Regression Trees

Regression Trees

- The tree will search for all combination of predictors and cutoff value to decide the best split
- In Regression tree, the best split is the split that minimizes

$$\underbrace{\sum_{i:\mathbf{x}_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2}_{\text{RSS of obs. in left branch}} + \underbrace{\sum_{i:\mathbf{x}_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2}_{\text{RSS of obs. in right branch}}$$

- $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ are the means of the responses falling in to the left branch and right branch, respectively.

Example

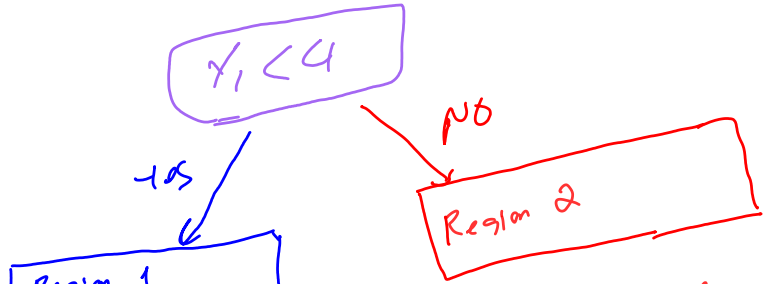
X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

Using the RSS to decide the best split among

- Split 1: Region 1 $X_1 < 4$, Region 2 $X_1 \geq 4$
- Split 2: Region 1 $X_2 < 2$, Region 2 $X_2 \geq 2$

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

Split 1



$Y = 1.2$
2.1
1.5
2.0
1.6

$Y = 3.0$.

$\Rightarrow \bar{Y}_2 = 3.0$

$\Rightarrow RSS_2 = (3.0 - \bar{Y}_2)^2$

$= 0$

$$\bar{Y}_1 = \frac{1.2 + 2.1 + 1.5 + 2.0 + 1.6}{5} = 1.68$$

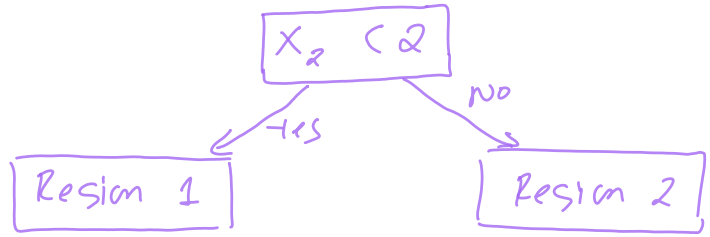
$$RSS_1 = (1.2 - \bar{Y}_1)^2 + (2.1 - \bar{Y}_1)^2 + (1.5 - \bar{Y}_1)^2 + (2.0 - \bar{Y}_1)^2 + (1.6 - \bar{Y})^2$$

$$= \boxed{.548}$$

The RSS of split 1 $= RSS_1 + RSS_2 = .548$

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

Split 2 :



$Y = 1.2, 2.1, 3.0, 1.6$

$\Rightarrow \bar{Y}_1 = \frac{1.2 + 2.1 + 3.0 + 1.6}{4}$

$= 1.975$

$Y = 1.5, 2.0$

$\Rightarrow \bar{Y}_2 = \frac{1.5 + 2.0}{2} = 1.75$

$RSS_1 = (1.2 - \bar{Y}_1)^2 + (2.1 - \bar{Y}_1)^2 + (3.0 - \bar{Y}_1)^2 + (1.6 - \bar{Y}_1)^2$

$= 1.8075$

$RSS_2 = (1.5 - \bar{Y}_2)^2 + (2 - \bar{Y}_2)^2$

$= 0.125$

$\Rightarrow$ RSS of split 2 $= RSS_1 + RSS_2$

$= 1.8075 + 0.125 = 1.9325$

The split with the smaller RSS is the better split. Thus, split 1 is better than split 2.

Suppose that your regression tree contain only one split which is the best split in the previous question. Calculate the R^2 of this regression tree on the training data.

we have:

$$R^2 = 1 - \frac{\text{RSS of the tree}}{\text{RSS of the baseline model } (\bar{Y})} = 1 - \frac{RSS}{RSS_0}$$

* RSS of the tree is the RSS of the split 1 or .548

* RSS of the baseline model $(\bar{Y})$: $\sum (Y - \bar{Y})^2$

$$\bar{Y} = \frac{1.2 + 2.1 + 1.5 + 3.0 + 2.0 + 1.6}{6} = 1.9$$

$$RSS_0 = (1.2 - \bar{Y})^2 + (2.1 - \bar{Y})^2 + (1.5 - \bar{Y})^2 + (3 - \bar{Y})^2 + (2 - \bar{Y})^2 + (1.6 - \bar{Y})^2$$

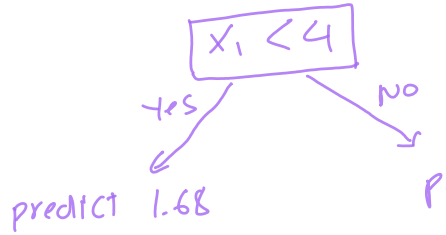
$$= 2$$

$$\Rightarrow R^2 = 1 - \frac{.548}{2} = 0.726$$

Use your regression tree to predict the y for the below testing data.
Calculate the R^2 of the tree on the testing data.

x_1	x_2	y
3	1	3.0
1	5	3.6
5	1	4.0
5	2	3.9

* The tree :



* The prediction

x_1	x_2	y	$\hat{y}$
3	1	3.0	3.0
1	5	3.6	3.0
5	1	4.0	1.68
5	2	3.9	1.68

* $R^2 = 1 - \dfrac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$

$\sum (y - \hat{y})^2 = (3-3)^2 + (3.6-3)^2 + (4-1.68)^2 + (3.9-1.68)^2$
$= 10.6708$

$\sum (y - \bar{y})^2 = 0.6075$

$\Rightarrow R^2 = 1 - \dfrac{10.6708}{0.6075} = -16.5651$