UNIVERSITY OF TEXAS AT AUSTIN

Homework Assignment 9

Regression trees.

Please, provide your **complete solutions** to the following problems. Final answers only, even if correct will earn zero points for those problems.

Problem 9.1. (10 points) Solve Problem **8.4.1** from page 361 from the textbook.

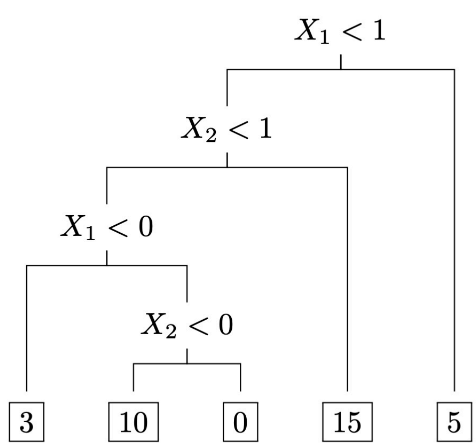
Solution: Solutions will vary.

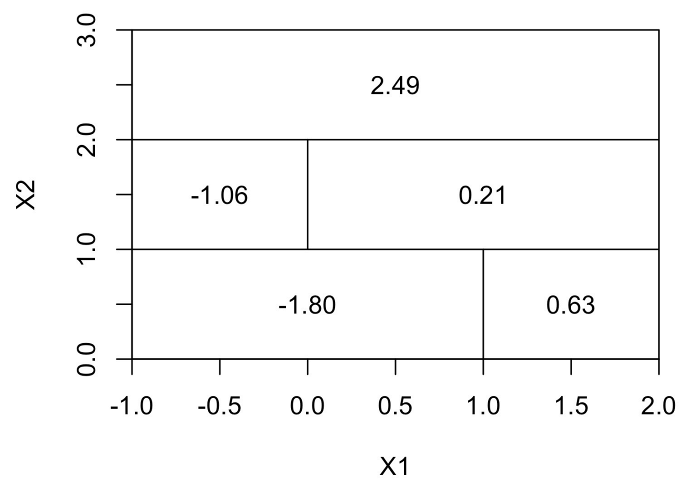
Problem 9.2. (5 points) Draw an example of a partition in the plane that **cannot possibly** correspond to recursive binary splitting.

Solution: Solutions will vary.

Problem 9.3. (10 points) Solve Problem **8.4.4** from page 362 from the textbook.

Solution: The tree and the partition should look something like this (up to various symmetries):





Problem 9.4. (10 points) *Source: An old SRM manual.*
Consider the following observations of (X, Y) with X being the predictor and Y being the response:

$$(0,8), \quad (1,5), \quad (3,8), \quad (6,6).$$

After one iteration of recursive binary splitting, there are two groups of observations. Find the members of the two groups.

Solution: Remember that - in general - the criterion for choosing the splits is to minimize the residual sum of squares (RSS)

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

where $\hat{y}_{R_j}$ stands for the average of the response variable in region R_j for $j = 1, \dots, J$. However, since this problem is computationally too complex, we resort to **recursive binary splitting**. Hence, as there is one predictor only in our current problem, we must make the split along its possible values. Every available split is **binary** and partitions the support of X into R and R^c. So, it creates an RSS with this structure

$$\sum_{i\in R}(y_i-\hat{y}_R)^2+\sum_{i\in R^c}(y_i-\hat{y}_{R^c})^2.$$

In this problem, we can now proceed "by hand" from the lowest to the highest observed value of the predictor.

If $(0,8)$ is the sole element in R, the mean response for the remaining points is

$$\frac{5+8+6}{3}=\frac{19}{3}.$$

The RSS is

$$\left(5-\frac{19}{3}\right)^2+\left(8-\frac{19}{3}\right)^2+\left(6-\frac{19}{3}\right)^2=\frac{14}{3}.$$

If $(0,8)$ and $(1,5)$ form the first region, the mean response in that region is $\frac{13}{2}$. The remaining points $(3,8)$ and $(6,6)$ are in the other region and their 7. So, the RSS is

$$\left(8-\frac{13}{2}\right)^2+\left(5-\frac{13}{2}\right)^2+(8-7)^2+(6-7)^2=\frac{13}{2}.$$

If $(0,8),(1,5)$, and $(3,8)$ are in the first region and only $(6,6)$ remains in the other region, then the average of the first region's values of the response variable is

$$\frac{8+5+8}{3}=7.$$

So, the RSS equals

$$(8-7)^2+(5-7)^2+(8-7)^2=6.$$

Overall, the smallest RSS corresponds to the first partition with $(0,8)$ in its own region, and the remaining points in the other region.

Problem 9.5. (15 points) *Source: Sample MAS-II.*
A data set contains six observations for two predictor variables, X_1 and X_2, and a response variable Y. Here is the table of observations:

X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

The following five splits are analyzed:

I. $R_1(1,1)=\{X\,|\,X_1<1\}$ and $R_2(1,1)=\{X\,|\,X_1\geq 1\}$
II. $R_1(1,4)=\{X\,|\,X_1<4\}$ and $R_2(1,4)=\{X\,|\,X_1\geq 4\}$
III. $R_1(2,0)=\{X\,|\,X_2<0\}$ and $R_2(2,0)=\{X\,|\,X_2\geq 0\}$
IV. $R_1(2,1)=\{X\,|\,X_2<1\}$ and $R_2(2,1)=\{X\,|\,X_2\geq 1\}$
V. $R_1(2,2)=\{X\,|\,X_2<2\}$ and $R_2(2,1)=\{X\,|\,X_2\geq 2\}$

Determine which split is chosen first.

Solution: First note that **I.** and **III.** do not constitute a meaningful partition of the predictor space (since all observations end up in R_2). Now, let's focus on the other proposed binary splits individually

In **II.**, the only point in R_2 is $(4,1,3.0)$. All the remaining points are in R_1. The mean of the observations in R_1 is

$$\frac{1.2+2.1+1.5+2.0+1.6}{5}=1.68.$$

So, the RSS equals

$$(1.2 - 1.68)^2 + (2.1 - 1.68)^2 + (1.5 - 1.68)^2 + (2.0 - 1.68)^2 + (1.6 - 1.68)^2 = 0.548.$$

In **IV.**, $(1, 0, 1.2)$ is the only point in R_1. The mean of the response in the remaining observations is

$$\frac{2.1 + 1.5 + 3.0 + 2.0 + 1.6}{5} = 2.04$$

So, the total RSS equals

$$(2.1 - 2.04)^2 + (1.5 - 2.04)^2 + (3.0 - 2.04)^2 + (2.0 - 2.04)^2 + (1.6 - 2.04)^2 = 1.412.$$

In **V.**, $(3, 2, 1.5)$ and $(2, 2, 2.0)$ are placed into R_2 while the remaining points end up in R_1. The mean of the response values in R_2 is 1.75. So, the contribution to the RSS from R_2 equals

$$(1.5 - 1.75)^2 + (2.0 - 1.75)^2 = 0.125$$

The mean of the response values in R_1 is

$$\frac{1.2 + 2.1 + 3.0 + 1.6}{4} = 1.975.$$

So, the contribution to the RSS from R_1 equals

$$(1.2 - 1.975)^2 + (2.1 - 1.975)^2 + (3.0 - 1.975)^2 + (1.6 - 1.975)^2 = 1.8075$$

The total RSS is 1.9325.

Since its RSS is the smallest of the ones proposed, **II.** is our final answer.