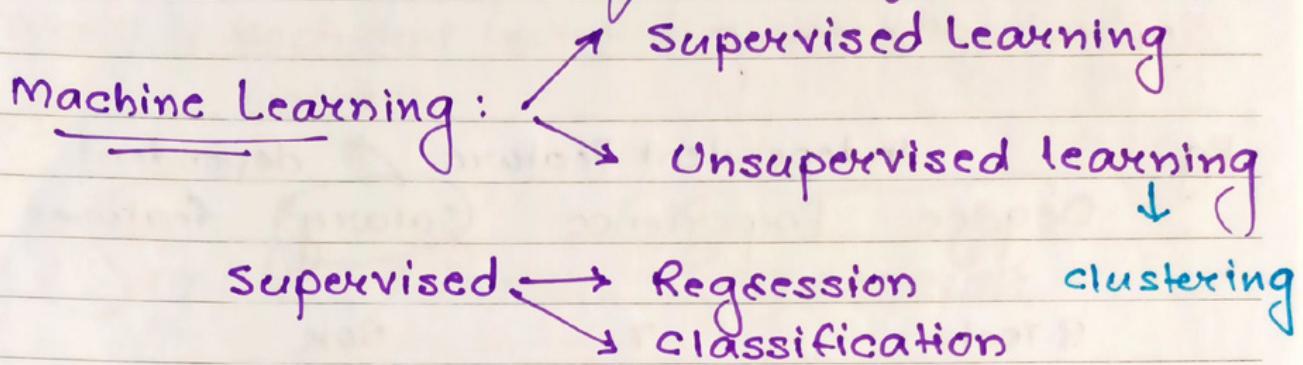


Machine Learning.



Supervised (output is known)

- Regression →
1. Linear Regression
 2. Polynomial Regression
 3. SVR
 4. Decision Tree
 5. Random Forest
 6. Xgboost , 7. KNN

Classification: →

1. Logistic Regression
2. SVM
3. Decision Tree
4. Random Forest
5. Naive Bayes
6. KNN

Unsupervised learning: (output is unknown)

clustering → DBScan, kmeans,
Hierarchical,
Silhouette scoring

Supervised Learning

Eg: Independent feature \rightarrow dependent features

Degree	Experience	Salary
B.Tech	1.7	50k
PhD	2	70k
-	-	-

O/p = continuous \rightarrow Regression problem

Eg. Independent Feat. \rightarrow dependent
No. of plays No. of study hours Pass/Fail Feat.

O/p = categorical = classification problem

- Flight Price Prediction \rightarrow Regression problem
- Algerian Fire Forest \rightarrow classification problem
- Predict Air Quality Index \rightarrow Regression Problem
- Rain Prediction \rightarrow classification Problem
- Buying day of a person \rightarrow classification problem

Note: Independent features is basically input features.

Simple Linear Regression.

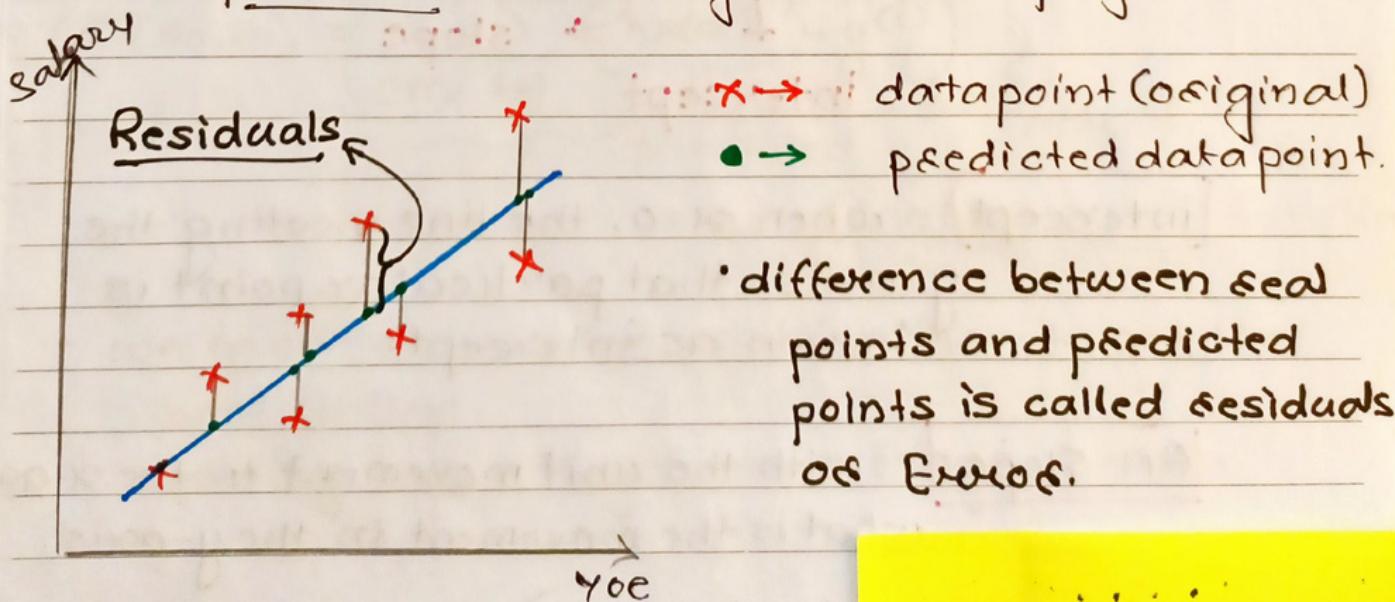
(one independent feature and one dependent feature)

Eg. aim: to create a model, which takes input as height and predict weight.

dataset: height, weight.

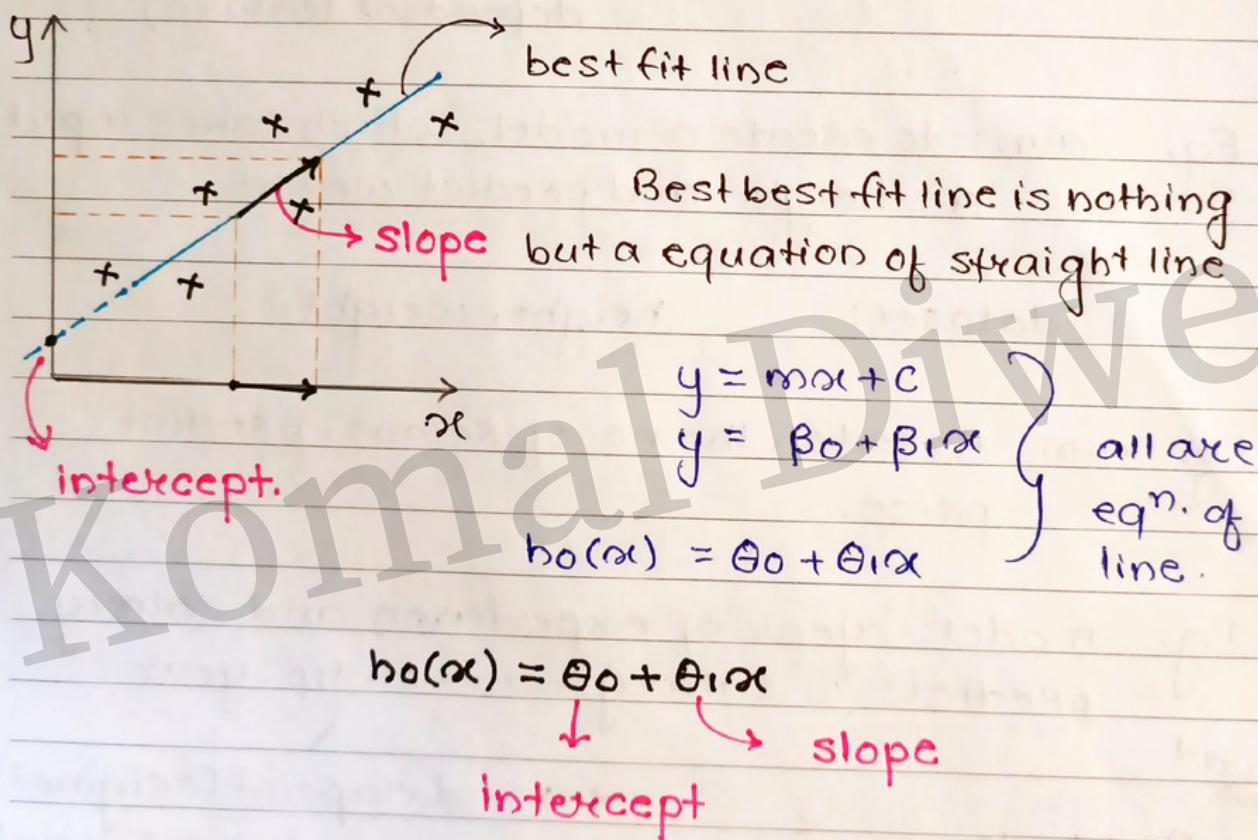
Eg. aim: Based on the no. of rooms, predict price.

Eg. model: year of experience and salary
predict: salary based on 1st year.



- Based on the training dataset, it finds the best fit line in such a way that the sum of difference between real points and predicted should be minimum.

First, we need to understand why are creating a straight line?



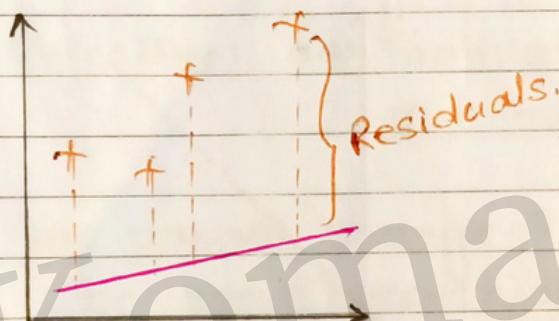
[Intercept]: when $x=0$, the line meeting the ~~line intersecting~~ y-axis, that particular point is ~~is called~~ known as intercept.

[θ_1 : Slope]: with the unit movement in the x-axis, what is the movement in the y-axis.

- By changing θ_0 and θ_1 , best fit line will be change.

θ_0, θ_1 : changing the values, will change the line.

after some iterations, we will get a best fit line which is known as training of the model.



we need to minimize this excess using equation called as cost function.

Cost Function:

for easy calculations (derivatives)

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m \left(\underbrace{h(\mathbf{x})^{(i)}}_{\text{predicted}} - \underbrace{y^{(i)}}_{\text{actual.}} \right)^2$$

mean square error. (one of the cost function)

- we need to minimize cost function to get the best fit line.
- dividing by m to get the average (mean).

Final aim:

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

• by changing the values of θ_0 and θ_1 .

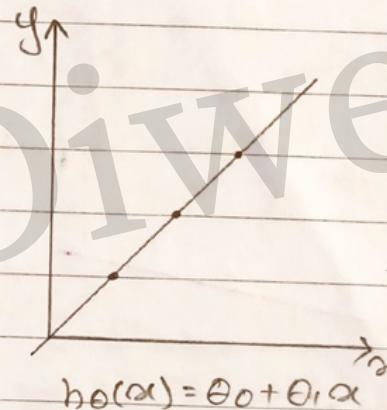
Eg. Training dataset.

x	y
1	1
2	2
3	3

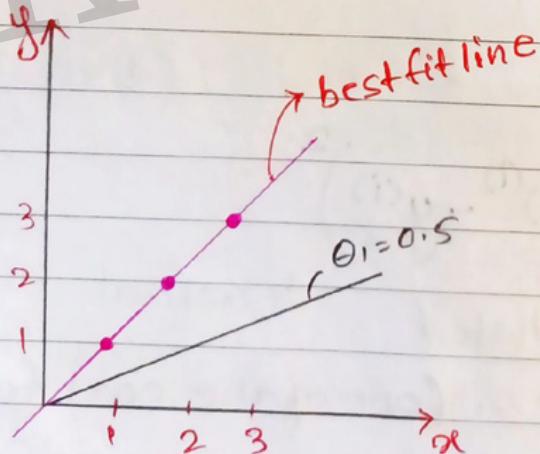
Let us consider

$$\theta_0 = 0$$

$$h_\theta(x) = \theta_1 x$$



Now, let's assume $\theta_1 = 1$
 $\therefore h_\theta(x) = x$.

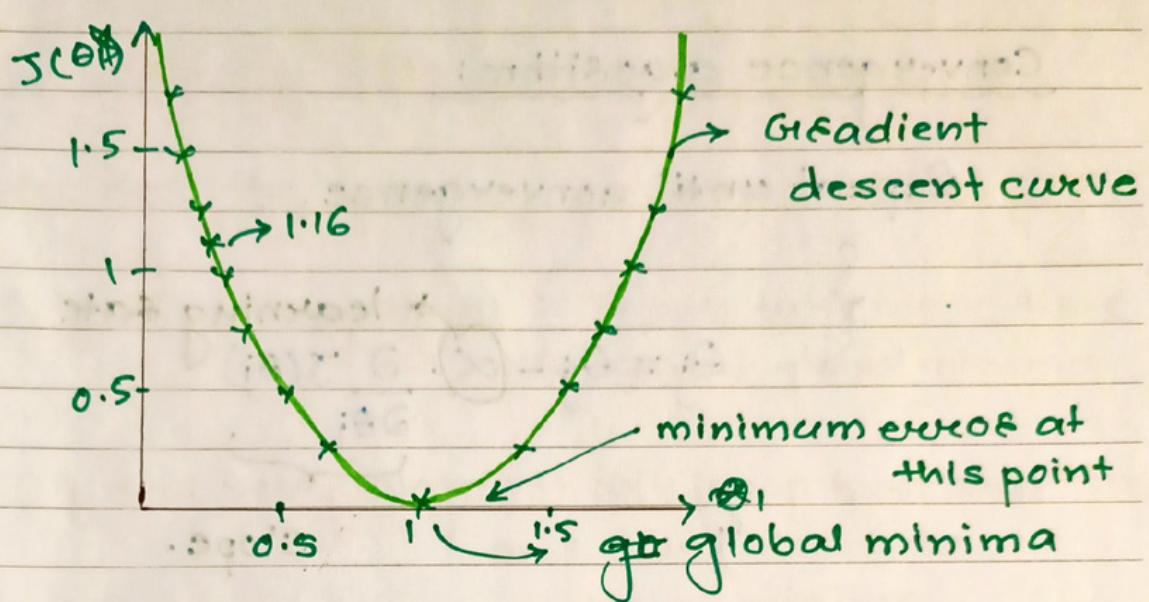


$$\begin{aligned} J(\theta_1) &= \frac{1}{m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2 \\ &= \frac{1}{3} [(0+0+0)] \\ &= 0. \end{aligned}$$

when $\theta_1 = 0.5$

$$\begin{aligned} J(\theta_1) &= \frac{1}{3} [(0.5-1)^2 + (1-2)^2 + (1.5-3)^2] \\ &\approx 1.16. \end{aligned}$$

Komal Diwe



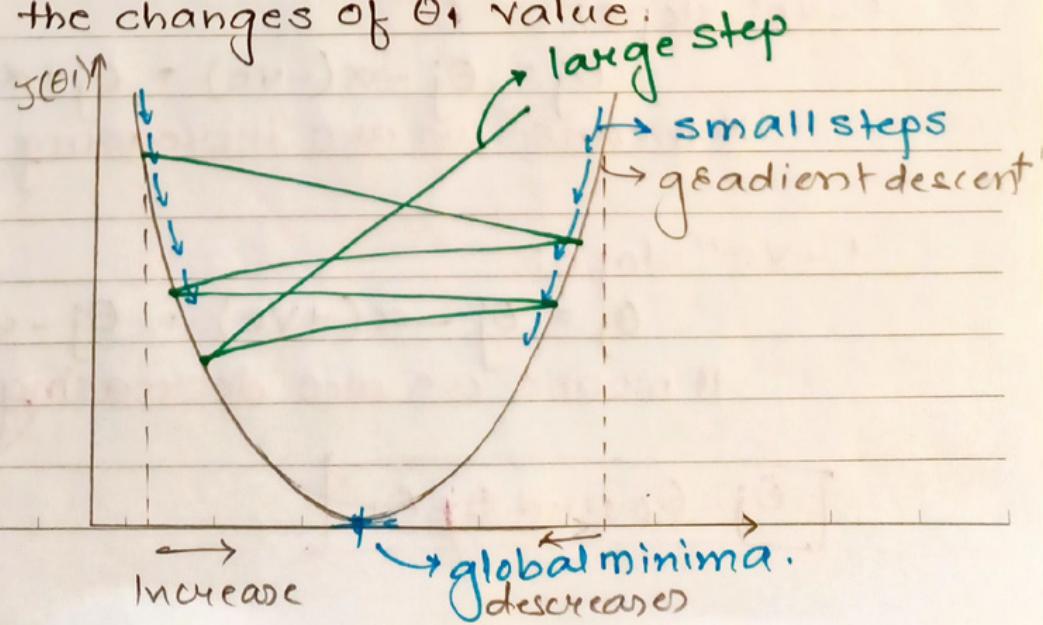
Objective: our main aim is to come near global minima.

we cannot change θ value manually, there should be some mechanism to change θ value to get the global minima.

To overcome this, we use **Convergence algorithm**.

Convergence algorithm:

- optimize the changes of θ_1 value.



Convergence algorithm:

Repeat until convergence,

{

$$\theta_j = \theta_j - \alpha \cdot \underbrace{\frac{\partial J(\theta)}{\partial \theta_j}}_{\text{slope.}}$$

learning rate

}

+ve or -ve slope:

left side of the line facing downwards
→ (-ve) slope

right side of the line facing upwards
→ (+ve) slope

(-ve) slope:

$$\theta_j = \theta_j - \alpha(-\text{ve}) = \theta_j + \alpha$$

It means we are increasing θ_j .

(+ve) slope:

$$\theta_j = \theta_j - \alpha(+\text{ve}) = \theta_j - \alpha$$

It means we are decreasing θ_j .

$$\boxed{\theta_j = \theta_0 \text{ and } \theta_j \neq \theta_1}$$

α , Learning Rate

It decides the speed of the convergence.

"if α is very small": It will take more time to reach global minima.

"if α is very large": It will jump here and there, and won't reach to global minima.

α , should be around 0.001 for smaller steps.

Convergence algorithm:

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

↓

predicted

actual value

meansquare error.

m = no. of datapoint.

difference between cost function and loss function.

cost function: we find errors for all the points and take average of it.

loss function: we find errors for observed points and if we find errors for all the points then it is loss function.

$$\begin{aligned} \text{loss function} &:= \left(\underbrace{h_{\theta}(x^{(i)})}_{\text{predicted value}} - \underbrace{(y^{(i)})}_{\text{actual value}} \right)^2 \\ &= (\hat{y}_j - y_j)^2 \end{aligned}$$

* for every single point we need to find loss function.

To achieve global minima,

derivative w.r.t $\theta_0, j=0$

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

$$\approx \frac{1}{2m} h_\theta(x) = \theta_0 + \theta_1 x$$

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

$$= \frac{1}{m} \sum_{i=1}^m ((\theta_0 + \theta_1 x)^i - y^i) \times 1$$

$$\frac{\partial}{\partial \theta_0} (1+x)^2$$

$$= 2[(1+x)^2] \times 1$$

$$= \frac{1}{m} \sum_{i=1}^m ((\theta_0 + \theta_1 x)^i - y^i)$$

derivative w.r.t. $\theta_1, j=1$

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

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

$$= \frac{1}{m} \sum_{i=1}^m ((\theta_0 + \theta_1 x)^i - y^i) x.$$

Repeat until convergence

$$\theta_0 := \theta_0 - \alpha \cdot \frac{1}{m} \sum_{i=1}^m (h_\theta(x)^i - y^i)^2$$

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

}

* learning rate: speed of convergence

Types of cost function:

- (1) MSE: mean square error
- (2) MAE: mean absolute error
- (3) RMSE: root mean square error.

Mean Square Error

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

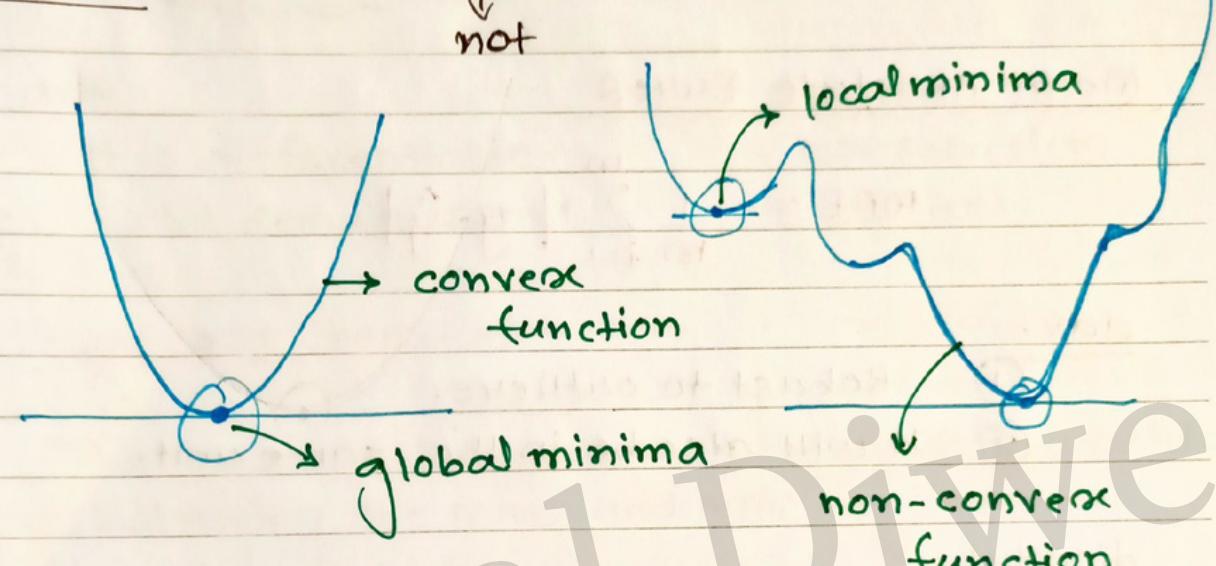
actual.
 $\therefore \hat{y} = \theta_0 + \theta_1 x$
predicted.

Adv

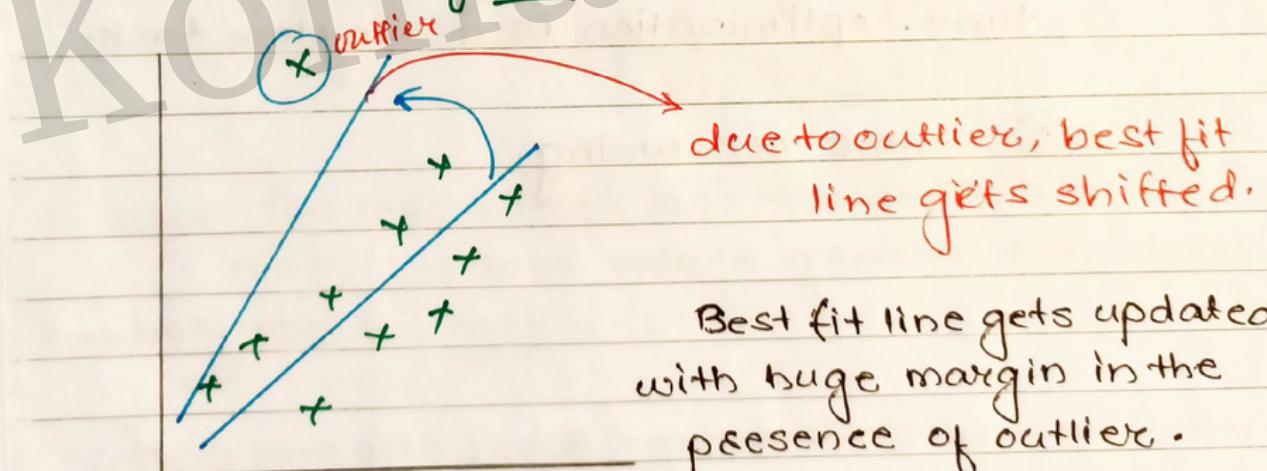
- This equation is differentiable.
- It also has one global minima.

disadv:

- This is robust to outliers.
not



∴ main adv. of MSE that it has convex function.



Best fit line gets updated
with huge margin in the
presence of outlier.

Eg. (salary) \rightarrow dependent feature

Experience \rightarrow independent feature

$$(y - \hat{y})^2 \text{ (lakhs)}^2$$

\hookrightarrow unit changing \rightarrow time complexities increased.

Komal Diwe

we don't do scaling for dependent feature.

$(y - \hat{y})^2 \rightarrow$ squared \rightarrow error \rightarrow penalized
 \downarrow
Increased.

Mean Absolute Error

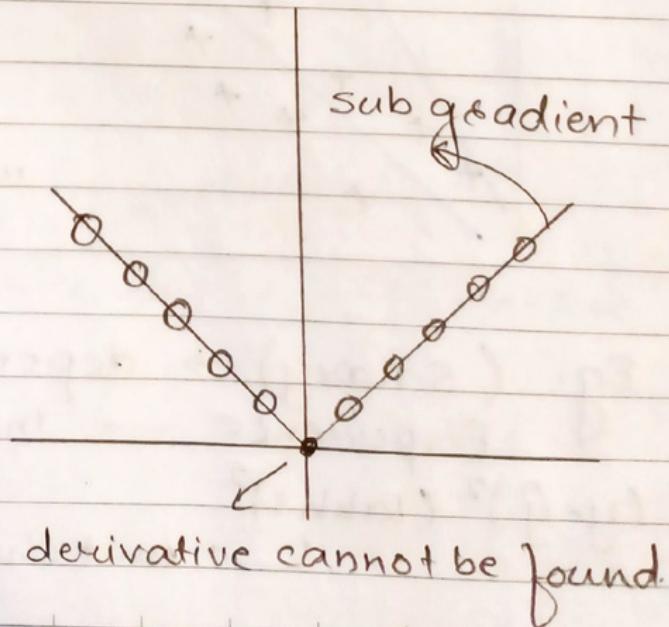
$$MAE = \frac{1}{m} \sum_{i=1}^m |y - \hat{y}|$$

adv:

- ① Robust to outliers.
- ② It will also be in the same unit.

disadv:

- ① convergence usually takes more time. optimization is a complex task.
- ② time consuming.



Root mean Square Error (RMSE)

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

adv:

- It is differentiable.
- unit remain same.

disadv:

- Not robust to outliers.

Huber Loss Function

The Huber loss offers the best of both worlds by balancing the MSE and MAE together.

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2} (y - f(x))^2 & \text{for } |y - f(x)| \leq \delta, \\ \delta |y - f(x)| - \frac{1}{2} \delta^2 & \text{otherwise.} \end{cases}$$

It says: for loss values less than delta, use the MSE, for loss values greater than delta, use the MAE.

using the MAE for larger loss values mitigates the weight that we put on outliers so that we still get a well-sounded model.

At the same time, we use the MSE for the smaller loss values to maintain a quadratic function near the centre.

This has the effect of magnifying the loss values as long as they are greater than 1.

Once the loss for those data points dips below 1, the quadratic function down-weights them to focus the training on the higher-error data points.

Note: use the Huber loss any time you feel that you need a balance between giving outliers some weight, but not too much.

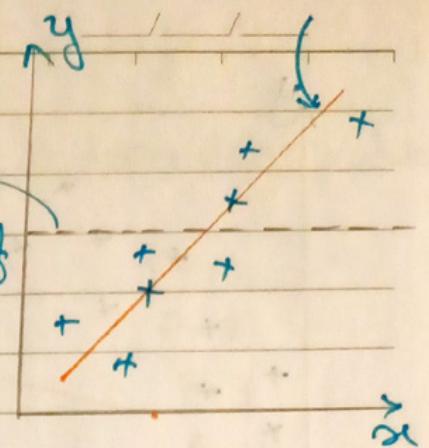
How can we check if a model is good or not?
→ using performance metrics.

Performance metrics

- ① R-squared
- ② Adjusted-R squared

R-squared: measures the performance of the model.

$$R^2 = 1 - \frac{SS_{\text{Res}}}{SS_{\text{Total}}} \quad \bar{y} \leftarrow \text{(average of } y\text{)}$$



$SS_{\text{Res}} = \text{Sum of Square Residuals}$

SST_{Total} = Sum of square average

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

low value

high value

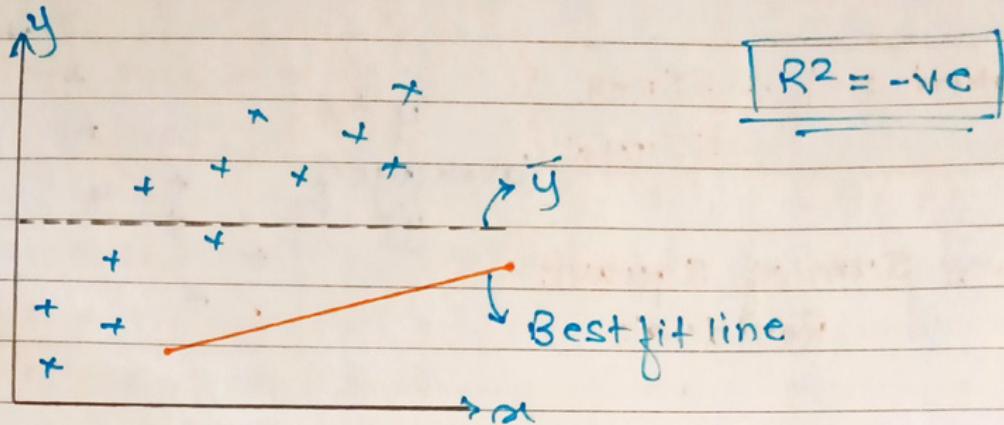
\bar{y} = average of y .

$$= 1 - \left\{ \frac{\text{small value}}{\text{Bigger value}} \right\} \rightarrow \text{small value}$$

R-squared \leq 1

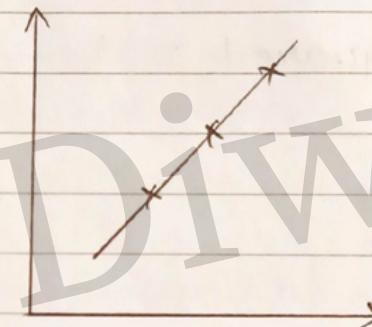
If R-squared = 0.85 = 85% accurate
= 0.75 = 75% accurate

* If R-squared is 'neg' then the model is not working is good.



Here $(y_i - \hat{y}) > (y_i - \bar{y}) \rightarrow R^2$ will be negative.

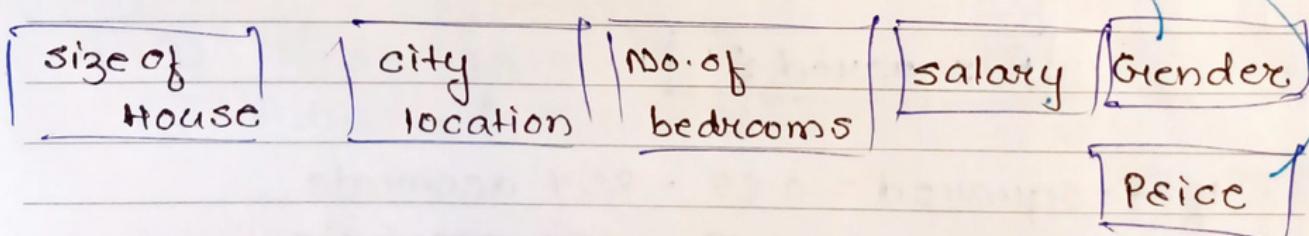
$$R^2 = 1 \rightarrow$$



Adjusted R-squared

Eg.

No direct correlation



Before, when only size of house was present as an independent feature to get the price and has some R^2 value. But with addition of city location R^2 got increase. after that with addition of no.of bedroom, gender the value got increase, but there is not no direct correlation between gender and price, the value shouldn't increased.

R^2 features added adj. R^2 Independent feature

65%. size of house 63%. $P=1$

75%. location 73%. $P=2$

88%. no. of bedrooms 86%. $P=3$

90%. Gender

no direct correlation w.r.t price

* very slightly increase, but this shouldn't happen.
To solve this problem we use Adjusted R^2 .

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

N = No. of datapoints

P = No. of independent features

From the above table

* Adjusted R^2 is the best metrics to evaluate the model.

Overfitting and Underfitting

dataset

| 1000 datapoints |

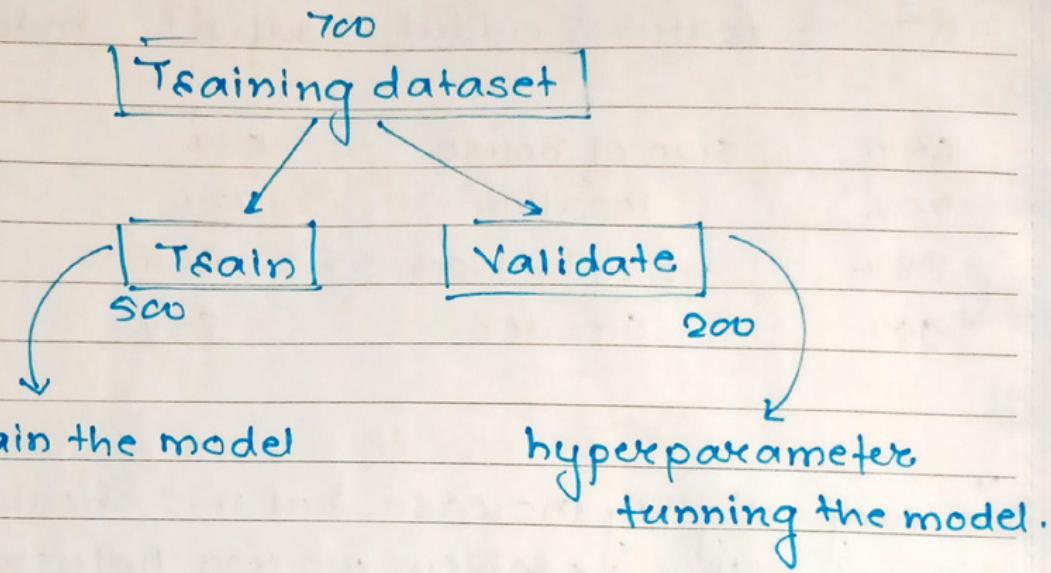
TRAINING

model training
(700)

TESTING

(300) model test

overfitting and underfitting (Bias and Variance)



MODEL

Generalize

Model

Ttrain Data → very good accuracy (90%)

Test Data → very good accuracy (85%)

our aim is to get good test and train accuracy.

Ttrain Data → very good accuracy 90% [low bias]

Test Data → very good accuracy 85% [low variance]

Overfitting

TRAIN → very good accuracy (eg.) [low Bias]

TEST → Bad accuracy (eg.) [High Variance]

Underfitting

TRAIN → model accuracy is low. [High bias]

TEST → model accuracy is low/high. [low & high variance]

- * we can solve this issue by performing hyperparameter tuning or by increasing no. of dataset.