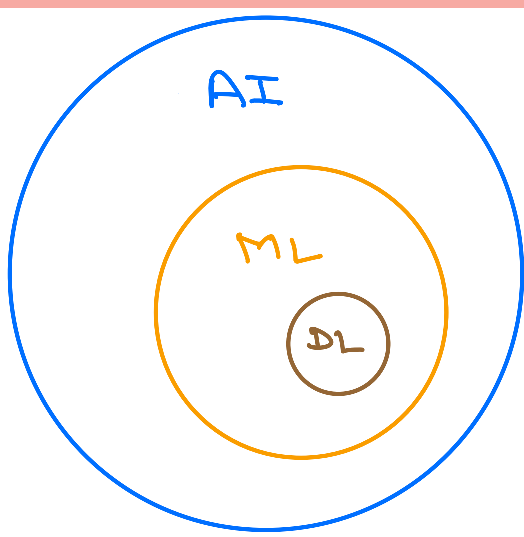
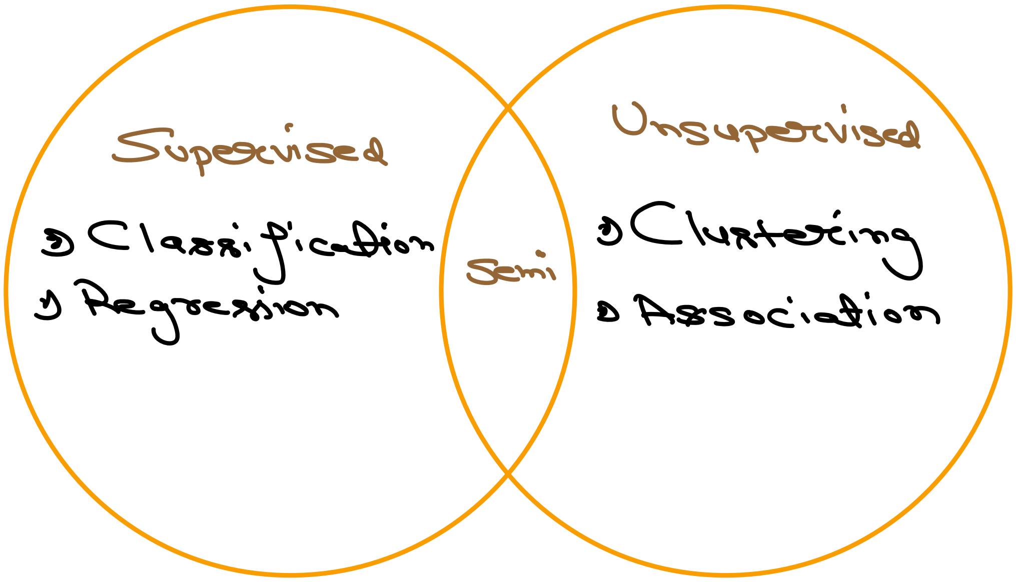
Linear Regression 3

Agenda

- Recap
- Adjusted r2_score
- Need for scaling
- Assumptions of Linear Regression
- Sklearn and statsModel implemention



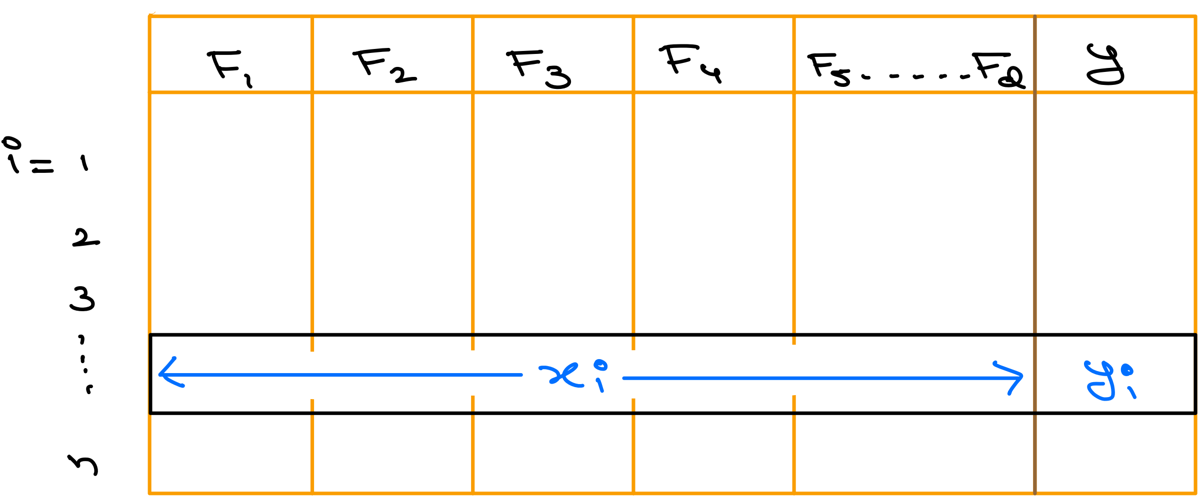
Types of ML



Regression

- Deals with prediction of continous numeric values.

Ex:- Stock price prediction
Car Value prediction



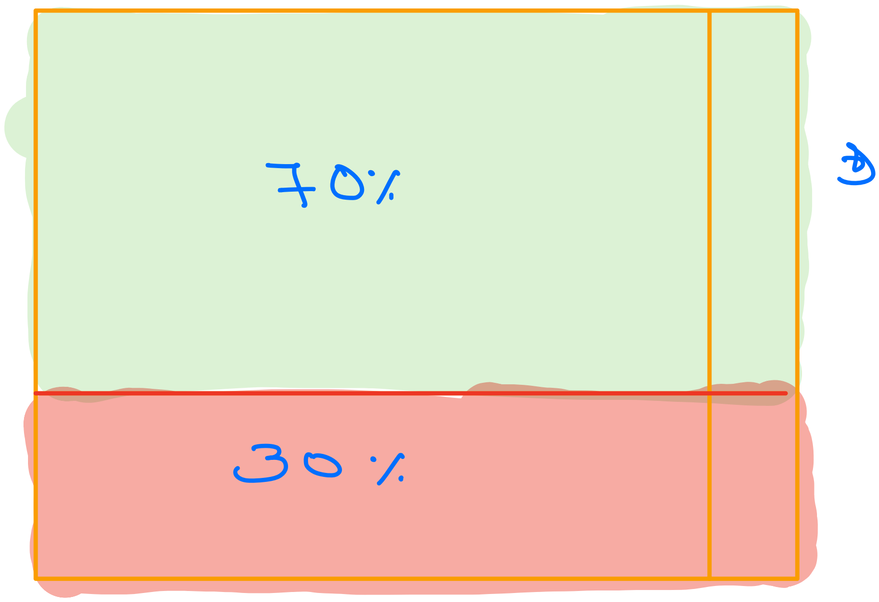
$n \Rightarrow$ no of rows/samples

$d \Rightarrow$ no of features

i^{th} sample $\Rightarrow x_i$

i^{th} label $\Rightarrow y_i$

Train Test Split



Linear Regression

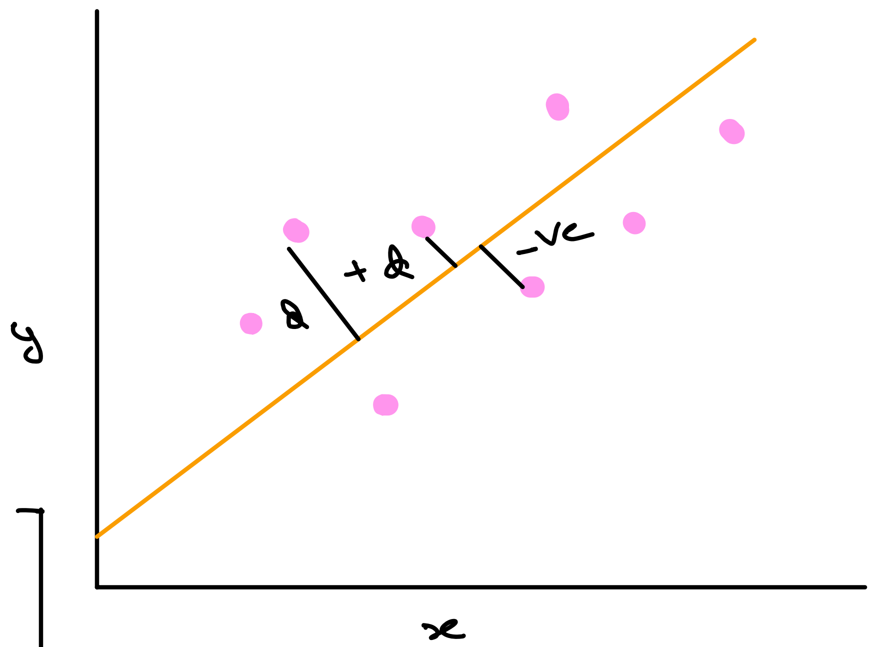
(Ordinary Least Square)

Single Variable L.R.

$$\hat{y} = w_0 + w_1(x)$$

Multi Variate L.R

$$\hat{y} = w_0 + w^T x$$



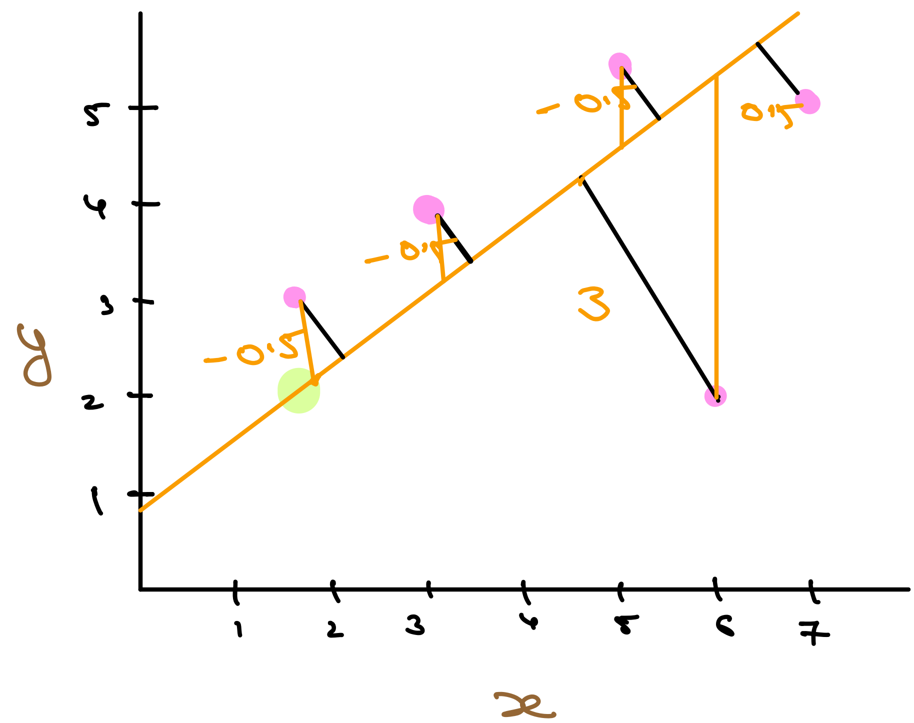
⇒ w is a Vector of d dims $\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$

⇒ x is a Vector of d dims $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix}$

Goal: Find value of w_0 and w_1 for Best Fit line

Error or Residual

i	x	y	$\hat{y}$	$y-\hat{y}$
0	1.5	3	2.5	-0.5
1	3.5	4	3.5	-0.5
2	6	2	5	3
3	5	5.5	5	-0.5
4	7	5	5.5	0.5



$$RSS = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \omega_1 x_i + \omega_0$$

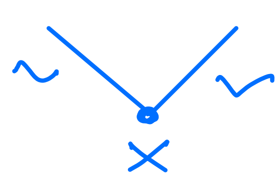
Loss/Cost Functions:

$$MSE \Rightarrow \frac{1}{n} RSS = \frac{1}{n} \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

$$RMSE \Rightarrow \sqrt{MSE}$$

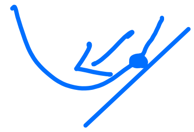
(unit same as y)

$$MAE \Rightarrow \frac{1}{n} \sum_{i=0}^{n} |y_i - \hat{y}_i|$$

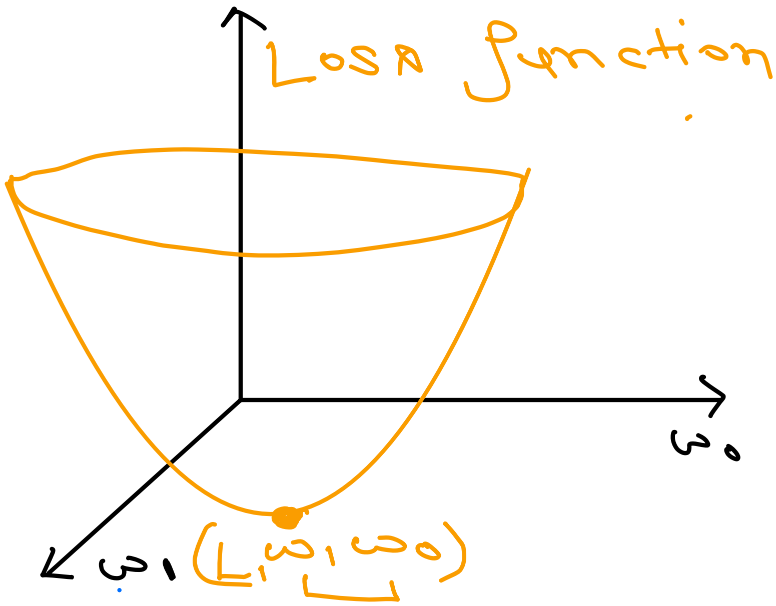


Optimization

$$L = \min_{w, w_0} \frac{1}{n} \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$



$\frac{\partial L}{\partial w_1} \leftarrow \Delta w_1$



$\frac{\partial L}{\partial w_0} \leftarrow \Delta w_0$

$w_1 = w_1 - \alpha \Delta w_1$

η, α > learning rate

How will the eqⁿ change for MLR?

$w_0 = w_0 - \alpha \Delta w_0$

$w_1 \Rightarrow w_1 - \alpha w_1$

$w_2 \Rightarrow w_2 - \alpha w_2$

$w_3 \Rightarrow w_3 - \alpha w_3$

$w_d \Rightarrow w_d - \alpha w_d$

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_d \end{bmatrix}$$

Global minima = Best value of $\langle \vec{w}$ and $w_0 \rangle$

How do we find Global Minima?

Gradient Descent

$w_0 = w_0 - \alpha \frac{\partial L}{\partial w}$ → $L = (y_i - \hat{y}_i)^2$

z^2

$\frac{dL}{dw} \Rightarrow \frac{dz^2}{dz} \times \frac{dz}{dw}$

$w_1 = w_1 - \alpha \frac{\partial L}{\partial w_1}$

$2z \times \frac{dz}{dw}$

⋮

$w_d = w_d - \alpha \frac{\partial L}{\partial w_d}$

Let's calculate partial derivatives manually:

for simplicity Let's assume only 2 features

i.e. $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$

$$L = [y - (w_0 + w_1 x_1 + w_2 x_2)]^2$$

↓

$$\frac{\partial L}{\partial w_0} \Rightarrow \frac{\partial z^2}{\partial w_0} \Rightarrow \frac{\partial z^2}{\partial z} \times \frac{\partial z}{\partial w_0}$$

$w_0 + w_1 x_1 + w_2 x_2$

$z \Rightarrow (y - \hat{y})$

$$\Rightarrow \frac{\partial (y - \hat{y})^2}{\partial (y - \hat{y})} \times \frac{\partial (y - \hat{y})}{\partial w_0}$$

$$\Rightarrow 2 \times (y - \hat{y}) \times \left(\frac{\partial y}{\partial w_0} - \frac{\partial w_0}{\partial w_0} - \frac{\partial w_1 x_1}{\partial w_0} - \frac{\partial w_2 x_2}{\partial w_0} \right)$$

$$\underset{0}{\downarrow} \quad -1 \qquad 0 \qquad 0$$

$$\frac{\partial (y - w_0 - w_1 x_1 - w_2 x_2)}{\partial w_0} \qquad -x_1$$

$$\Rightarrow -2 \times (y - \hat{y})$$

Similarly

$$\frac{\partial L}{\partial w_1} = -2(y - \hat{y}) \times x_1$$

and

$$\frac{\partial L}{\partial w_2} = -2(y - \hat{y}) \times x_2$$

$$\frac{\partial L}{\partial w_d} = -2(y - \hat{y}) \times x_d$$

$$\boxed{\frac{\partial L}{\partial w_0} = \frac{1}{n} \sum_{i=1}^{n} -2(y_i - \hat{y}_i) \qquad \frac{\partial L}{\partial w_d} = \frac{1}{n} \sum_{i=0}^{n} -2(y_i - \hat{y}_i) \times x_{id}}$$

y

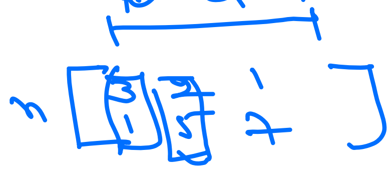
Tricks to Speedup Training

1) Vectorization

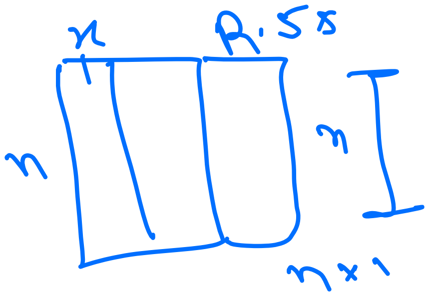
In code we can Vectorize $\frac{\partial L}{\partial w_0}$ using • product

$$\frac{1}{n}\sum_{i=1}^{n} -2\,(y_i-\hat{y}_i)\times x_{id}$$

for $d=1$



$$-2\,(y-\hat{y})\cdot X$$



$n\times 1$ $\qquad$ $n\times d$

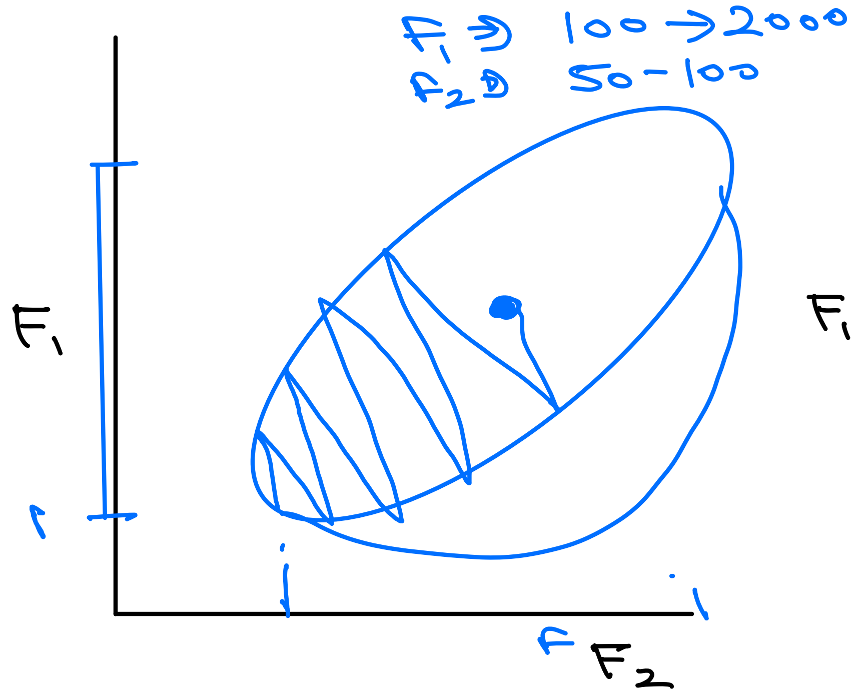
d derivatives over full dataset in single ops

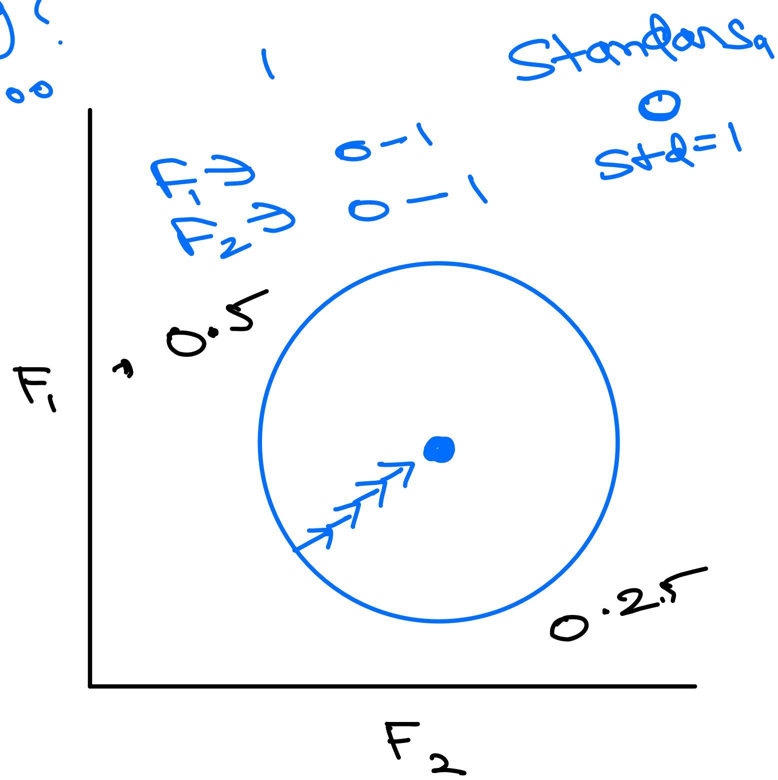
$$X^T\cdot(y_i-\hat{y}) \Rightarrow$$ Replacement of For Loop

$d\times 1$ $\qquad$ $d\times n\cdot n\times 1 \Rightarrow d\times 1$

$$\boxed{\frac{\partial L}{\partial w} \Rightarrow -\frac{2}{n}\left(X^T\cdot(y-\hat{y})\right)}$$

② Scaling why?



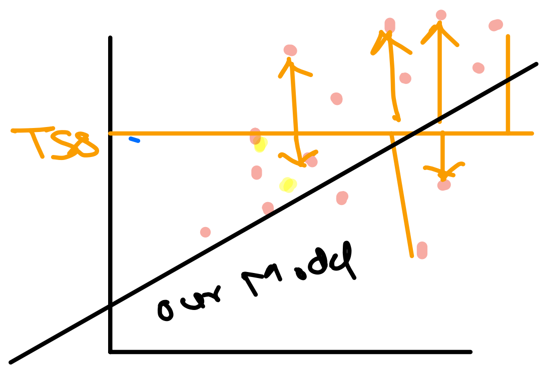


→ Scaling Ensures No feature dominates

Metric for Evaluating performance of Linear Regression

Q: What does RSS or MSE of 10000 mean?

$$R2\_score = 1 - \frac{RSS}{TSS}$$



$$\sum_{i=0}^{n} \left(y_{mean} - y_i\right)^2$$

⇒ How good is LR model from mean Model

Q Range of R-2 Score?

$(-\infty, 1]$

Q Problem with R-2 Score

Case 1 : d-feature + 1 more relevant feature

r2_score ⇒ ↑ increase

Case 2 : d - feature + 1 irrelevant feature

r2_score is same

↓ x_d

$w_d = 0$ ↑

Q : How to mitigate this

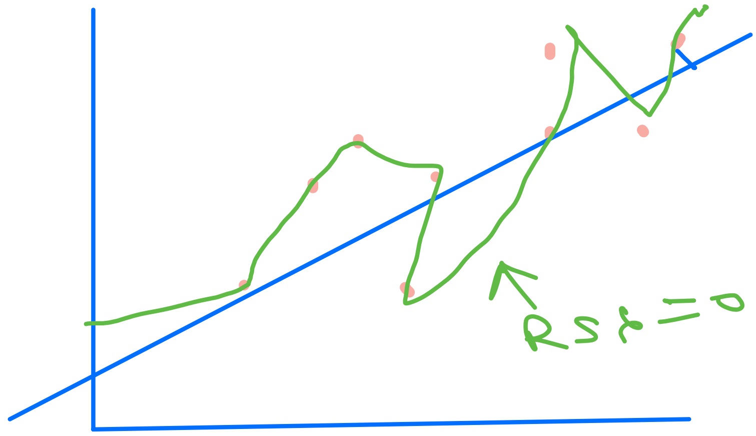
$$\text{adj\_r2\_score} \Rightarrow \left[1 - \frac{(1 - \text{r2\_score}) \times (n-1)}{n - d - 1} \right]$$

n ⇒ no of row

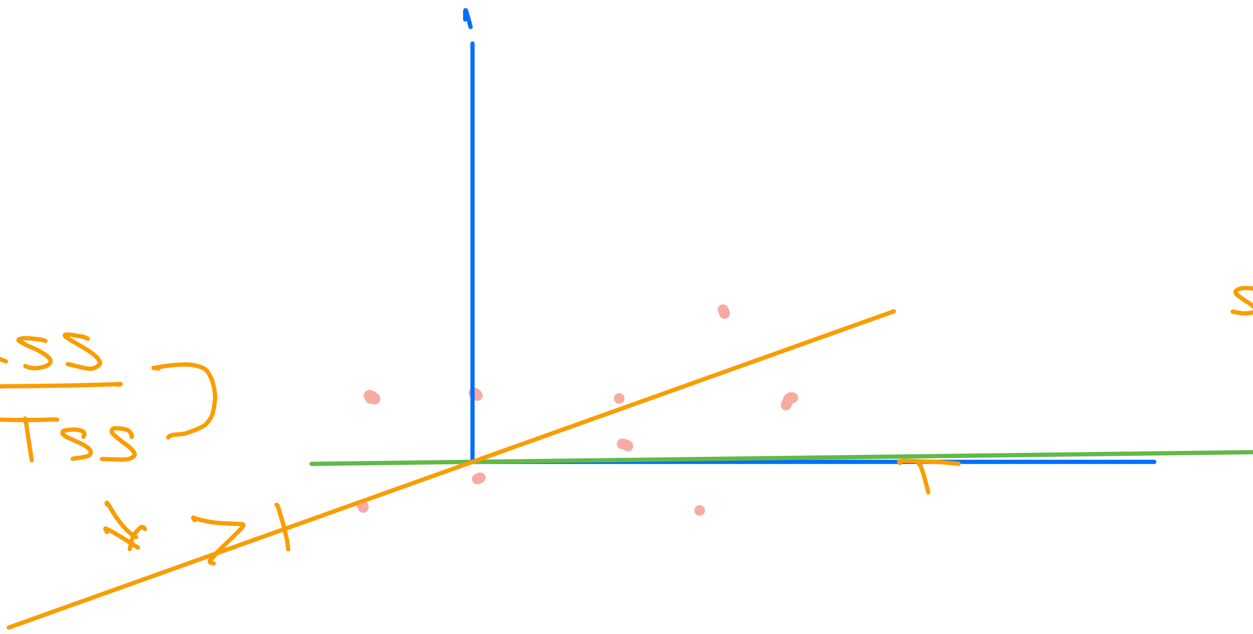
d ⇒ no of features

Case 1 ⇒ adj_r2_score ↑

Case 2 ⇒ adj_r2_score ↓



RSS



$1 - \frac{RSS}{TSS}$

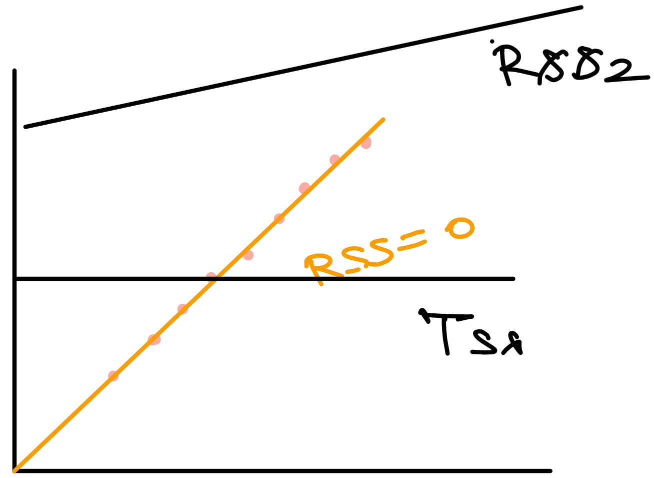
$X > 1$

r2_score ⇒ -ve

i

1 d

xlj2



$r2 = 1 - \frac{RSS}{TSS} \Rightarrow 1$

-4000

-2000

⇒ Transformer

Generative AI

Text audio Video Image

⇒ diffusion Models