day - 52

when you apply regression algorithm then how could you know how good is the algorithm.

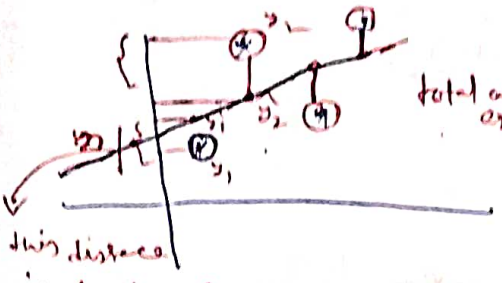
basically we check the efficiency of regression algorithm.

to find this there is many matrics

Regression metrics

1) MAE (mean absolute error)
2) MSE (mean squared error)
3) RMSE (root " " ")
4) R2 score (coefficient of determination)
5) adjusted R2 score.

(1) MAE



$$\text{total abs error} = \sum_{i=1}^{n} |(y_i - \hat{y}_i)|$$

$$\text{mean absolute error (MAE)} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

↳ neglecting the sign

advantage :-

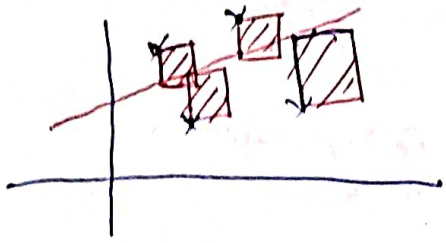
1) same unit (like if the error is in LPA, if its error the data is in LPA)

2) Robust to outlier (it can handle the outliers. MAE can also effect on if outlier present but compared to others it is less.

disadvantage :- not differentiable at 0

∴ we do many optimization techniques where we use differentiation

(2) MSE



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

↳ In previous data set we use mse as our loss function.

advantage :-

|→ we can use as loss function

Disadvantage :-

|→ (unit) :- |→ for outliers it give more attention (not robust to outliers)

Rmse : $\sqrt{mse}$

advantage

↳ same unit

disadvantage

↳ mse is not that robust to outliers.

(4) **R2 score** (coefficient of determination)
↳ goodness of fit

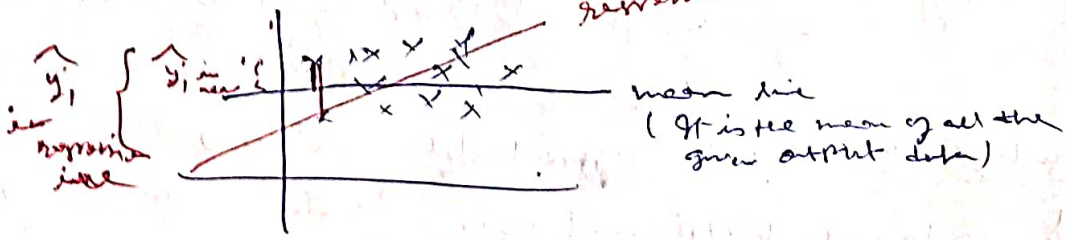
cgpa | Package (lpa)

NOTE:-

If we don't know what is cgpa then for same body how can you tell the Package simply by Previous Package mean

this is not the best but we can say something about the package

but when you bring cgpa it- bring us the regression line



So, R2 score calculate how good your linear regression line better than mean line

$$R_2 = 1 - \frac{SS_R}{SS_m}$$

SS_r = Sum of Squared error in regression line

SS_m = " " " " " mean "

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

Now how to interpret R2 score

$R2 = 0 \Rightarrow \frac{[\quad]_{reg}}{[\quad]_{tm}} = 1 \Rightarrow [\quad]_{reg} = [\quad]_{tm}$

i.e the mistakes did in regression line is same as the mistakes did the mean line

$R_2 = 1 \Rightarrow \frac{[\quad]}{[\quad]} = 0 \Rightarrow [\quad] = 0$

$\Rightarrow$ i.e regression line doesn't make any mistakes

So, after calculate R2 it should goes toward 1

$R_2 = -ve = \frac{[\quad]}{[\quad]} > 1$

$[\quad] > [\quad]$

$SSR > SSm$

So, SSR line made more mistakes than SSm

R2 → 0.80 → cgpa explains 80% of variance in lpa column

cgpa | lpa

i.e it gives explanation of 80% but for rest 20% we don't know

for ex:- cgpa | iq | lpa

80% explanation gives via cgpa & iq

rest 20%, we don't know it may be good interview skills.

(R2 × 100)% gives amount of explation by input column.

ex:- In R2 score
cgpa / lpa

If we add more input - then R2 score will increase.

this is not the issue

but - if your give irrelevant - column
it - doesn't impact - on lpa but - sometime
R2 score gets increase. for ex:- there is temperature column

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

n = no of rows
k = total no of independent - column.

so, if we add temp column $k \uparrow$

$\therefore (n-1-k) \downarrow$

$[\quad] \uparrow$

$1 - [\quad] \downarrow$

now if we add $\frac{\text{iq}}{\text{it is relevant}}$

$k \uparrow$, $(n-1-k) \downarrow$, $R^2 \uparrow$, $(1-R^2) \downarrow$

but $(1-R^2)$ decrease more faster than $(n-1-k)$

$[\quad] \downarrow$

$1 - [\quad] \uparrow$

Ex) It is very useful when there are multiple column
ex:- Multiple regression