

This problem involves the `OJ` data set which is part of the `ISLR2` 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 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 in order to get a detailed text output. Pick one of the terminal 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 a 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?