ML Algorithms: Regression

Supervised Learning: Regression, Classification

Unsupervised Learning: Clustering

Linear Regression:

Predicting a real number as outcome. E.g -

- We have a a few (x,y) values.

X	Y
0	2
1	3
2	5
3	4
4	6

- X – independent variable and Y – dependent variable
- Consider a line equation: y = ax+b

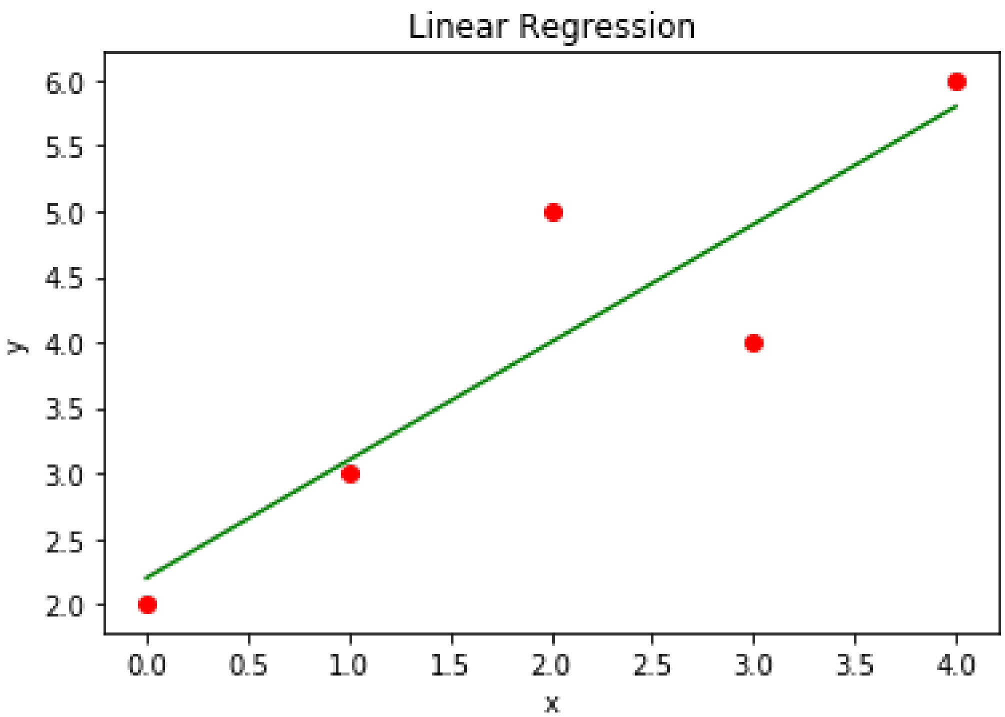
$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

-
- Thus, the equation **: y' = 0.9x+2.2. : Regression Equation**
- Now, using this equation we can find the values of y':

X	Y
0	2.2
1	3.1
2	4
3	4.1
4	5.8

- We can also predict the value of y from the regression equation, given the value of x.
- The green Line is the line y'=0.9x+2.2. The red dots are the actual values.



- <u>**Linear Regression Performance Metrics/Cost Function/Error Estimation :**</u>
  1. **Error:** Actual value – predicted value = y – y'
  2. **Total error:** Σ(y-y')
  3. **Mean error:** 1/n * Σ(y-y')
  4. **Mean squared error:** 1/n * $\Sigma(y-y')^2$
  5. **Mean absolute error:** 1/n * |Σ(y-y')|
  6. **RMSE(Root Mean Square Error):** $\sqrt{\frac{\Sigma(y-y')^2}{n}}$
  7. **R square value (R^2):** $1-\frac{\Sigma(y_i-y_i')^2}{\Sigma(yi-\bar{y})^2}$

Variability between actual and predicted value.

Total variability in y.

R^2 : Lies between 0-1

When its value to close to 0: It indicates poor fit.

Values close to 1: Indicates a good fit.(However can not adequately conclude so)

Adjusted R^2 : $1 - \frac{\Sigma\left(y_i - y_i'\right)^2/(n-p-1)}{\Sigma(yi-\bar{y})^2/(n-1)}$, where p: no, of columns.

- When there is 1 independent variable: Simple linear regression
- When there are more than 1 independent variables: Multi Linear Regression.