HW 05: Trees

Due Thursday, Nov. 10 at 11:59pm

Exercise 1

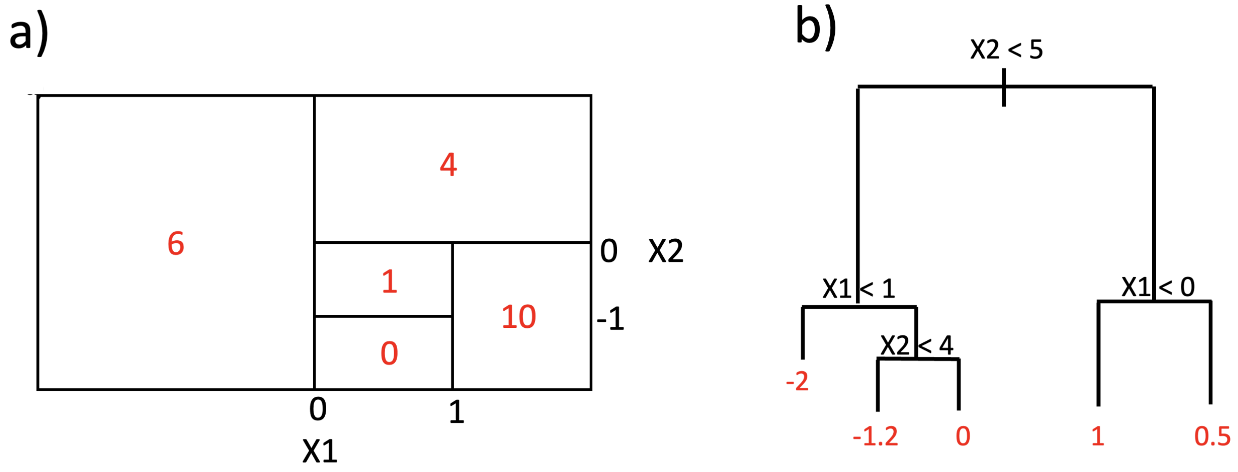

Figure 1: Figure 1

a) Sketch the tree corresponding to the partition of the predictor space illustrated in Figure 1a. The red numbers inside the boxes indicate the mean of the training Y within each region.

b) Create a diagram similar to Figure 1a using the tree illustrated in Figure 1b. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

c) Suppose I have new a observation with predictors $(X1 = 1.5, X2 = 4.5)$. What is your estimate $\hat{y}$ using the tree in Figure 1a? What about when using the tree in Figure 1b?

Exercise 2

Step through example of greedy approach for regression tree

Exercise 3

Step through example of greedy approach for classification tree using Gini and/or entropy

Exercise 4

Suppose we decide to perform bagging for classification. We have data on about mushrooms, the response has classes `toxic` and `non-toxic`. We fit $B = 10$ classification trees. For each bootstrapped sample and a specific value of X, we have the following 10 estimates of $\Pr(\text{toxic}|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in class. 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?

Exercise 5

a) How do you expect the random forest model to perform on data with correlated features relative to data without correlated features? Why?

b) Where does the "random" in random forests come from? That is, during which part(s) of the algorithm are we including randomization?