HW 05: Trees

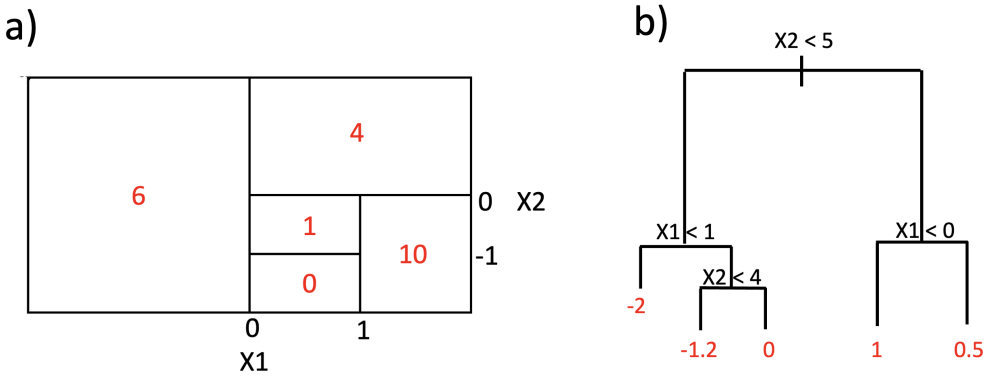
Due Thursday, Nov. 10 at 11:59pm

Exercise 1

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

b) Create a diagram similar to Figure a using the tree illustrated in Figure b. 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 a? What about when using the tree in Figure b?



Exercise 2

Suppose we decide to perform bagging for classification. We have data about mushrooms, where 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 3

a) How do you expect the random forest model to perform on data with correlated features relative to data without correlated features? Why? Consider the following aspects of the model: the prediction performance and the model interpretability.

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

Exercise 4

In this exercise, we will step through an example of the greedy approach for classification tree using Gini index. Suppose we are trying to classify a person's undergraduate `major` into either `"CS"`, `"Econ"`, or `"Math`", based on two predictors: their preferred programming `language` and their `salary` two years post-graduation (in $10,000s). Our data is as follows:

salary	language	major
12	R	Econ
15	Python	CS
9	Python	Math
17	Python	CS
11	R	Econ
10	Python	Math
13	R	Econ
12	Python	CS

We will perform recursive binary splitting using a greedy approach, where the splits are chosen based on which split yields the best node purity at each step. We will use the Gini index to define node purity:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

where $\hat{p}_{mk}$ represents the proportion of training observations in the m-th region that are from the kth response class. Region m corresponds to the segment of predictor space under the given split. At every step, we choose the split that yields the lowest impurity, averaged across the child nodes.

To decide the next split, we have candidate decisions. That is, for each predictor X_j, we can divide the predictor space into regions $\{X_j < s\}$ (left child) and $\{X_j \geq s\}$ (right child) for quantitative X_j. For categorical predictors, we segmend the predictor space into $\{X_j = a\}$ (left child) and $\{X_j \neq a\}$ (right child), where a is a category of X_j.

For a given split on predictor X_j, we calculate the Gini indices for both the left and right child nodes. Call these G_l and G_r, respectively. Then take a *weighted average* of the Gini indices to determine the overall quality of this particular split. The weighted average is $w_l G_l + w_r G_r$, where w_l is the proportion of observations in the left child node, and w_l is the proportion of observations in the right child node (note: for any given split, $w_l + w_r = 1$). We will choose the split that yields the *lowest weighted average.*

If a given node is completely pure (i.e. the Gini index of that node is 0), set it to be a terminal node.

For quantitative X_j, we typically consider a range of s values to segment the predictor and choose the s that yields the lowest impurity. For the purposes of this homework, we will not take this approach. Instead, let s be the median value in the current segment of the predictor space. If you have an even number of observations n, let the s be the average of the two middle values. Example: if the predictor values for X_j are `1, 4, 3, 2`, then `s = 2.5`. If the predictor values for X_j are `3, 2, 5, 4, 1`, then `s = 3`.

a) We will begin by constructing the root node. Calculate the the weighted Gini indices for each of the splits `salary < s` (where `s` is defined above) and `language = "Python`.

b) Based on your weighted Gini indices in (a), what is the first split (root node) in your decision tree?

c) Finish building your tree. Once you are finished, draw your resulting tree. Don't forget to add the predicted classes for each of the terminal nodes.