework, and submit it at the beginning of the class. Make sure that it is stapled, and your name is typed on it.
4. Attach your R code as an Appendix. Make sure to provide comments on what your code is doing. Keep it clean and clear!
   associated with them. If you run into problems you may find it easier to work with the auto data.

5. You can refer to Section 8.3 (Lab: Decision Trees) in the course book and the facebook group for this class if you need some help with the R code necessary to finish this assignment.

This problem involves the OJ data set, which is part of the ISLR package.

(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 except for Buy as predictors. Use the summary() function to produce summary statistics about the tree, and describe the results obtained. What is the training error rate? How many terminal nodes does the tree have?
(c) Type in the name of the tree object (model) in order to get a detailed text output. Pick one of the nodes, and interpret the information displayed.
(d) Create a plot of the tree, and interpret the results.
(e) Predict the response on the test data, and produce the confusion matrix comparing the test labels to the predicted test labels. What is the test error rate?
(f) Apply the cv.tree() function to the training set in order to determine the optimal tree size.
(g) Produce a plot with tree size on the x-axis and cross-validated classification error rate on the y-axis.
(h) Which tree size corresponds to the lowest cross-validated classification error rate?
(i) 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.
(j) Compare the training error rates between the pruned and unpruned trees? Which is higher?
(k) Compare the test error rates between the pruned and unpruned trees? Which is higher?