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

continuous

	X_1	X_2	Y
A	1	0	1.2
B	2	1	2.1
C	3	2	1.5
D	4	1	3.0
E	2	2	2.0
F	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

$X_1 < 4$

yes: $Y = 1.2$, 2.1, 1.5, 2.0, 1.6

no: $Y = 3.0 \rightarrow \bar{Y}_2 = 3.0$

$\Rightarrow RSS_2 = 0$

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

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

$$= \boxed{.548}$$

$$RSS = RSS_1 + RSS_2 = \boxed{.548}$$

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

$X_2 < 2$ $\qquad$ $X_2 \geq 2$

$Y = 1.2$, 2.1, 3.0, 1.6 $\qquad$ $Y = 1.5$, $Y = 2.0$

$$\bar{Y}_1 = 1.975$$

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

$$RSS_1 = \sum (Y_i - \bar{Y}_1)^2 = 1.8075$$

$$RSS_2 = (1.5 - 1.75)^2 + (2 - 1.75)^2 = 0.125$$

$$\text{Total } RSS = RSS_1 + RSS_2 = \boxed{1.9325} > RSS \text{ of split 1} \Rightarrow \boxed{\text{Split 1 is better.}}$$

Split 1

$\boxed{x_1 < 4}$

yes

$y = 1.2$, 2.1, 1.5, 2.0, 1.6

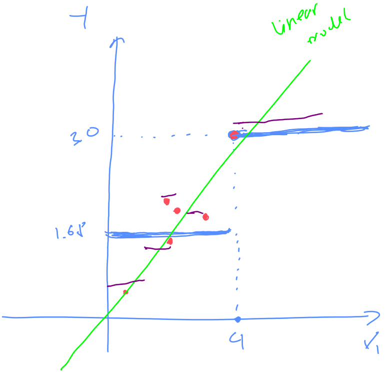
predict by the mean $\bar{y}_1 = 1.68$

no

$y = 3.0 \rightarrow \bar{y}_2 = 3.0$

predict by the mean

$\bar{y}_2 = 3.0$



For example, if $x_1 = 0$, $x_2 = 100$, then this tree will predict $y = 1.68$

Model Evaluation

Classification: Misclassification, ROC, Sensitivity ....

Regression:

$R^2 =$

$RSS = \sum (y_i - \hat{y}_i)^2$

$\sum |y_i - \hat{y}_i|$

$MAE = \sum \frac{|y_i - \hat{y}_i|}{n}$

(mean absolute error)

....

True value	predicted value
y_1	$\hat{y}_1$
y_2	$\hat{y}_2$
⋮	⋮

Linear model and regression model

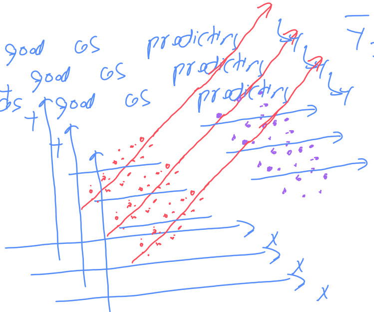
$$R^2 = 1 - \frac{RSS}{\text{Total SS}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

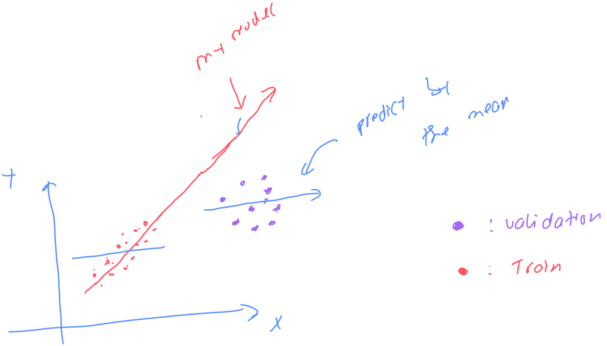
* $R^2 = 1 \iff RSS = 0$

* $R^2 = 0 \iff RSS = TSS$

The model is as good as predicting $\bar{Y}$
The model is as good as predicting $\bar{Y}$
The model is as good as predicting $\bar{Y}$

* $R^2 < 0$ (?)





In training

$R^2 > 0$

In validation :

$R^2 < 0$

classify 0 1 : 2%

predict 1 → 0

regression 2.5

$R^2 = -100$