2-Day
15-Oct-2022

Loss function vs cost function

In case of cost function all the point are considered.

In case of loss function it is for every observation

$$\text{loss function} = (h_\theta(x)^{(i)} - y^{(i)})^2$$

$$= (\hat{y}_{(i)} - y_{(i)})^2$$

$\hat{y}_{(i)}$ → Preselicted, $y_{(i)}$ → Actual

$$J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x)^{(i)} - y^{(i)})^2$$

→ cost function
→ MSE

→ convergence Algorithm

repeat until ~~Algorithm~~ convergence {

$$\theta_j : \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_j)$$

}

$$\Rightarrow \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{\partial}{\partial \theta_0} \left[\frac{1}{m} \sum_{i=1}^{m} (h_\theta(x)^{(i)} - y^{(i)})^2 \right]$$

$$\Rightarrow h_\theta(x) = \theta_0 + \theta_1 x$$

$j=0$

$$= \frac{\partial}{\partial \theta_0}\left[\frac{1}{m}\sum_{i=1}^{m}\left((\theta_0+\theta_1 x)^{(i)} - y^{(i)}\right)^2\right]$$

$$= \frac{2}{m}\sum_{i=1}^{m}\left[(\theta_0+\theta_1 x) - y^{(i)}\right] * 1$$

$j=1$

$$= \frac{\partial}{\partial \theta_1}\left[\frac{1}{m}\sum_{i=1}^{m}\left[(\theta_0+\theta_1 x)^{(i)} - y^{(i)}\right)^2\right]$$

$$= \frac{2}{m}\sum_{i=1}^{m}\left((\theta_0+\theta_1 x)^{(i)} - y^{(i)}\right] * x$$

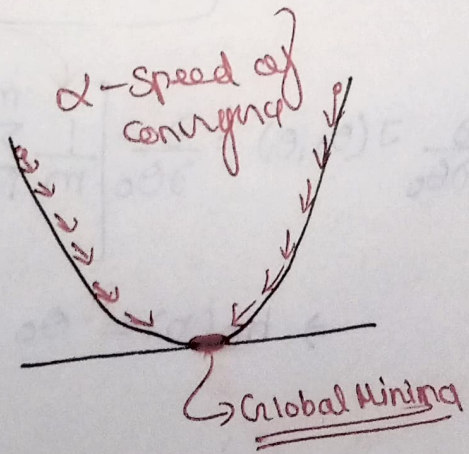
Repeat until convergence

{

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x)^{(i)} - y^{(i)}\right)x^{(i)}$$

}



→ L2 loss

Cost functions ① MSE ② MAE (L1 loss) ③ RMSE

① MSE {Mean Squared Error} (use when no outliers or treat them)

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

$\hat{y} = \theta_0 + \theta_1 x$

↳ Predicted value

↳ quadratic Eqn — Plot → Parabola

↓

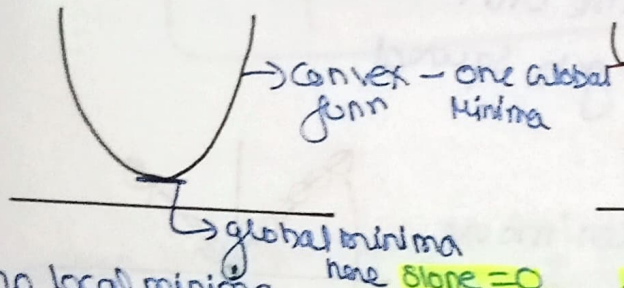
always get parabola not having local minima.

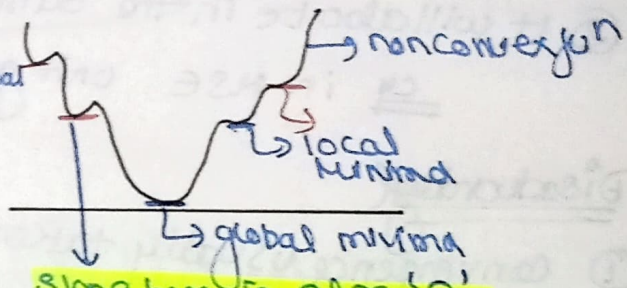
Advantage

① This Eqn is Differentiable

- Diff. needed to calc slope & update θ_0 & θ_1

② This Eqn also has one global minima





- no local minima

here slope = 0

∴ θ_0 & θ_1 not get updated

slope here fq also '0'

∴ θ_0 & θ_1 will not get updated & get stuck at local minima & never reach global minima!

Disadvantage

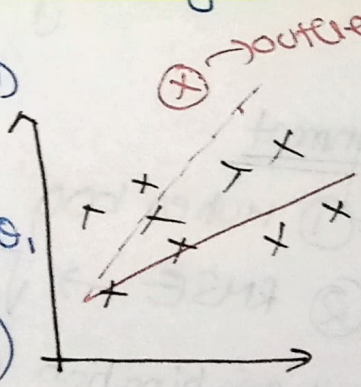
1) This is not robust to outliers

- coz of outliers → cost fun ↑↑↑

↓

update θ_0, θ_1

∴ Remove outliers



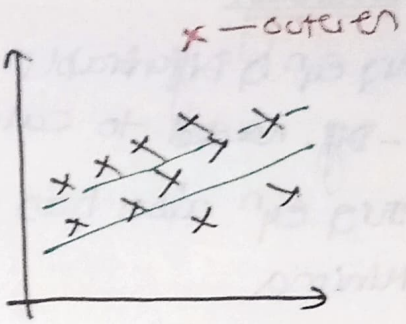
2) Penalizing the Error (changing the unit)

Eg (real salary – predic salary)^2 in Lakhs

= Error^2 (lakhs)^2

② Mean Absolute Error (can be use with outliers)

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

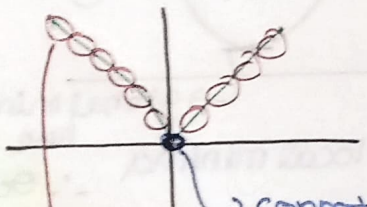


Advantage

① Robust to outliers
② It will also be in the same unit
as in MSE unit gets squared

Disadvantage

① Convergence usually takes more time. optimization is a complex task
② Time Consuming



Derivative is calculated by Sub gradient Concept
- Divide Region by Region & find derivative & then come near to the Minimum.

→ cannot find derivative if point is zero.

Assignment

① Huber loss
② RMSE → $\sqrt{MSE}$ ⇒ Advantage & disadvantage
→ combination of MSE & MAE

③ Huber loss (Smooth Mean absolute Error)

- Combination of MSE & MAE. It takes good properties of both the loss function by being less sensitive to outliers & differtible at minima.
- when error is less, MSE part of huber is utilized & when error is large, the MAE part of huber loss is used.
- A hyperparameter is use to tell loss function to switch from MSE to MAE.

$$\text{loss} = \begin{cases} \frac{1}{2} * (y-\hat{y})^2 & \text{if } (|y-\hat{y}| \le \delta) \\ \delta * |y-\hat{y}| - \frac{1}{2} * \delta^2, & \text{otherwise} \end{cases}$$

④ Root Mean Squared Error (RMSE)

$$L = \sqrt{\frac{1}{N}\left[\sum (\hat{y} - y)^2\right]}$$

other loss functions in ML:-

1. Sum of Errors $L = \sum (\hat{y} - y)$
2) Sum of absolute Error $L = \sum (|\hat{y} - y|)$
3) Sum of ~~Errors~~ Squared Error, $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
4) Mean Bias Error (MBE)

$$MBE = \frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)$$

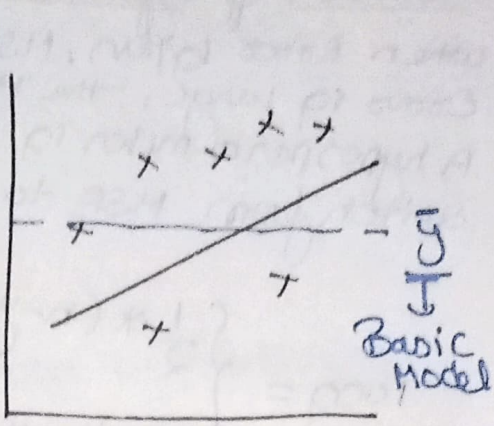
$\hat{y}$ - predicted

y - actual value

Performance Matrix

R Squared – Measure performance of Model that have been created.

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}}$$



SS_{Res} = Sum of squared Residual

SS_{Total} = Sum of squared avg.

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

$\sum (y_i - \hat{y}_i)^2$ → low value

$\sum (y_i - \bar{y})^2$ → high value

} if model fitted well

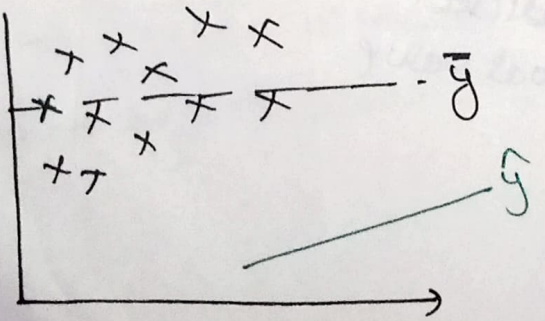
$$= 1 - \left\{\frac{\text{Small no.}}{\text{Bigger no.}}\right\} \rightarrow \text{small no.}$$

$R \leq 1$ – if model fitted well

⇒ 0.85 → 85% accurate

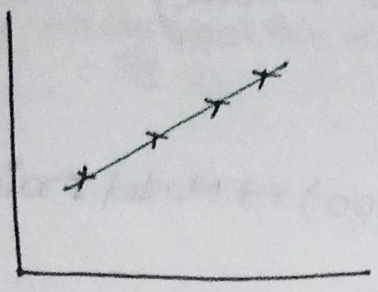
⇒ 0.75 → 75% accurate

Q Can R squared value be -ve → **Model is very Bad**



$R^2 = -ve$

* $R^2 = 1$



Adjusted R Square → will going to evaluate accuracy based on very important features

Dependent feature

Size of house	City	No. of Rooms	Gender	Price

- Coz of addition of more independent feature the R^2 may inc which is Desired. But there may be such feature which have no correlation b/w feature & output but there may be inc in accuracy but inc in accuracy is very slight. In order to solve such problem adjusted R^2 is used.

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

N = no. of data points
P = no. of independent feature

eg

R^2	65%.	P=1
R^2	75%.	P=2
R^2	88%.	⋮
R^2	88	P=4

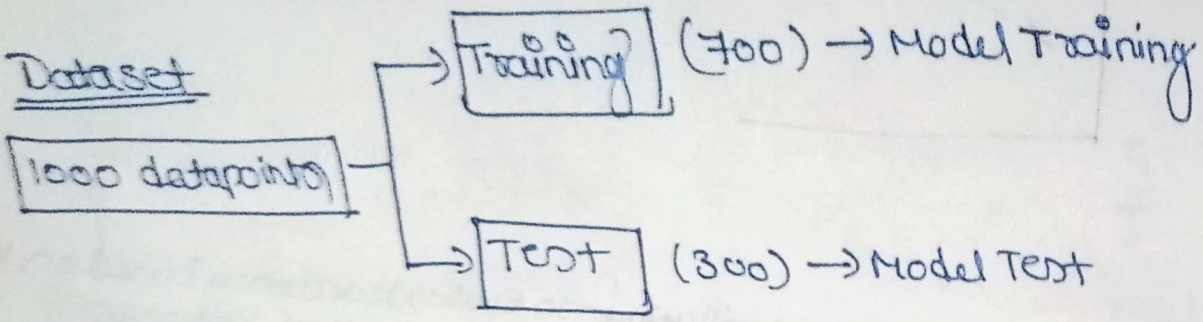
Adj R^2 = 63%.] less than R^2
Adj R^2 = 73%.

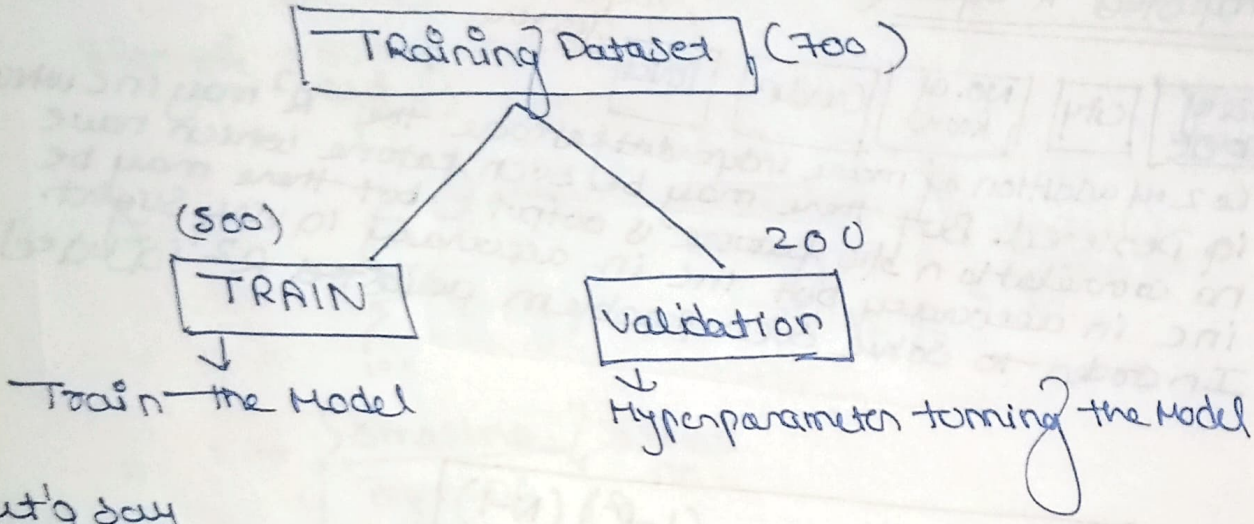
Adjusted R^2 = 85%.
↓
Dec adjusted R^2 Coz that feature is not correlated.

P=4 ← eg gender added ← less correlated feature

- In inc unnecessary columns R^2 will inc but Adj R^2 will dec.

Overfitting & underfitting (Bias and Variance)





Let's say
- For a Model

Train Data - very good Accuracy (90%.) [Low Bias]

Test Data - very good Accuracy (85%.) [low variance]

Train Data Accuracy ——— Bias

Test Data Accuracy ——— Variance

*

TRAIN - very Good Accuracy [90%] [Low Bias]

Test - Bad Accuracy [50%] [High variance]

↓

Overfitting

*

TRAIN - Model Accuracy is low [High Bias]

TEST - Model Accuracy is low/high [Low or High variance]

↓

Underfitting