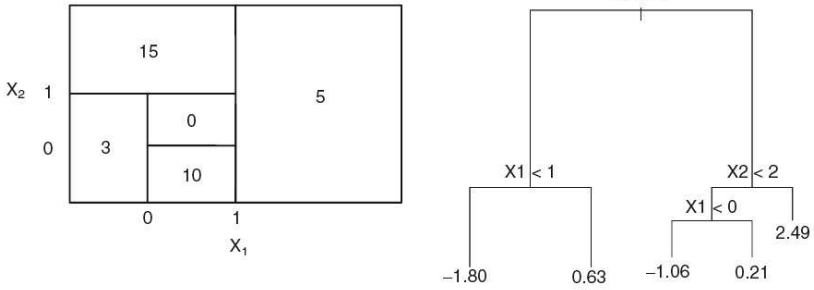
Stat 627/427 (Statistical Machine Learning)

Trees (Chap. 8)

1. (**Trees; Chap. 8, # 4, p. 332**) Look at the figure below.

   (a) Sketch the tree corresponding to the partition of the predictor space in the left-hand panel of the figure. The numbers inside the boxes indicate the mean of Y within each region.

   (b) Look at the tree in the right-hand panel of the same figure. Based on this tree, create a diagram similar to the left-hand panel of this figure. Divide the predictor space into the correct regions, and indicate the mean for each region.



2. (**Bagging; Chap. 8, # 5, p. 332**) Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $\boldsymbol{P}\{\text{Class is Red} \mid X\}$: 0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, and 0.75.

   There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

3. (**Project; Chap. 8, # 9, p. 334**)

   This problem involves the **OJ** (orange juice) data set which is part of the ISLR package. `library(ISLR); attach(OJ);` To find its description, type `?OJ`

   (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 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?