Performance metrics used in Regression: To calculate the accuracy of a Regression model we use performance metrics.

There are two tecniques:

① R squared ② Adjusted R squared

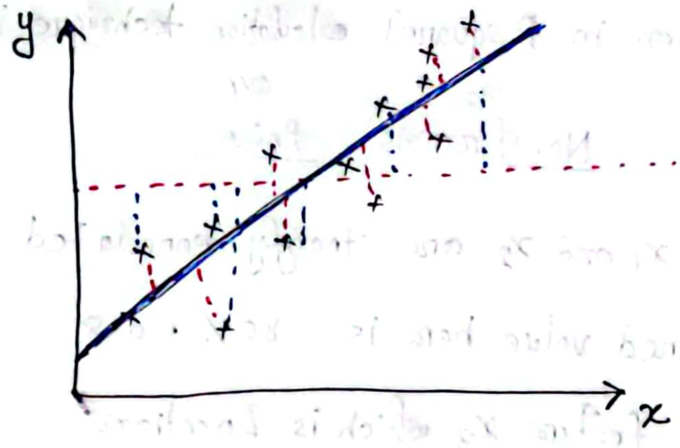
R squared Formula:

$$Rsquared = 1 - \frac{SSRes}{SSTotal}$$

SSRes = Sum of square residual {Error}

SSTotal = Sum of square total

SSRes = The total error $\Sigma(y_i - \hat{y}_i)^2$ $[\Sigma(\text{actual point} - \text{predicted point})^2$



SSTotal → We take a line (straight) from the average value of y.

Now the summation of $(y_{avg} - \text{actualpoints})^2$ would be SS Total

$$\Sigma(y_{avg} - y_i)^2$$

So, we can write the R squared Formula like below →

$$R\ squared = 1 - \sum_{i=1}^{n} \frac{(y_i - h_\theta(x))^2}{(y_i - \bar{y_i})^2}$$

$h_\theta(x) = \hat{y}_i$ (predicted point)

$\bar{y_i}$ = y mean

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

R squared ranges between 0 to 1.

② Adjusted R squared:

One of the problem in R squared calculation technique is. Suppose

x_1 House size	x_2 No. of rooms	O/P Price

Both the feature x_1 and x_2 are strongly correlated with output (Price)

Suppose, R squared value here is = 80% = 0.8

Let's add anothe feature x_3 which is Location

x_1 House Price	x_2 No. of rooms	x_3 Location	O/P Price

Here 3 of features are strongly correlated with output. For that R squared will increase even more. Suppose that become → 90%

Now, add another feature, which is gender (x_4)

x_1	x_2	x_3	x_4	O/P
House size	No. of rooms	Location	Gender	Price

In the case of gender, this feature is not highly correlated or important for output prediction. Although for this feature also R squared a bit. Let's say it increased 1% and became 91%.

So, no matter what feature we are adding, R squared is increasing. That is not right for the model accuracy. For gender feature R squared should not increase. Adjusted R squared solve this specific problem.

Formula of Adjusted R squared:

$$1 - \frac{(1-R^2)(N-1)}{N-P-1}$$

N = Num of datapoints

R^2 = R squared

P = Num of independent features.

We should perform both R squared and adjusted R squared to know that not every feature is important.

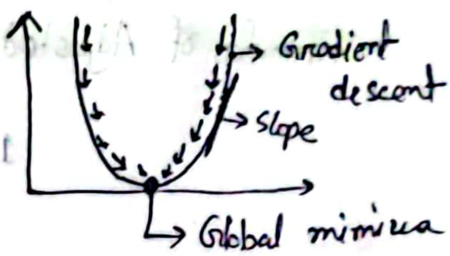
Cost Functions (MSE, MAE, RMSE)

1. Mean Squared Error (MSE)
2. Mean Absolute Error (MAE)
3. Root Mean Squared Error (RMSE)

Previously we already discussed about Mean squared Error whose Formula was →

$$J(\theta_0, \theta_1) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - h_\theta(x)_i \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \rightarrow \text{quadratic Equation}$$

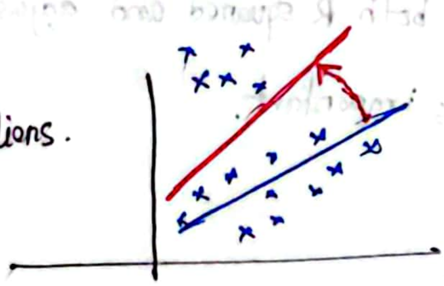


Advantage of MSE:

1. Equation is differentiable.
2. It has only one Local/Global minima

Disadvantage of MSE:

1. Not Robust to outliers.



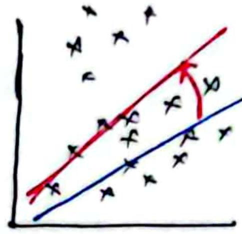
If dataset has outlines, Best fit line moves away from where it should be to ~~and~~ the side of outliers a bit. (can't handle situation)

② It is not in the same unit.

As we are squaring the errors, then the unit will also be squared.

(ii) Mean Absolute Error (MAE):

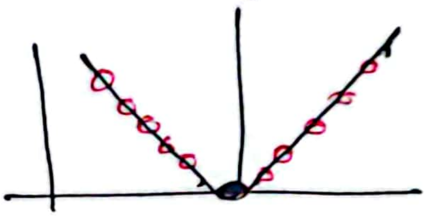
Formula: $\frac{1}{n} \sum_{i=1}^{n} | y_i - \hat{y}_i |$



Advantage:

① Robust to outliers →

② It will be in the same unit



Disadvantage:

① Convergence usually take more times.

(iii) Root Mean squared Error (RMSE):

Formula, $RMSE = \sqrt{MSE}$

$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - h_\theta(x)_i)^2}$$

Advantages:

① Same Unit

② Differentiable

③ 1 Global minima

Disadvantage:

① Not Robust to outliers

When you have outliers → USE MAE

When you don't have outliers → USE MSE, RMSE