Memo No. _____

Date / /

To achieve global minima:

derivative w.r.to θ_0, $j=0$

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

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

for $j=0$

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

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

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

derivative w.r.t θ_1, $j=1$

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

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

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Repeat untill convergence:

{

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

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

}

* 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:

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

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

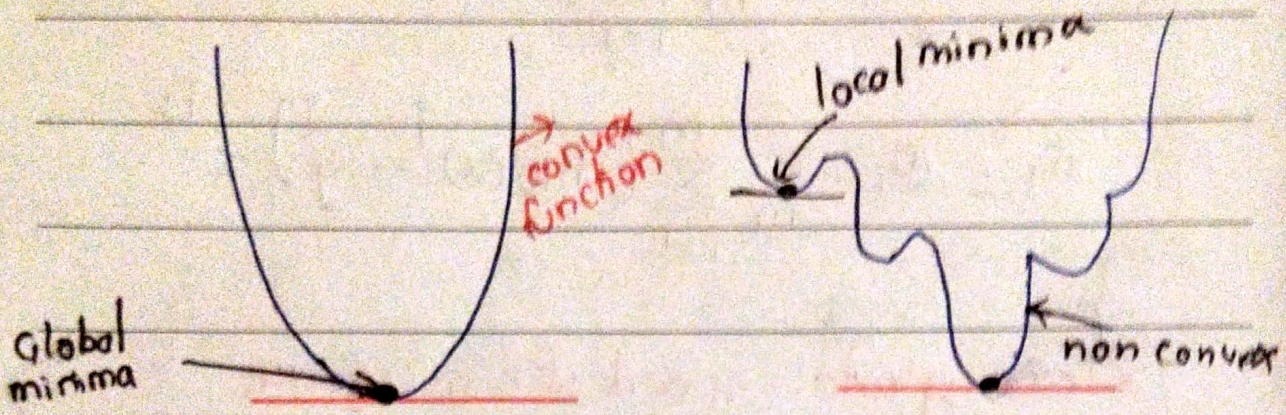
↑ predicted

Memo No.

Date / /

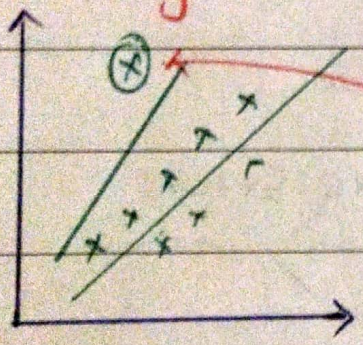
Advantage:- ① This equation is differentiable.

② It has only one global minima.



∴ main aim = It has convex function.

Disadvantage:- !



due to outlier, best fit line gets shifted.

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

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

↳ unit^2 → unit changing

↓

time complexities increasing

Memo No. _____

Date / /

We wont do scaling for depending features

$(y - \hat{y})^2 \Rightarrow$ got squared $\Rightarrow$ error $\Rightarrow$ penalized (increase) value

Mean Absolute Error

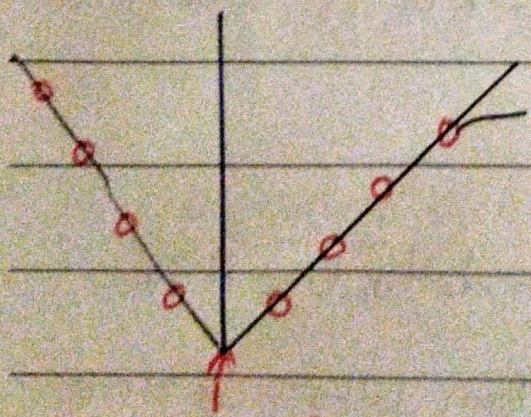
$$MAE = \frac{1}{m} \sum_{i=1}^{m} |y - \hat{y}|$$

Advantage:

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

Disadvantage:

1. Convergence usually takes more time.
   Optimization is a complex task.
2. time consuming



subgradient

→ generalize the derivative to convex function which are not necessary to be differentiable.

→ Convex analysis method.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Root Mean Square Error (RMSE)

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

Advantage	Disadvantage
• It is differentiable.	• Not robust to the outliers
• unit remains same	

Huber loss function:

The huber loss offers the best of world 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 MSE

→ for loss values greater than delta, use MAE.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

* MAE for larger loss

Using MAE for larger loss values mitigate the weight that we put on outliers so that we still get a well-grounded model.

MSE for smaller loss Why?

→ to maintain a quadratic function near the centre.

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

Memo No. _____

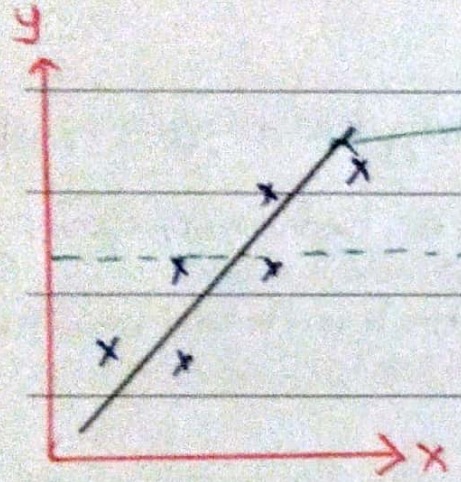
Date / /

1) R-Squared:- Measure the performance of the model.

$$\text{R-Squared} = 1 - \frac{SS_{RES}}{SS_{Total}}$$

SS_{RES} = Sum of square residuals

SS_{Total} = Sum of square average



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

low value → numerator

high value → denominator

$$= 1 - \frac{\text{small value}}{\text{large value}} \quad \text{small value}$$

$\therefore$ R squared ≤ 1

Disadvantage of R^2 : → Even additional input variable shows no rels the o/p variable, Rsquared ↑

Memo No.

Date /

If R^2 of the model is 0.50, then approximately half of the observed variation can be explained by the model's input.

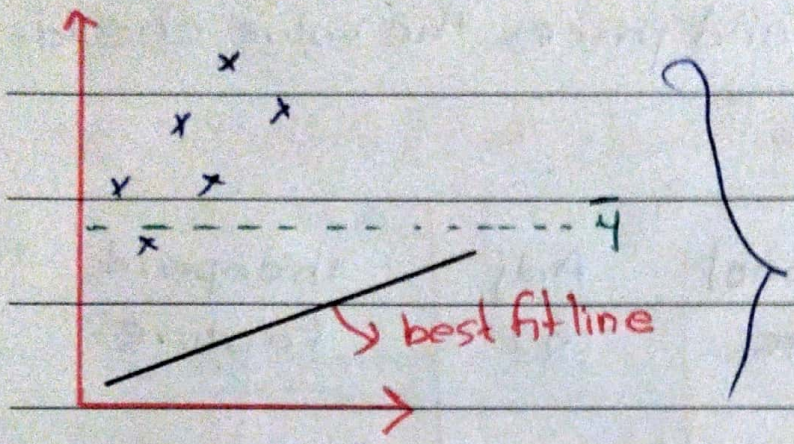
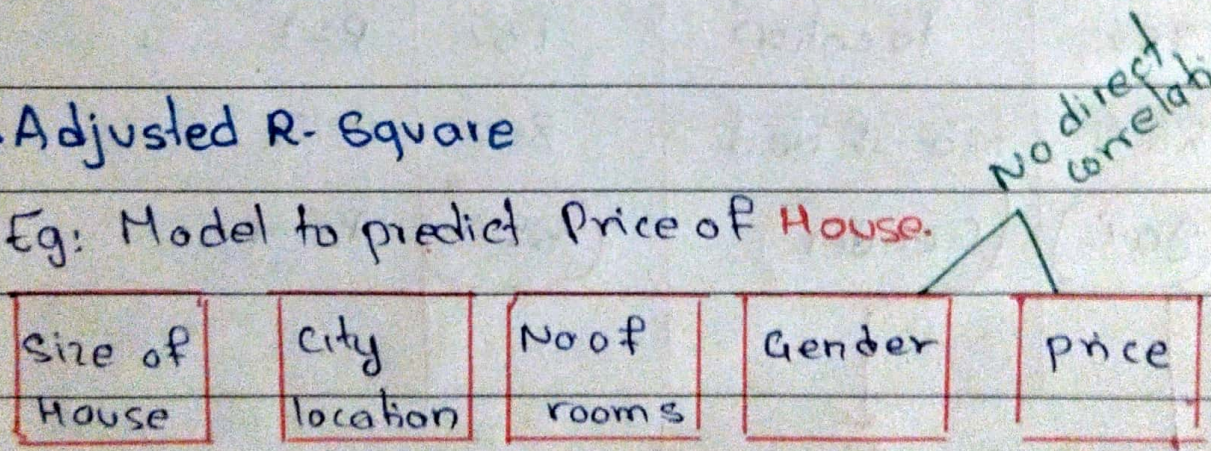


Fig: When R^2 = -ve.

→ model is not working good.

2) Adjusted R-Square

Eg: Model to predict Price of House.



Before, when only size of house was present, as an independent feature it has some R-squared value. After, addition of features like city

Memo No. _____

Date / /

location, No. of bedrooms R^2 value got increased. After, addition of Gender as well, the R^2 value increased but there is no direct correlation between gender and price, the value should-nt be increased.

R^2	Additional features	Adj R^2	Independ feature
65%	size of house	63%	P = 1
75%	location	73%	P = 2
88%	No. of bed	86%	P = 3
90%	Gender	85%	P = 4

No direct relation with price

very slight increase, but this should-nt be happen. To solve this, we use Adjusted R^2.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

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

N = No. of data point

P = No. of independent features

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