Simple Linear Regression

ML Type

① supervised learning
i> Regression
ii> classification

② Unsupervised Learning
i> clustering
ii> Association

③ Reinforcement leaving

Introduction

When we study ML, we first learn linear Regression because

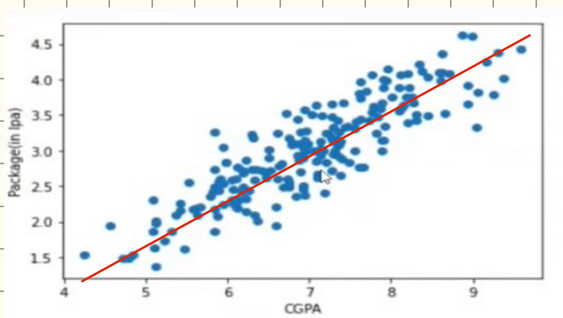
i> it is easy to understand and
ii> learning it makes other algorithm easier to understand later

⇒ Simple linear Regression, there is one input and one output.

Cgpa	Package
7.8	17.8

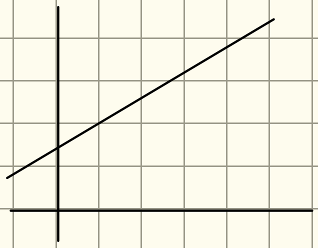
Simple linear regression

Cgpa	Package
7.8	17.8
8.1	20.1
7.5	25.5
8.9	16.5



LR :- First, we need to check if the data is linear or sort of linear (real world data). In 99% of cases, the data we get is sort of linear. We just need to draw a line on that sort of linear data. This is called linear regression. We have to draw the line in such a way that the mistake' are minimum.

Intuition



$y = mx + b$

$$\boxed{package = m \times cgpa + b}$$

How to find m and b

$y = mx + b$

↓

(m, b)

nikaalne ke tarike

① closed form solution

↓

direct formula

$\boxed{OLS}$

② non-closed form

↳ gradient descent

Direct formula

$$\boxed{b = \bar{y} - m\bar{x}}$$

y → package

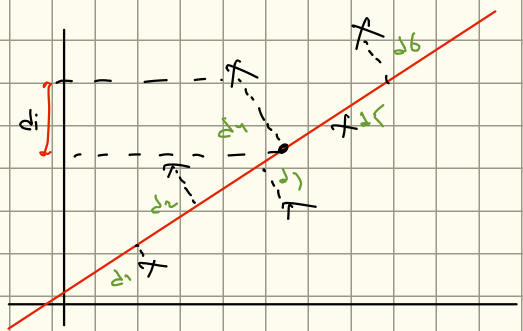
x → cgpa

$$\boxed{m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$\bar{y}$, $\bar{x}$ → these are mean value

x_i, y_i — these are current row cgpa & package

Scratch se derived formula



Error

↓

$E = d_1 + d_2 + d_3 - - - d_n$

$E = d_1^2 + d_2^2 + - - - d_n^2$

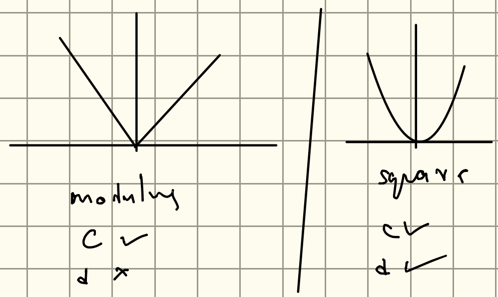
$$\boxed{E = \sum_{i=1}^{n} d_i^2}$$ → Error function / Loss function

Ham aisa (m,b) chahiye jise E minimise ho jaye

$E = |d_1| + |d_2| - - - |d_n|$

(kyu nahi kar rahe hai

We are squaring the distance instead of taking the modulus because later we have to do differentiation. We know that the graph of modulus is continuous but not differentiation



$$d_i = (y_i - \hat{y}_i)$$

y_i ↓ actual data and package

$\hat{y}_i$ ↳ mera model ya package

$$\boxed{E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

↓ we need value of m and b such that when we put them in this formula, the formula become minimum.

$$\boxed{\hat{y}_i = mx_i + b}$$

$$\boxed{E(m, b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2}$$

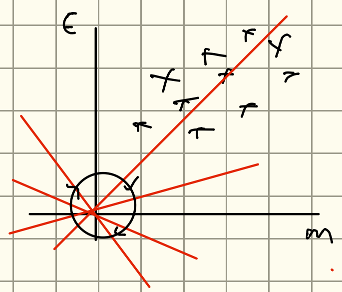
*

$$\boxed{E(m, b) = \sum_{i=1}^{n} (y_i - mx_i - b)^2}$$ minimum

(m, b)

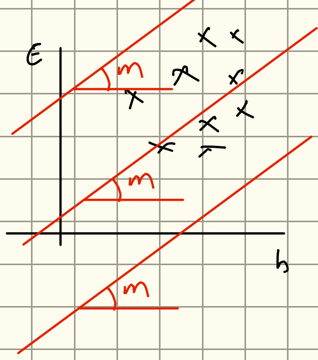
$\boxed{b = 0}$

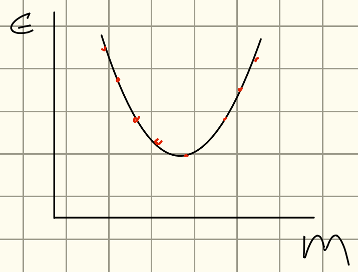
$$E(m) = \sum_{i=1}^{n} (y_i - mx_i)^2$$

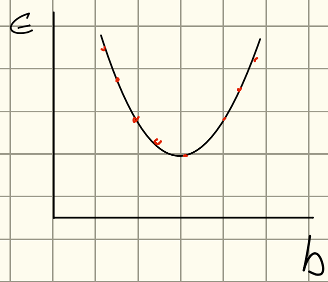


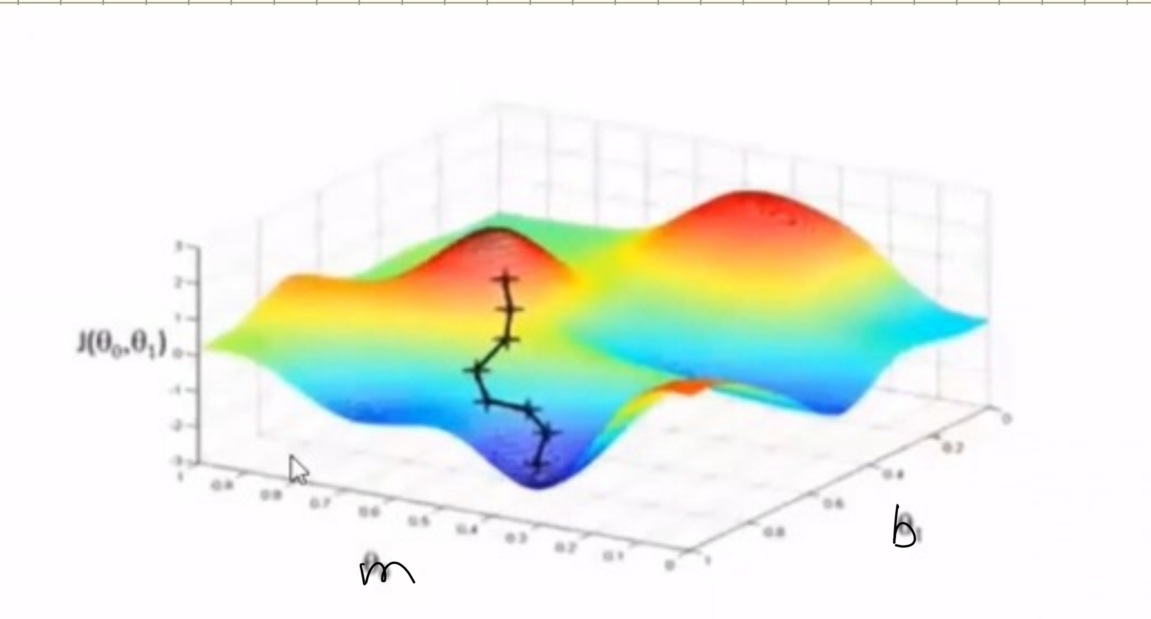
$\boxed{m = 1}$

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









$$\boxed{E(m,b) = \sum_{i=1}^{n} (y_i - m x_i - b)^2}$$

$$\frac{\partial E}{\partial b} = \frac{d}{db} \sum_{i=1}^{n} (y_i - m x_i - b)^2 = 0$$

$$= \sum \frac{d}{db} (y_i - m x_i - b)^2 = 0$$

$$= \sum -2 (y_i - m x_i - b) = 0$$

$$= \sum (y_i - m x_i - b) = 0$$

$$\frac{\sum y_i}{n} - \frac{\sum m x_i}{n} - \frac{\sum b}{n} = \frac{0}{n}$$

$$\downarrow$$

$$\overline{y_i} - m\overline{x_i} - \frac{\cancel{n} b}{\cancel{n}} = 0$$

$$\overline{y_i} - m\overline{x_i} = b$$

$$\boxed{b = \overline{y} - m\overline{x_i}}$$

$$E = \sum (y_i - m x_i - \overline{y} + m\overline{x})^2$$

$$\frac{\partial E}{\partial m} = \sum \frac{d}{dm} (y_i - m x_i - \overline{y} + m\overline{x})^2 = 0$$

$$= \sum 2 (y_i - m x_i - \overline{y} + m\overline{x})(-x_i - \overline{x}) =$$

$$= \sum -2 (y_i - m x_i - \overline{y} - m\overline{x_i})(x_i - \overline{x}) = 0$$

$$= \sum (y_i - m x_i - \overline{y} + m\overline{x})(x_i - \overline{x}) = 0$$

$$= \sum \left[(y_i - \overline{y}) - m(x_i - \overline{x})\right](x_i - \overline{x}) = 0$$

$$= \sum \left[(y_i - \overline{y})(x_i - \overline{x}) - m(x_i - \overline{x})^2\right] = 0$$

$$= \sum (y_i - \overline{y})(x_i - \overline{x}) = m \sum (x_i - \overline{x})^2$$

$$m = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Regression Metrics

① MAE (mean absolute error)

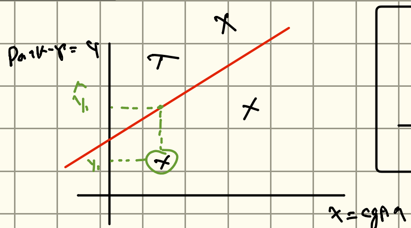
② MSE (mean square error)

③ RMSE (Root mean square error)

④ R2 Score

⑤ Adjusted R2 error.

MAE



$$\frac{|y_1-\hat{y}_1| + |y_2+\hat{y}_2| + \dots + |y_n-\hat{y}_n|}{n}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{n}$$

① Advantage

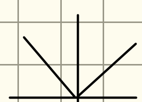
i) When we calculate MAE, we get a number called the loss. The unit of this number is the same as the output y. for example, if x is CGPA and Y is package (LPA), then MAE will also be in LPA.

Because the unit of MAE and the output are the same, it becomes easy to communicate.

ii) Robust outliers.

② Disadvantage

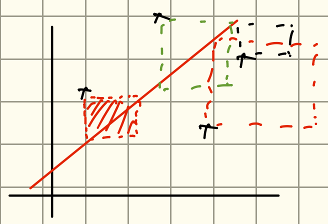
i) MAE me modulus function aa jata hai



→ these graph is not differentiable at 0, this is biggest problem jiski wajah se MSA ko aana padha.

↳ iss problem ko solve karta hai MSE

MSE



$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

Advantage

① we can use MSE as a loss function because it is differentiable

Disadvantage

① When we get the MSE value, its unit is $(LPA)^2$, while the output unit is just LPA.

This make it difficult to interpret.

② Robust to outliers.

RMSE

$$RMSE = \sqrt{MSE}$$

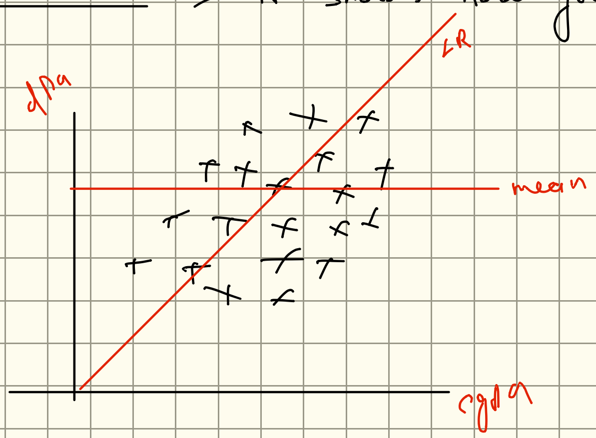
⇒ these all properties is similar to MSE

⇒ The only benefit is that its unit is the same as the output unit.

rmse – LPA
Y → LPA

R2 Score

→ it shows how good the model is performing



⇒ when we take admission in a collage and have data of 100 students, we may not have their cgpa, only their package. if someone asks, "What will be my package?" we can tell the mean of the packages.

But if we also have cgpa, we can use it and apply linear regression.

$$R2 = 1 - \frac{SS_R}{SS_m} = 1 - \frac{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_R}{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_m}$$

The R2 score tells us how much better the linear regression line is compared to the mean line.

This is R2 Score

SS_R = sum of squared error in the Reg line

SS_m = sum of squared error in the mean line

When R2 = 0

it mean the regression line is making the same amount of error as the mean line

$$R2 = \boxed{1 - \frac{\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]_R}{\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]_m}} = 1$$

When R2 = 1

it means the regression line is making no error all all.

$$\boxed{1 - \frac{\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]_R}{\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]_m}} \Longrightarrow 0$$

=> R2 score can also be negative when the regression line makes more errors than the mean line.

$$\boxed{R2 = 1 - \boxed{\frac{SS_R}{SS_m}}} \rightarrow \geq 1$$

$$\boxed{SSR > SSm}$$

Adjusted R2 Score

=> when we add input columns, the R2 score starts to increase.

cgpa | pac => R2 score = .80

iq | cgpa | pac => R2 = .90

=> There is no problem with this behavior.

The problem comes when we add irrelevant columns, like temperature - for example, the temp on the day of the interview.

Even for such inputs, the R2 may increase or stay the same, but it should have decreased.

To handle this, we use adjusted R^2 score.

$$\boxed{R^2_{adj} = 1 - \left[\frac{(1-R2)(n-1)}{(n-1-K)}\right]}$$

n = no. of rows

K = no of independent columns.