

Regression Trees

Example

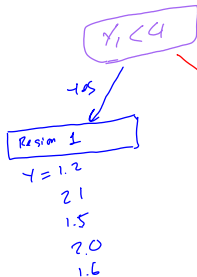
X_1	X_2	Y
1	0	1.2
2	1	2.1
3	2	1.5
4	1	3.0
2	2	2.0
1	1	1.6

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



$$Y = 3.0$$

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

$$\Rightarrow RSS_2 = (3.0 - \bar{Y}_2)^2 = 0$$

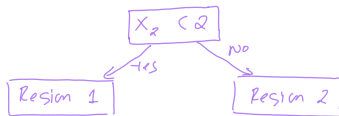
$$\bar{Y}_1 = \frac{1.2 + 2.1 + 1.5 + 2.0 + 1.6}{5} = 1.68$$

$$RSS_1 = (1.2 - \bar{Y}_1)^2 + (2.1 - \bar{Y}_1)^2 + (1.5 - \bar{Y}_1)^2 + (2.0 - \bar{Y}_1)^2 + (1.6 - \bar{Y}_1)^2 = .548$$

$$\text{The RSS of split 1} = RSS_1 + RSS_2 = .548$$

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 :



$$Y = 1.2, 2.1, 3.0, 1.6$$

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

$$= 1.975$$

$$\begin{aligned} RSS_1 &= (1.2 - \bar{Y}_1)^2 + (2.1 - \bar{Y}_1)^2 + (3.0 - \bar{Y}_1)^2 \\ &\quad + (1.6 - \bar{Y}_1)^2 \\ &= 1.8075 \end{aligned}$$

$$Y = 1.5, 2.0$$

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

$$\begin{aligned} RSS_2 &= (1.5 - \bar{Y}_2)^2 + (2.0 - \bar{Y}_2)^2 \\ &= 0.125 \end{aligned}$$

$$\begin{aligned} \Rightarrow \text{RSS of split 2} &= RSS_1 + RSS_2 \\ &= 1.8075 + 0.125 = 1.9325 \end{aligned}$$

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

Suppose that your regression tree contains 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:

$$R^2 = 1 - \frac{\text{RSS of the tree}}{\text{RSS of the baseline model } (\bar{y})} = 1 - \frac{\text{RSS}}{\text{RSS}_0}$$

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

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

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

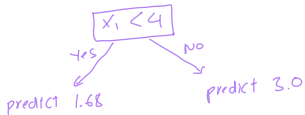
$$\begin{aligned} \text{RSS}_0 &= (1.2 - \bar{y})^2 + (2.1 - \bar{y})^2 + (1.5 - \bar{y})^2 + (3.0 - \bar{y})^2 + (2.0 - \bar{y})^2 + (1.6 - \bar{y})^2 \\ &= 2 \end{aligned}$$

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

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	1.68
1	5	3.6	1.68
5	1	4.0	3.0
5	2	3.9	3.0

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

$$\begin{aligned} \sum (y - \hat{y})^2 &= (3 - 1.68)^2 + (3.6 - 1.68)^2 + (4 - 3)^2 + (3.9 - 3)^2 \\ &= 7.238 \end{aligned}$$

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

$$\Rightarrow R^2 = 1 - \frac{7.238}{0.6075} = -10.914$$