Spring 2020

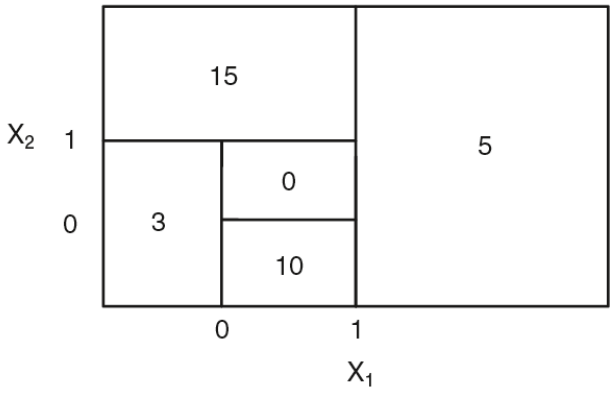
Quantitative Analysis
Chung-Ming Kuan

Problem Set 10

Due: 5/25

Part One: Hand-Written Exercise

1. Sketch the tree corresponding to the partition of the predictor space illustrated in the following figure. The numbers inside the boxes indicate the mean of Y within each region.



2. Suppose we have five equal sized data set containing red and green classes. We then apply a classification tree to each data set and, for a specific value of X, produce 5 estimates of $Pr(\text{Class is Red}|X)$: 0.1, 0.2, 0.55, 0.6, and 0.75. There are two common ways to "combine" these results together into a single class prediction. One is the majority vote approach. 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?

Part Two: Computer Exercise

1. Consider the Gini index, classification error, and cross-entropy in a simple classification setting with two classes. Please use `R` to create a single plot that displays each of these quantities as a function of $\hat{p}_{m1}$. The x-axis should display $\hat{p}_{m1}$, ranging from 0 to 1, and the y-axis should display the value of the Gini index, classification error, and entropy.

2. Load the `Boston` data set in `R` and create a new variable `High`, which is a binary response and equals "*yes*" when `medv` $>$ 22 and "*no*" otherwise. Please answer the following questions:

(a) Let `medv` be our variable of interest and all the other 13 variables in the data set, except for `High`, be our predictors.
Please fit a `regression tree` that has the optimal number of terminal nodes, chosen by 10-fold CV.

(b) Let `High` be our variable of interest and all the other variables in the data set, except for `medv`, be our predictors.
Please fit a `classification tree` that has the optimal number of terminal nodes, chosen by 10-fold CV. (Use Gini index to guide the tree growing process, while using the misclassification error to guide the pruning process)