Page No.
Date

Lecture - 49
Simple Linear Regression - (SLR)

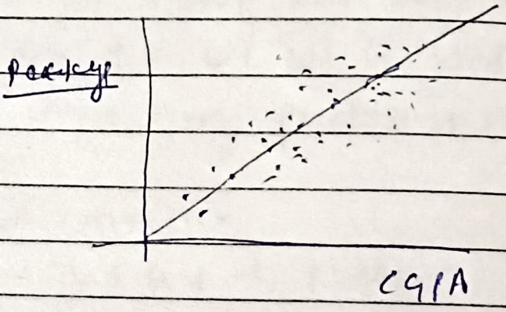
(LR) → Supervised (Regression)

- SLR ⇒ 1 input column & 1 o/p col
- MLR → multiple " " " "
- PLR

→ SLR :-

CGPA	Placement
6.6	3.01
7.1	4
6.2	4.2

i) Plot the graph



- This graph is sort of linear (Not exactly linear).
- Draw a line which passes closely to every points present. This is called best fit line
- Eqⁿ of line is: $y = mx + b$
  ⇒ package $= m * CGPA + b$

'm' tells about ~~pa~~ weightage

→ there are two ways to find m, c
i) Closed form Solⁿ -
  ↳ we can derive mathematical formula
  ↳ doesn't have diff and integration
  ↳ we use OLS (ordinary least Squares)
↳ scikit learn uses it

Page No.			
Date			

ii) Non-closed form

↳ we use approximation to find m, c

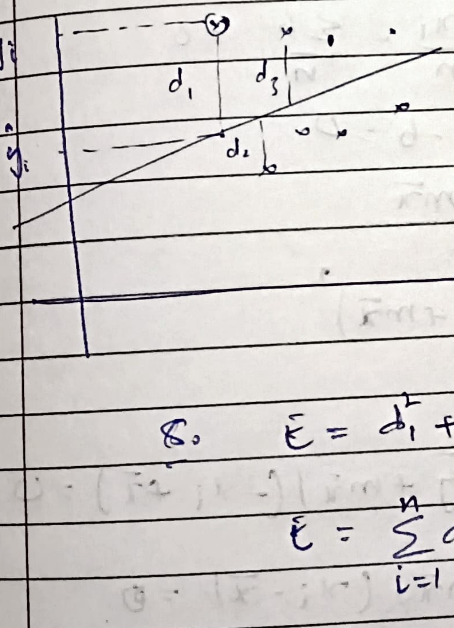
↳ ~~called~~ technique is Gradient descent

↳ used when dealing with higher

Descent

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

<u>Derivation</u> -



$\Rightarrow E = d_1 + d_2 + d_3 + \dots d_n$

↳ since d can be below line so add square and we don't do mode because its not diff-able at x=0

So $E = d_1^2 + d_2^2 + \dots + d_n^2$

$E = \sum_{i=1}^{n} d_i^2$

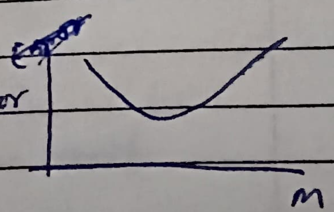
$\rightarrow d_i = (y_i - \hat{y}_i)$ So, $E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

we want that m, b which minimizes error

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

i) Keeping $b = 0$.

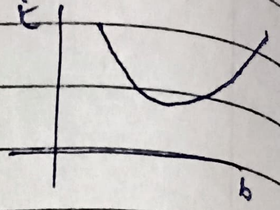
$E(m) = \sum_{i=1}^{n} (y_i - mx_i)$ $\rightarrow$ Error



Page No.
Date

ii) Keeping m constant

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



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

$$\frac{\partial E}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \frac{\partial}{\partial b} (y_i - mx_i - b)^2 = 0$$

$$= \frac{\sum -2(y_i - mx_i - b)}{n} = \frac{0}{n}$$

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

$$\Rightarrow \bar{y}_i - m\bar{x}_i - b = 0$$

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

$$\rightarrow E = \sum (y_i - mx_i - \bar{y} + m\bar{x})^2$$

$$\frac{\partial E}{\partial m} = 0$$

$$\rightarrow \sum 2(y_i - mx_i - \bar{y} + m\bar{x})(-x_i + \bar{x}) = 0$$

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

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

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

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

Regression Matrics

↳ It tells how good a regression model predictions are

1) MAE (Mean Absolute Error)

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

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

↳ Advantage -

- unit of MAE = unit of o/p
- Robust to outlier

↳ Disadvantage.

↳ graph (mode) is not diff^able at 0.

2) MSE (Mean Square Error)

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

↳ Advantage

↳ can use at loss fn (as diff^able)

Disad ⇒ ↳ MSE unit = $(o/p)^2$

↳ Not Robust to outlier

3) RMSE → $\sqrt{MSE}$

↳ output has same unit

Page No.
Date

4) R2 Score (coeff of determinant)

↳ tells how well regression model explains data

↳ also called goodness of fit

↳ $R2 = 1 - \frac{SS_R}{SS_m}$

$$R_2 = 1 - \frac{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_{Reg}}{\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right]_{mean}}$$

$R_2 = 1$ → It means Regression line passes through all pts

$R_2 = 0$ → we are just using avg

So R_2 should be close to 1

R → -ve means our regression is doing even more mistakes

- CGPA | lac

$R2 = 0.8$

means 80% of variation in lac is explained by CGPA

Remaining 20% depends on other factor

- Adjusted R^2 score

→ Suppose by CGPA | package we get R^2 score as 0.8

Now we add irrelevant columns like temp.

Page No.
Date

So R_2 score should decrease but it either remains same or increases

So

$$R^2 \text{ adjust} = 1 - \left[\frac{(1-R_2)(n-1)}{(n-1-k)} \right]$$

$n \rightarrow$ # rows , k = indep. columns

Lecture - 50