

## ML cont

### Agenda

1. Ridge
2. Lasso
3. elastic net
4. LoR

$$\mathcal{E} = \frac{1}{2m} \sum (y_i - \hat{y}_i)^2$$

$$\mathcal{E} =$$

## Adjusted $\sigma^2$

If we add more columns, even random unrelated col<sup>n</sup>,  $\Rightarrow$  variance is gonna increase  $\Rightarrow \sigma^2 \uparrow \uparrow$

temp	america	study	Sleep	Score
53F				
60F				
78F				

$\Downarrow$   
 $\sigma^2$

the score get increased or even can stay as it is which is not an accepted behaviour.

## Adjusted $\sigma^2$

$$\sigma_{adj}^2 = 1 - \left[ \frac{(1-\sigma^2) \frac{(n-1)}{\text{no. of rows}}}{(n-1-R) \frac{\text{no. of indep. features}}{\text{no. of rows}}} \right]$$

1) irrel. feature like say temp. added

$R \uparrow \uparrow$

$$\Rightarrow (n-l-R) \downarrow \downarrow \quad (n-l) \text{ const}$$

$$1 - \left( \frac{(1-\gamma^2)(n-l)}{n-R-1} \right) \downarrow \downarrow \downarrow$$

$\gamma^2 \uparrow$   
 $\Rightarrow (\text{const}) \text{ dec.}$        $\Rightarrow \text{const.}$

$$\frac{N}{\Phi} \quad \Phi \downarrow \downarrow \quad \Rightarrow \frac{N}{\Phi} \uparrow \uparrow$$

$$\Rightarrow 1 - \text{large oval} \quad \uparrow \uparrow$$

$\Rightarrow$  adjusted  $R^2$  gonna decrease.

2. Some relevant feature say Sleep hours

$$1 - \left[ \frac{(1-\gamma_2)(n-1)}{n-1-\gamma_2} \right]$$

$\gamma_2 \uparrow \uparrow \uparrow$   
 $\downarrow \downarrow \downarrow$   
 $\xrightarrow{\text{Want}}$

dec in  $(1-\gamma_2)$  >> dec in  $(n-1-\gamma_2)$

$$\frac{12}{4} = 3 \quad \frac{8}{3} = 2.66$$

$$\Rightarrow 1 - \text{circle} \downarrow \downarrow \downarrow$$

$$\Rightarrow \gamma_{adj}^2 >>$$

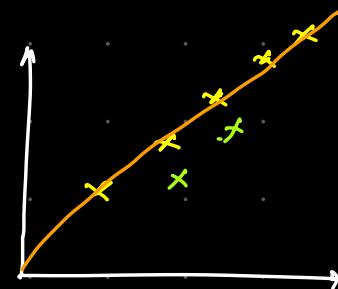
## Assumption of - LR

1. Linear relation b/w data point
2. Indep. feature should have normal distribution.
3. Take care of multicollinearity
4. Homoscedasticity
5. Heteroscedasticity
6. Feature scaling

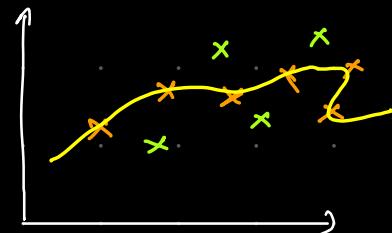
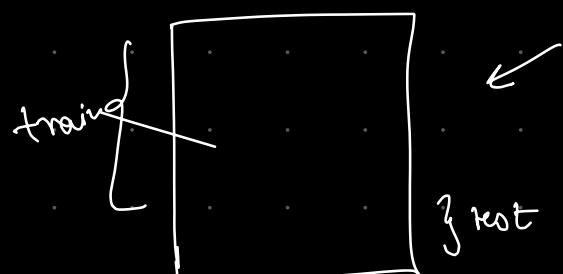
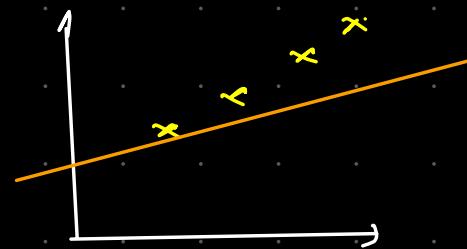
## 1. Overfitting in Alg Θ:

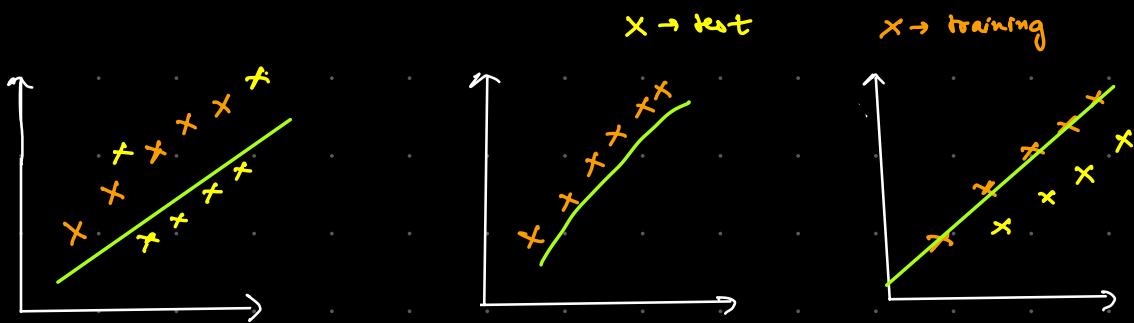
$$5+4 = 9 \quad \checkmark$$

$$5+2 = 7 \times$$



## 2. Underfitting





M1

underfitting

M2

$\Rightarrow$  Bias  $\Rightarrow$  Training data

$\Rightarrow$  Variance  $\Rightarrow$  Test data

M3

overfitting

Underfitting  $\Rightarrow$  training error  $\uparrow \uparrow$

$\Rightarrow$  high bias

M1

high error in training data

$\Downarrow$

bias

Now for test data, we can have

2 combination.

1. low variance

2. high variance

M<sub>2</sub>

low bias  $\Rightarrow$  training error

test data  $\Rightarrow$  low variance

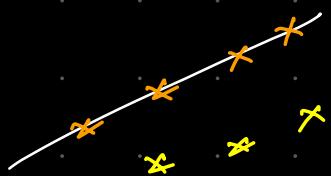
no under fitting & no overfitting  
 $\Rightarrow$  accepted / required model

M<sub>3</sub>

low bias

error in test ↑↑

$\Rightarrow$  high variance

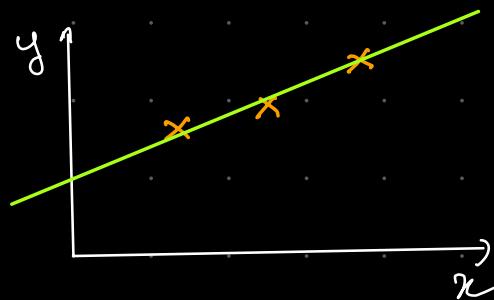


low bias + high variance ✓

So the aim should be to get a  
model like M<sub>2</sub> which is generalized

$\Rightarrow$  low bias + low variance

## Ridge regression



$$J/\epsilon = \sum (y_i - \hat{y}_i)^2$$

$J \rightarrow 0$  in case of overfitting

Low bias + high variance  $\rightarrow$  like M3

How to fix this?

$$\epsilon_{\theta_0, \theta_1} = \sum (y_i - \hat{y}_i)^2 + \lambda (\text{slope})^2$$

We have added a penalty

$\lambda \rightarrow$  hyperparameter

slope  $\rightarrow$  coefficient

What exactly will this do?

$$\lambda = 1$$

Our aim is to reduce cost  $J^n$   
to global minima.

Overfitting  $\Rightarrow J = 0$

$J \rightarrow 0 \Rightarrow$  line is fixed

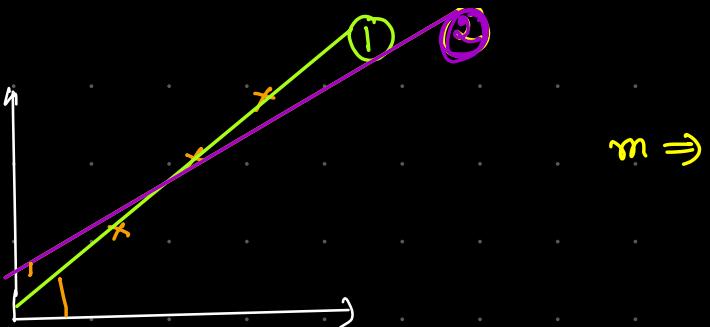
In ridge regression we are adding a component.

$$\lambda(\text{slope})^2$$

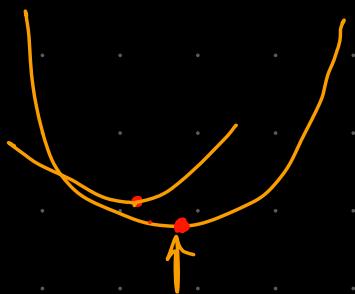
When error is 0 & say  $\lambda = 1$   $m = 15$

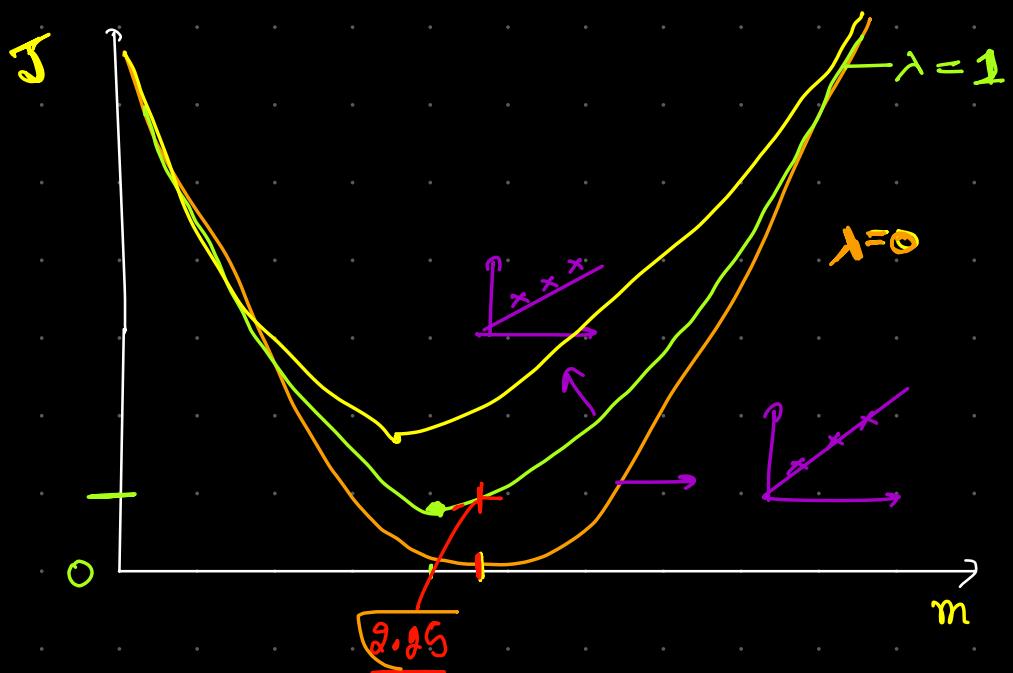
$$J = \boxed{0} + \underline{(2.25)}$$

$\Rightarrow$  model will have to train more  
as we need to reach global min<sup>n</sup>

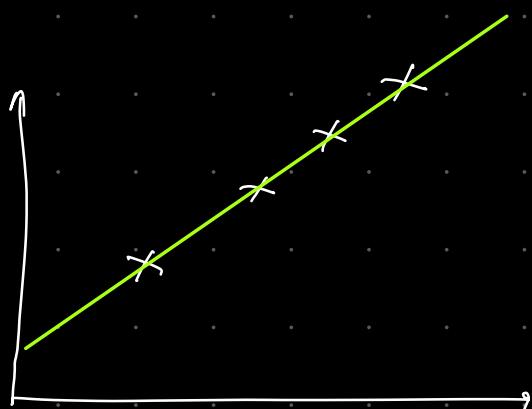


$$\begin{aligned}
 J &= \text{small value} + \frac{1}{(0.8)^2} \\
 &= \boxed{\text{small value} + 0.64} < \underline{2.25} \\
 &\quad \uparrow \\
 &\quad \text{Actual min}
 \end{aligned}$$





$$\frac{(y_i - \hat{y}_i)^2}{\text{Small value}} + \lambda (\text{slope})^2 \xrightarrow{\lambda = 1 (0.8)} (6.64 \pm \text{small}) \ll 2.25$$



relation b/w  $\lambda$  & slope?

$\lambda \uparrow \uparrow$

slope  $\downarrow \downarrow$

$\Rightarrow$  ridge regression helps us in  
avoiding overfitting

$\lambda \Rightarrow$  hyperparameter

## Lasso Regression [ $L_1$ regularization]

Only cost function we will be changing

$$J = \frac{1}{2m} \sum (y_i - \hat{y}_i)^2 + \lambda (\text{slope})$$

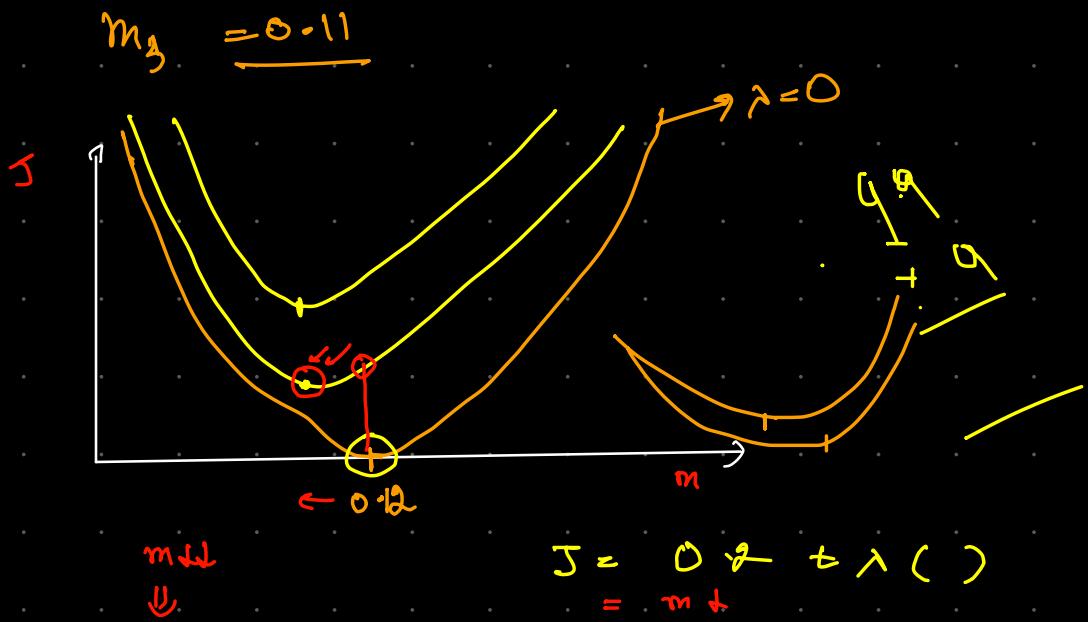
$\Rightarrow$  feature selection

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + C$$

$$= m_1(\text{study}) + m_2(\text{sleep}) + \frac{m_3(\text{temp})}{\text{temp}} + C$$

$m_3 = 0.11 \Rightarrow$  it means for change  
of temp just  
0.11 unit change





Global minima gets shifted

Ridge  $\rightarrow$  overfitting

Lasso  $\Rightarrow$  feature selection.

Elastic Net  $\frac{(\ell_1 + \ell_2)}{\text{lasso + ridge}}$

$$\Sigma = \sum (y_i - \hat{y}_i)^2 + \frac{\lambda_2 (\text{slope})^2}{\text{ridge}} + \frac{\lambda_1 |\text{slope}|}{\text{lasso}}$$

$\lambda_1$  &  $\lambda_2$  are different

$\Rightarrow$  reducing overfitting & feature selection

When to use ridge and lasso?

Multicollinearity

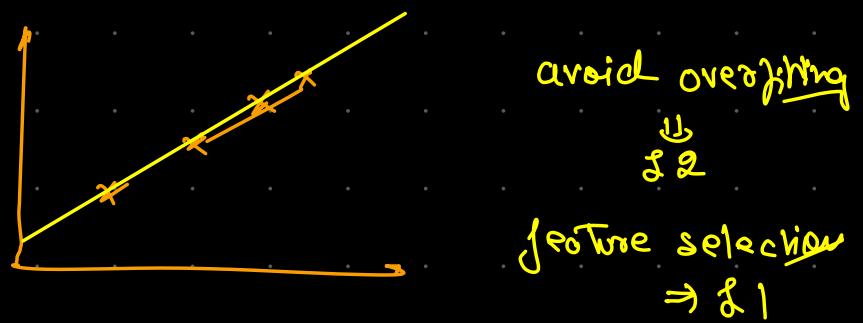


f<sub>1</sub>, f<sub>2</sub> & f<sub>3</sub> are highly correlated.

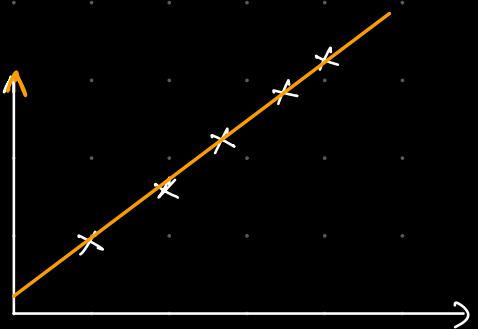
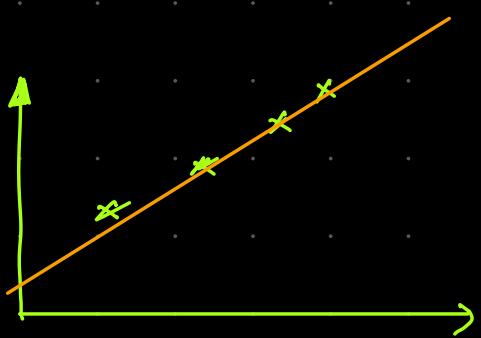
feature selection  
(L1)

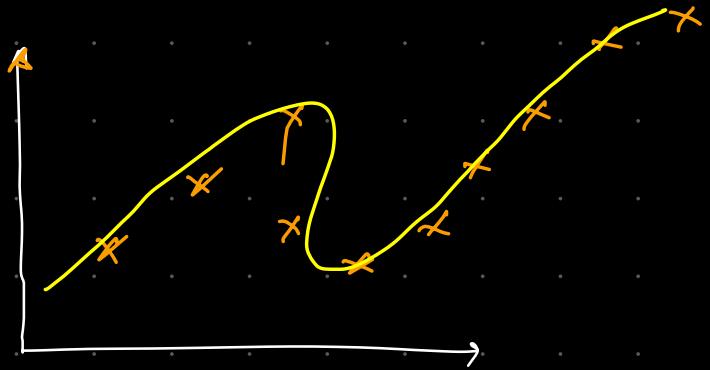
lot of similar feature  
 $\Rightarrow$  overfitting

800+



⇒ Elastic net



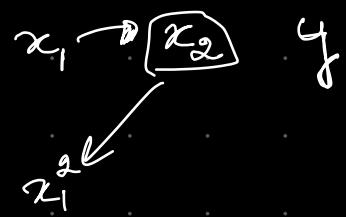


$$x_1^2 \leftarrow x_1^{\text{pred}} \quad y \Rightarrow \text{result}$$

$x_1$



$$\begin{aligned}
 x_f &= m_1 x_1 + m_2 x_2 + c \\
 &= m_1 x_1 + m_2 x_2 + m_3 x_3 + c \\
 &= x_1^2 \\
 &= m_1 x_1 + m_2 x_1^2 + c + m_3 x_3
 \end{aligned}$$



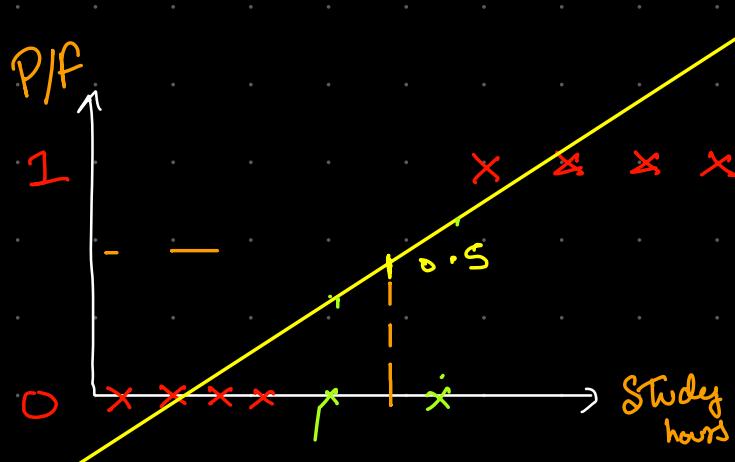
$$y = mx + c$$

$$y = m_1 x_1 + m_2 \underline{x_2} + c$$

$$y = m_1 x_1 + m_2 x_2^2 + c$$

# Logistic Regression [Classification]

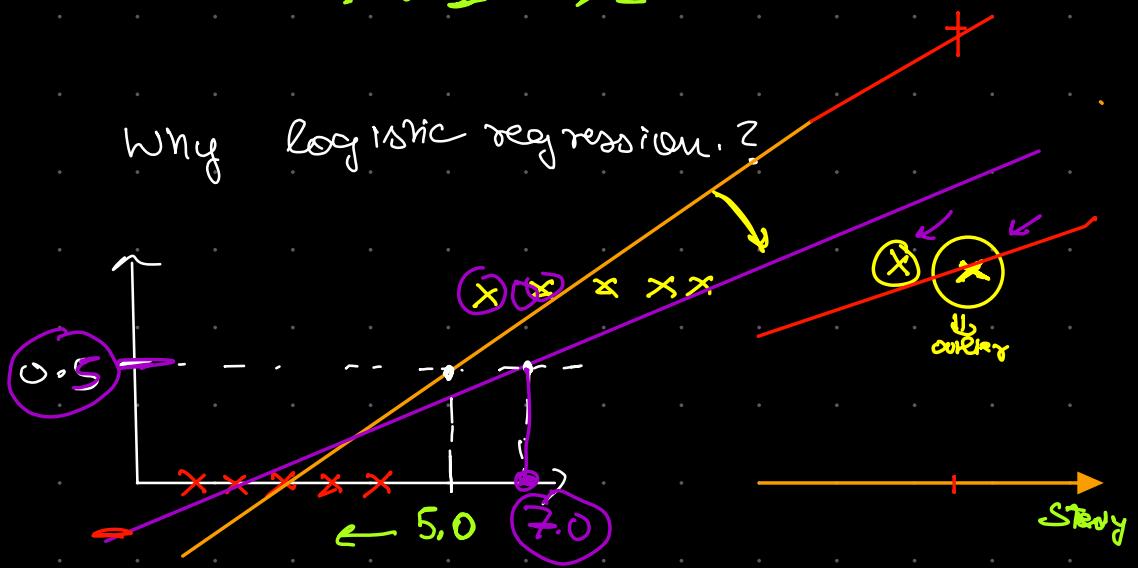
Binary [1]



$$0.5 < \Rightarrow 0$$

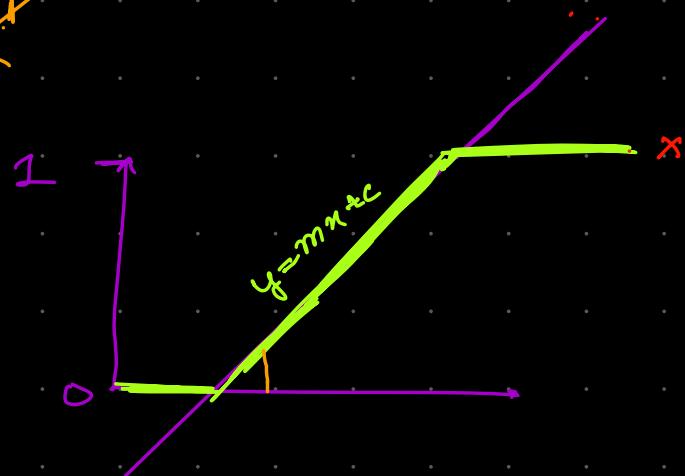
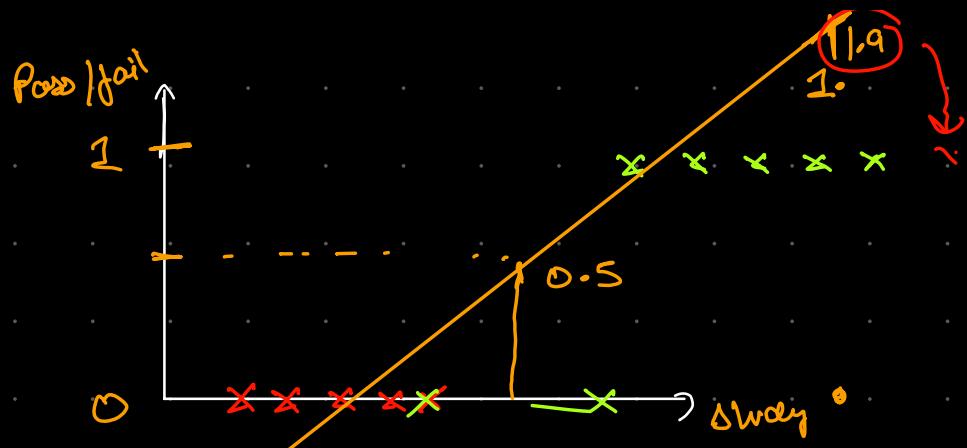
$$> 0.5 \Rightarrow 1$$

Why Logistic regression?



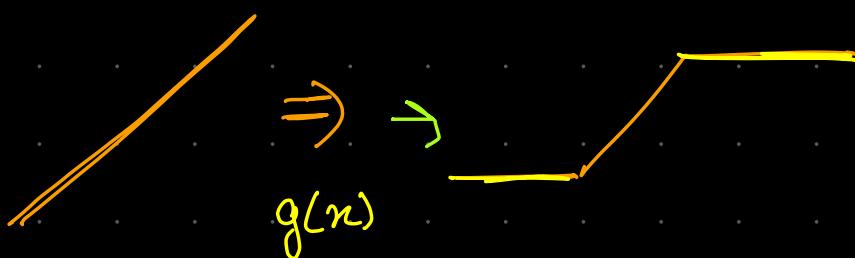
in training data

Outliers moved our line a lot



Our line equation is gonna be still

$$y = \underline{mx + c}$$



$$x \xrightarrow{f(x)} x^2$$

$$f(x) = y = \left( \begin{matrix} mx + c \\ \Downarrow \\ z \end{matrix} \right)$$



$$\underline{f(n)} = g(z)$$

$g(z)$  = sigmoid or activation  
function

$$g(z) = \frac{1}{1 + e^{-z}} \quad z = mx + c$$

$\curvearrowleft$

activation function

linear regression

$$y = mx + c$$

$$\varepsilon = ( )^2$$

$$y = \frac{1}{1 + e^{-(mx + c)}}$$

$$\varepsilon =$$

$$y = mx + c \propto x$$

$$y = \frac{1}{1 + e^{-(mx+c)}}$$

$$y \underset{(m, c)}{\sim} \frac{1}{1 + e^{-(mx+c)}}$$

$$y = \{0, 1\}$$