* Logistic Regression

Supervised learning

Regression | Classification → Logistic Regression

eg: # of hours studied → Predict Pass or fail

X (# hours)	Pass/Fail (1 / 0)
2	0
8	1
3	0
4	0
6	1

Train → model

New data → model → 1/0

Acc ↑

eg: Predict if a person will fraud / not fraud.

Salary (K)	Age (years)	Fraud (1,0)
20	19	1
22.5	20	1
30	30	0
-	-	-
-	-	-

50K, 20 → model → 1/0

eg: Predict diabetic/not

Cholesterol level	diabetes (1/0)
100	1
250	1
80	0
70	0

85 → model → 1/0

eg: Cancer | Not cancer based on size of tumor

Size of tumor	Cancer / Not Cancer
1.2	0
3.2	1
4.2	1
-	0
-	-
-	

eg: Spam | ham

Classification → binary classification → o/p → 2 categories

Classification → multiclass classification → o/p > 2 categories

Age	Salary	Credit Score
20	50k	Good
25	30k	Fair
20	31k	Good
30	50k	bad

2 1 0

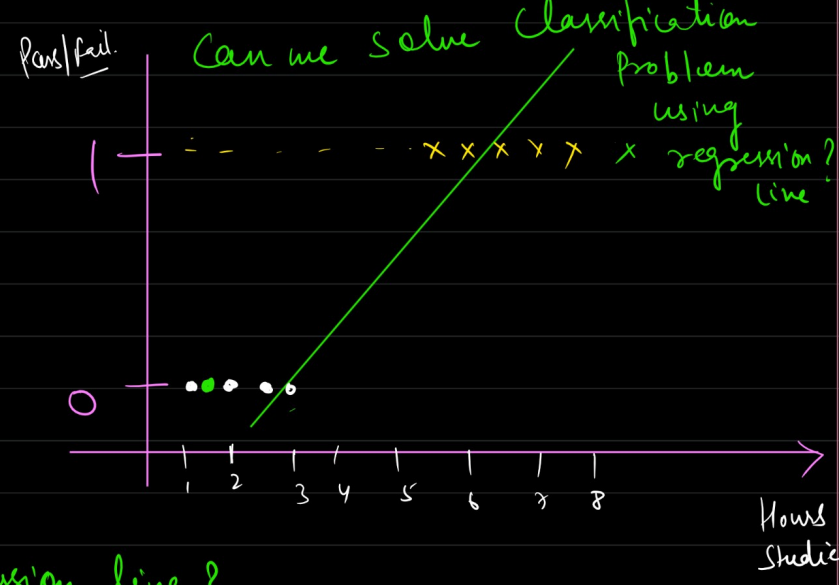
Good, Fair, bad

(multi class classification)

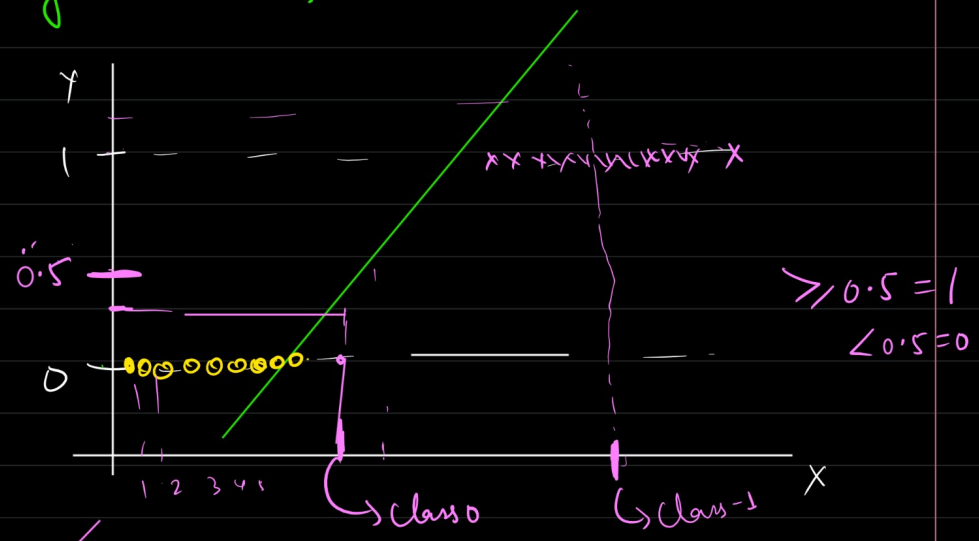
* There are different Algos to solve Classification Problem

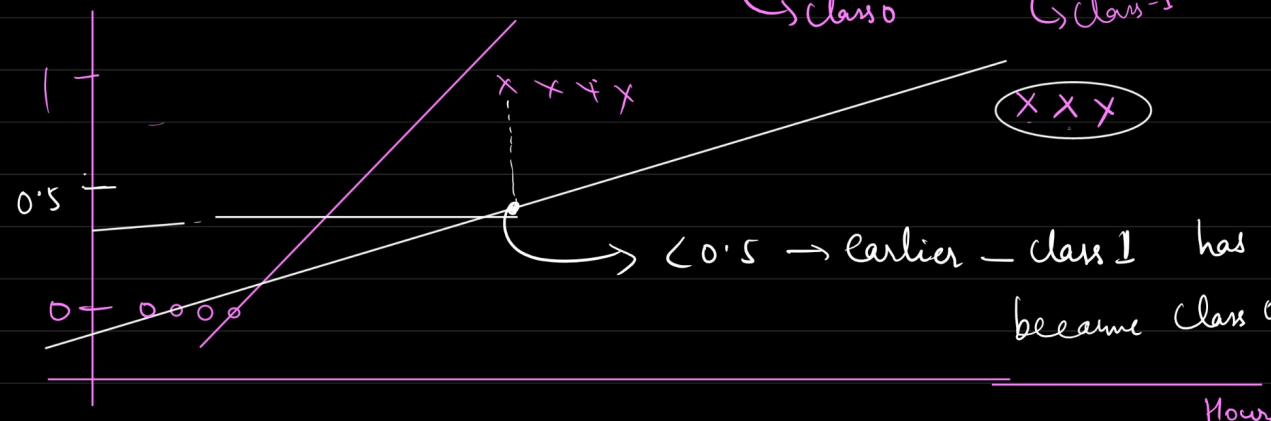
* Logistic regression

X (# hours)	Pass/Fail
2	0
8	1
3	0
4	0
6	1



Q. Why we cannot use regression line?





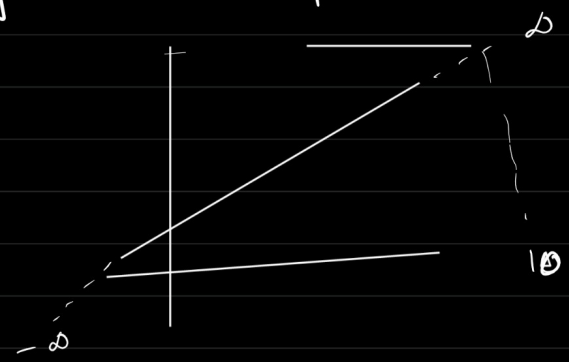
① Best fit line will change due to presence of outliers.

② Regression line range is $-\infty$ to ∞

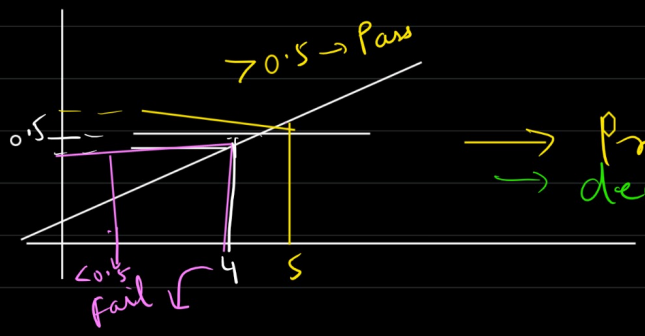
↓ 0 to 1 ↓

⇓

0 to 1

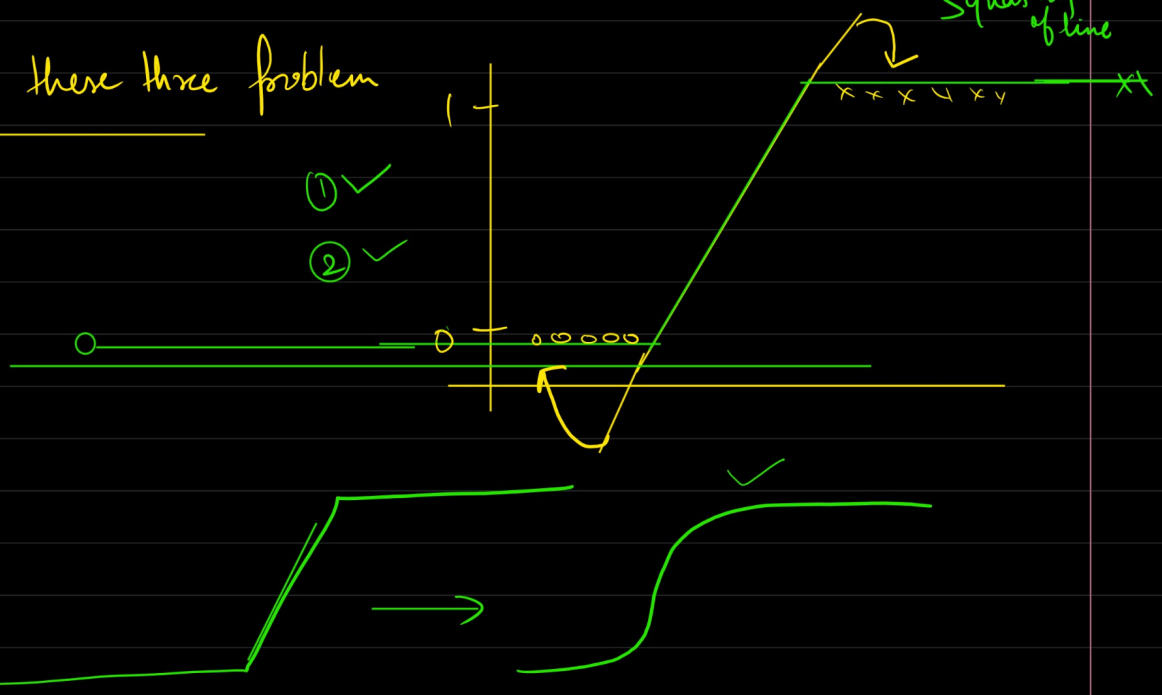


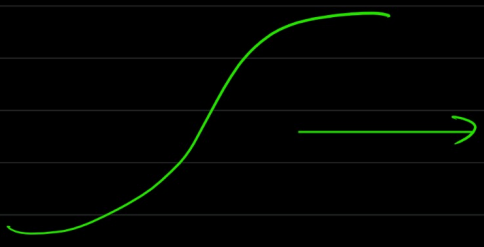
③



→ Prediction changes suddenly
→ decision boundary is not changing gradually.

* How to solve these three problem



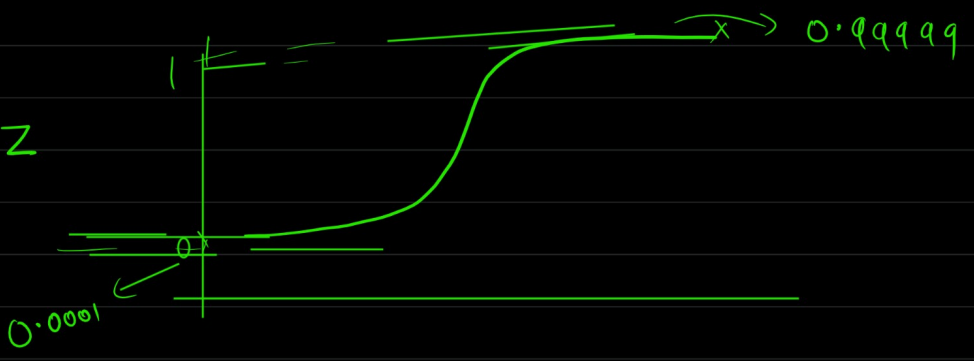


⟹ Sigmoid fn.

⇓

logistic function

$$\sigma = \frac{1}{1+e^{-z}}$$



0 to 1

$-\infty$ to ∞ $\longrightarrow$ 0 to 1

$$\sigma = \frac{1}{1+e^{-z}}$$

$$z = -\alpha \Rightarrow \frac{1}{1+e^{-(-\alpha)}} = \frac{1}{1+e^{\alpha}} = \frac{1}{1+\alpha} = \frac{1}{\alpha} = 0$$

$$z = \alpha \Rightarrow \frac{1}{1+e^{-\alpha}} = \frac{1}{1+\frac{1}{e^{\alpha}}} = \frac{1}{1+\frac{1}{\alpha}} = 1 \quad \left(\frac{1}{\alpha} \to 0\right)$$

$3^{-1} = \frac{1}{3}$

$h_\theta(x) = \theta_0 + \theta_1 x_1 \longrightarrow$ Best fit line

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

σ ↓ Logistic fun

$(\theta_0 + \theta_1 x)$ ↓ z

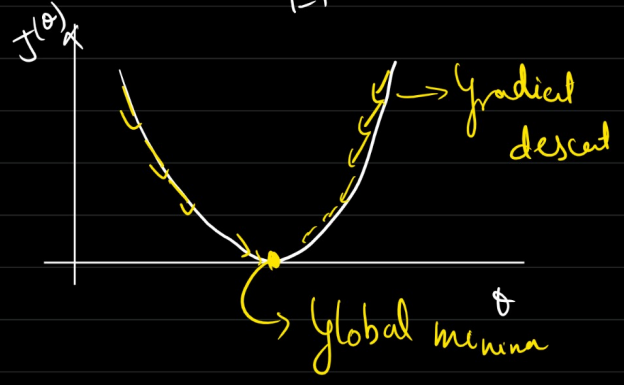
$$\boxed{h_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x_1)}}}$$

→ Logistic Regression model

* To get optimal θ_0 & θ_1, minimise the cost function.

Linear Regression model

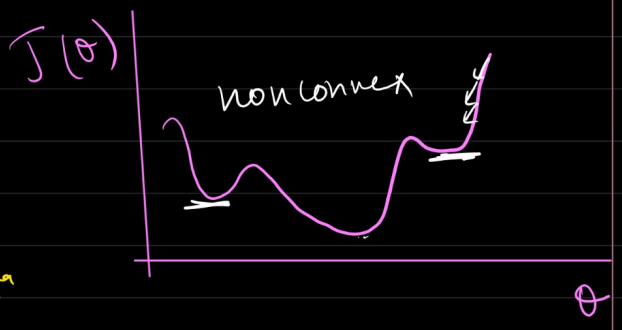
$$J(\theta_0,\theta_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x)_i)^2$$



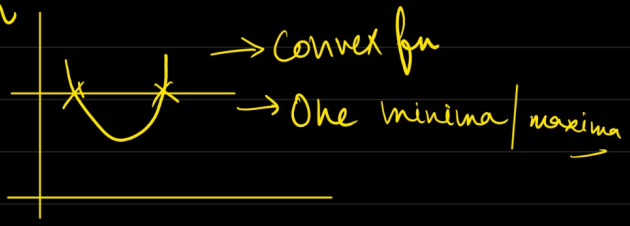
logistic Regression

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

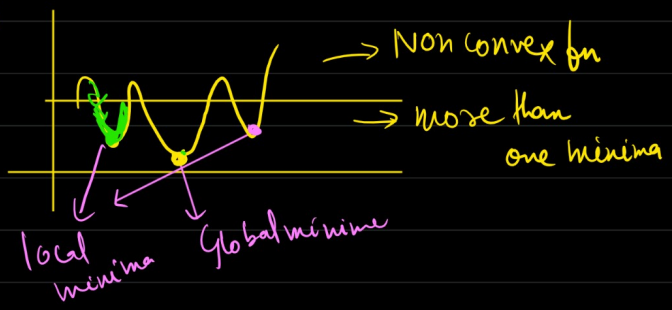
$$h_\theta(x_i) = \frac{1}{1+e^{-\theta_0+\theta_1 x_1}}$$



* Convex fun



* non convex



log loss function

$$J(\theta_0,\theta_1) = -y_i \log(h_\theta(x)_i) - (1-y_i)\log(1-h_\theta(x))$$

where $h_\theta(x)_i = \frac{1}{1+e^{-\theta_0+\theta_1 x}}$

y_i → actual value

$$J(\theta_0, \theta_1) = -y_i \log(h_\theta(x)_i) - (1-y_i)\log(1-h_\theta(x))$$

$$J(\theta_0, \theta_1) = \begin{cases} -\log(h_\theta(x_i)) & \text{if } y_i = 1 \\ -\log(1-h_\theta(x)_i) & \text{if } y_i = 0 \end{cases}$$

To minimize the cost function $J(\theta_0, \theta_1)$ by changing θ_0 & θ_1

Convergence Algorithm

Repeat until convergence

$$\{ \quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta} J(\theta_0, \theta_1)$$

To get optimal θ_0 & θ_1

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1)}}$$

for multiple variable

$$h_\theta(x) = \frac{1}{1+e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n)}}$$

multivariate logistic regression