Fin7047
Spring 2023

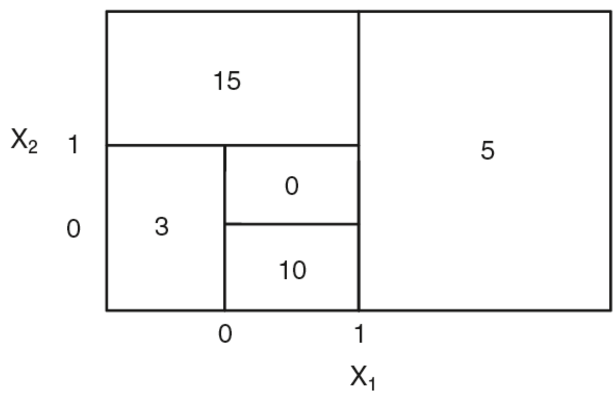
Quantitative Analysis
Chung-Ming Kuan

Problem Set 10

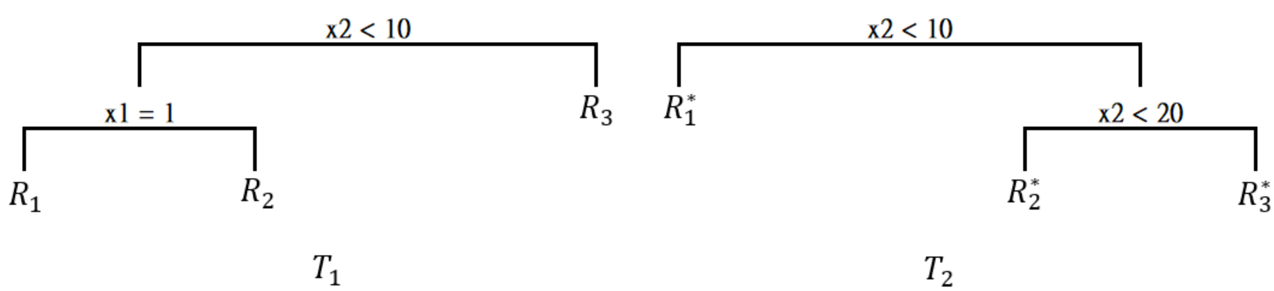
Due: 5/29

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. For each node, the predictors X_j and the cutpoint s split the predictor space into the regions $\{X|X_j \leq s\}$ and $\{X|X_j > s\}$, $j = 1, 2$.



2. Consider a 3-class classification problem for a sample of 900 observations, with 300 in each class. A node j with a, b, and c observations belonging to, respectively, calss I, II and III is denoted as $R_j(a, b, c)$. Suppose a tree T_1 contains three terminal nodes $R_1(200, 50, 50)$, $R_2(50, 200, 50)$ and $R_3(50, 50, 200)$. While another tree T_2 yields three alternative terminal nodes $R_1^*(100, 0, 0)$, $R_2^*(50, 250, 100)$ and $R_3^*(150, 50, 200)$.



   (a) If these two trees were our only options, which one would be chosen if we utilize classification error rate to guide the tree growing process?

(b) Continue with part (a), which tree would be chosen if we utilize Gini index to guide the tree growing process?

(c) Suppose we have a testing set with five observaitons as below. What's the probability that a random observation from this testing set is *correctly* classified with the tree you choose in part (a)?

	x_1	x_2	class
$obs.1$	1	20	III
$obs.2$	0	5	II
$obs.3$	1	4	I
$obs.4$	1	59	I
$obs.5$	0	0	III

3. A data set with five observations is shown below. Let y be the variable of interest and x_1 and x_2 be predictors (Note that y of $obs.5$ is unknown). A student wants to use boosting on this data set with the following settings:

- The number of trees, B, is equal to 2.
- The shrinkage parameter, λ, is equal to 0.6.
- The number of splits in each tree, d, is equal to 1.
- The boosting makes a prediction via *averaging* at each nodes.

Suppose we've known that the splits are $x_1 \leq 4.8$ and $x_2 = 1$, respectively, for the two trees. Please calculate the prediction of the boosting for $obs.3$. (Hint: Your answer should include a.)

	x_1	x_2	y
$obs.1$	2	10	6
$obs.2$	3	1	9
$obs.3$	4	1	12
$obs.4$	10	1	4
$obs.5$	10	10	a

Part Two: Computer Exercise

1. Load the `Boston` data set in `R` and answer the following questions with `set.seed(1)`.
   Let `medv` be our variable of interest and all the other 13 variables in the data set be our predictors.

   (a) Please fit a regression tree that has the optimal number of terminal nodes, chosen by 100-fold CV, and plot the tree.

   (b) Suppose we are trying to fit this data set using a boosting model. For $\lambda = 0.1$, and d (interaction depth) $= 1, 2, 3, 4$, please choose the best number of trees for each model, using 10-fold CV, ranging from 1 to 1000.

   (c) Among the four models from (b) ($\lambda = 0.1$, $d = 1, ..., 4$ with corresponding optimal number of trees), which one yields the smallest 10-fold CV error?

2. Please load the `heart` data set from the package `kmed`.
   Let `class` be our variable of interest and set it to 0 if it's 0 and 1 otherwise (that is, `class` only has 2 values, 0 and 1), and all the other 13 variables in the data set be our predictors. Some data processing has been done in your answer sheet.

   (a) Grow a classification tree with the gini index as the splitting criterion and then determine the optimal node for this tree using 10-fold CV with `set.seed(1)`.

   (b) Plot the OOB errors of random forest with m being 2, 4, 8 and 13, respectively, based on the `heart` data with `set.seed(1)`.