Logistic Regression. (Classification Problem)

eg.

Study	Play	Pass/Fail.
1	8	Fail
2	7	Fail.
3	7	Fail
6	3	Pass
1	4	Pass.

Classification.

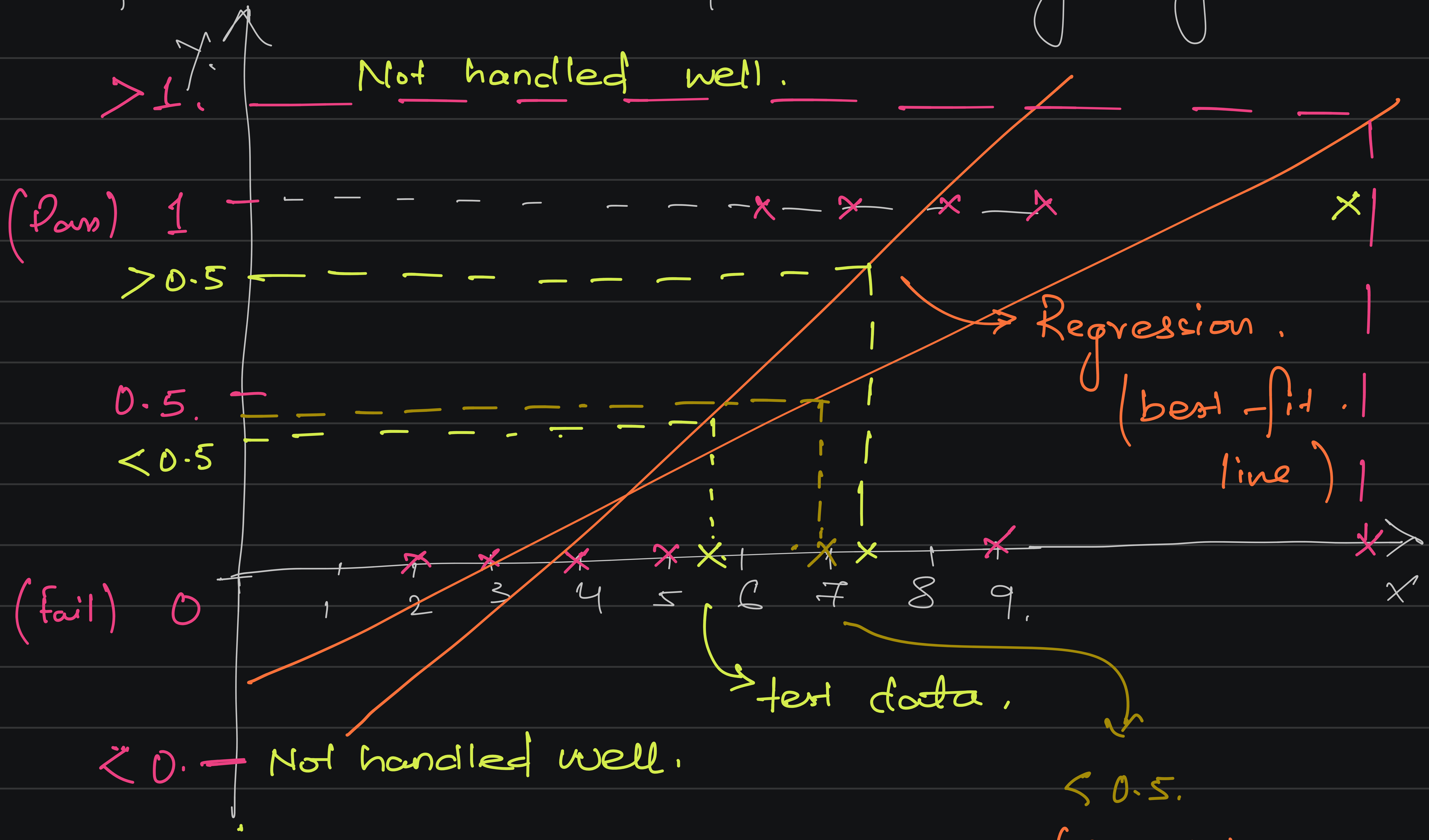
Pass. → Outlier.

eg. Study hours → independent; Pass/Fail. → output.

Study hours.	Pass/Fail.
2.	Fail
3	Fail
4	Fail
5	Fail
6	Pass
7	Pass
8	Pass
9	Fail

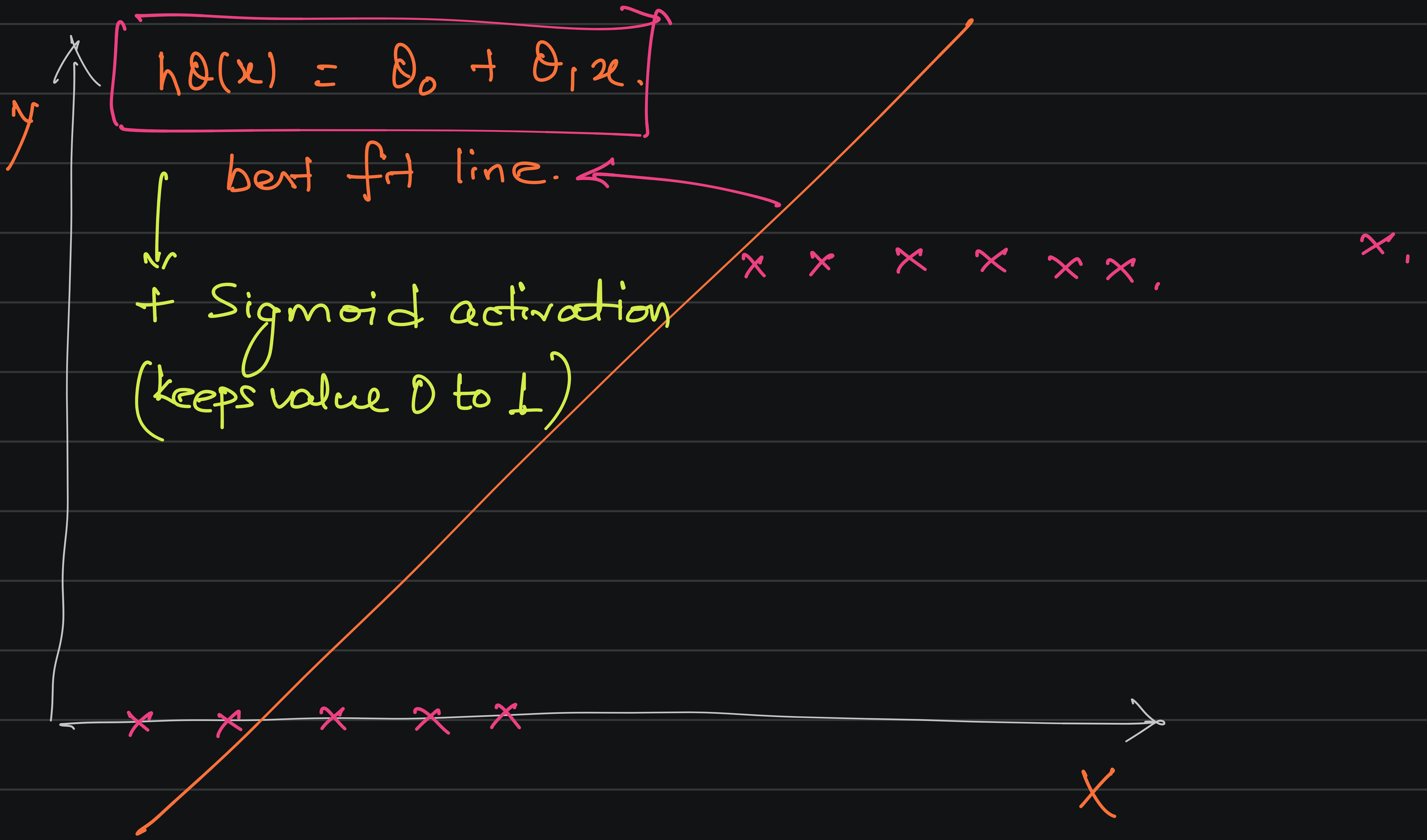
Fail → Outlier.

Q Can we solve this problem using regression?



$y \leq 0.5 = 0.$

$y \geq 0.5 = 1.$



1. $Z = h_\theta(x) = \theta_0 + \theta_1 x$ * Create best fit line.

2. Sigmoid fn $= \left[\dfrac{1}{1+e^{-z}}\right]$

* Squashing

Always value $= 0$ to 1.

3.

Linear regression cost fn.

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

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

MSE.

1 global minima. ← Convex function.

Logistic regression cost fn.

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

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

Sigmoid activation added.

Let $z = \theta_0 + \theta_1 x$

$$h_\theta(x) = \sigma(z)$$

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

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

$$\Downarrow$$

0 to 1.

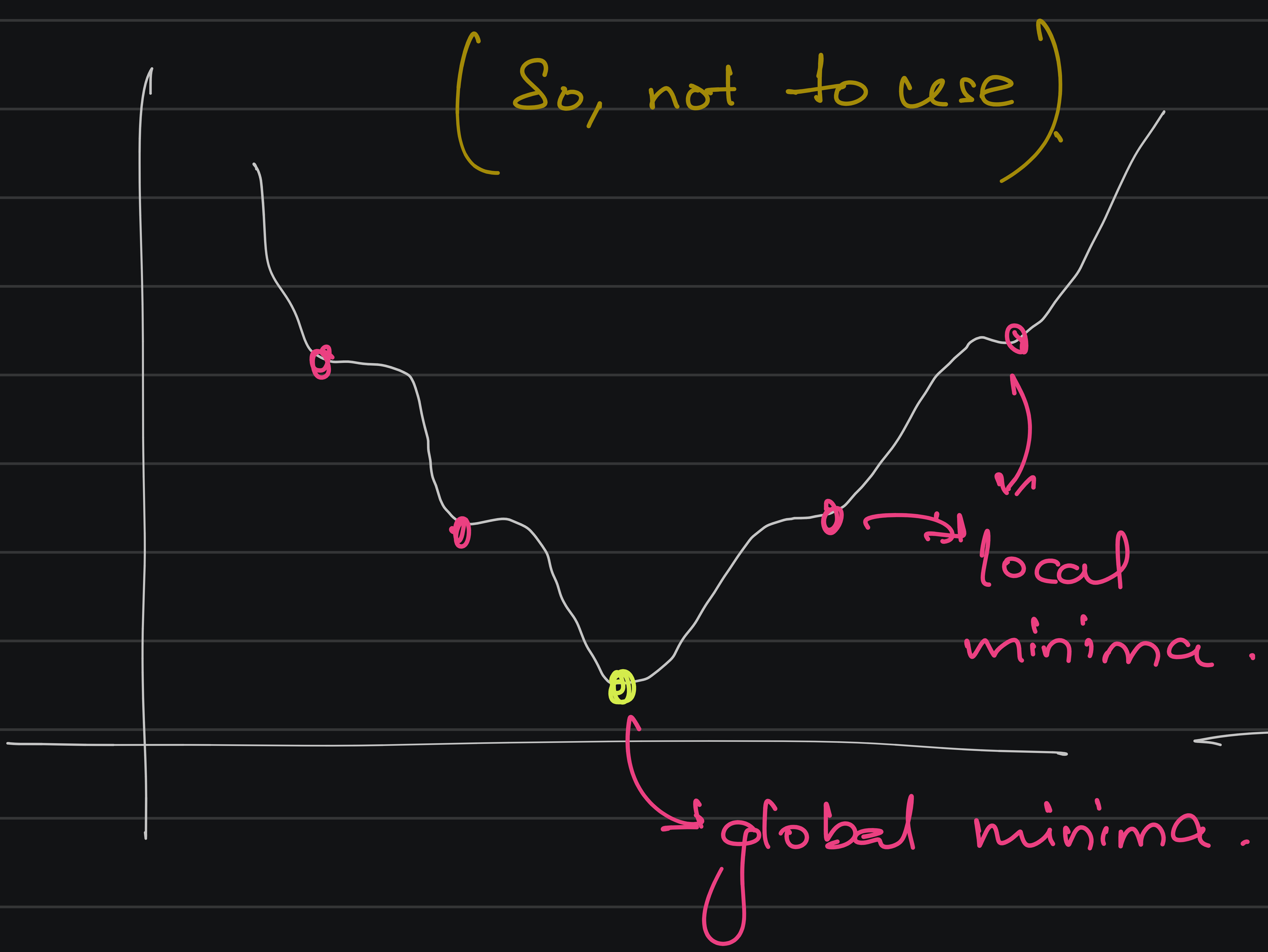
threshold.

if $\leq 0.5 \Rightarrow 0$,

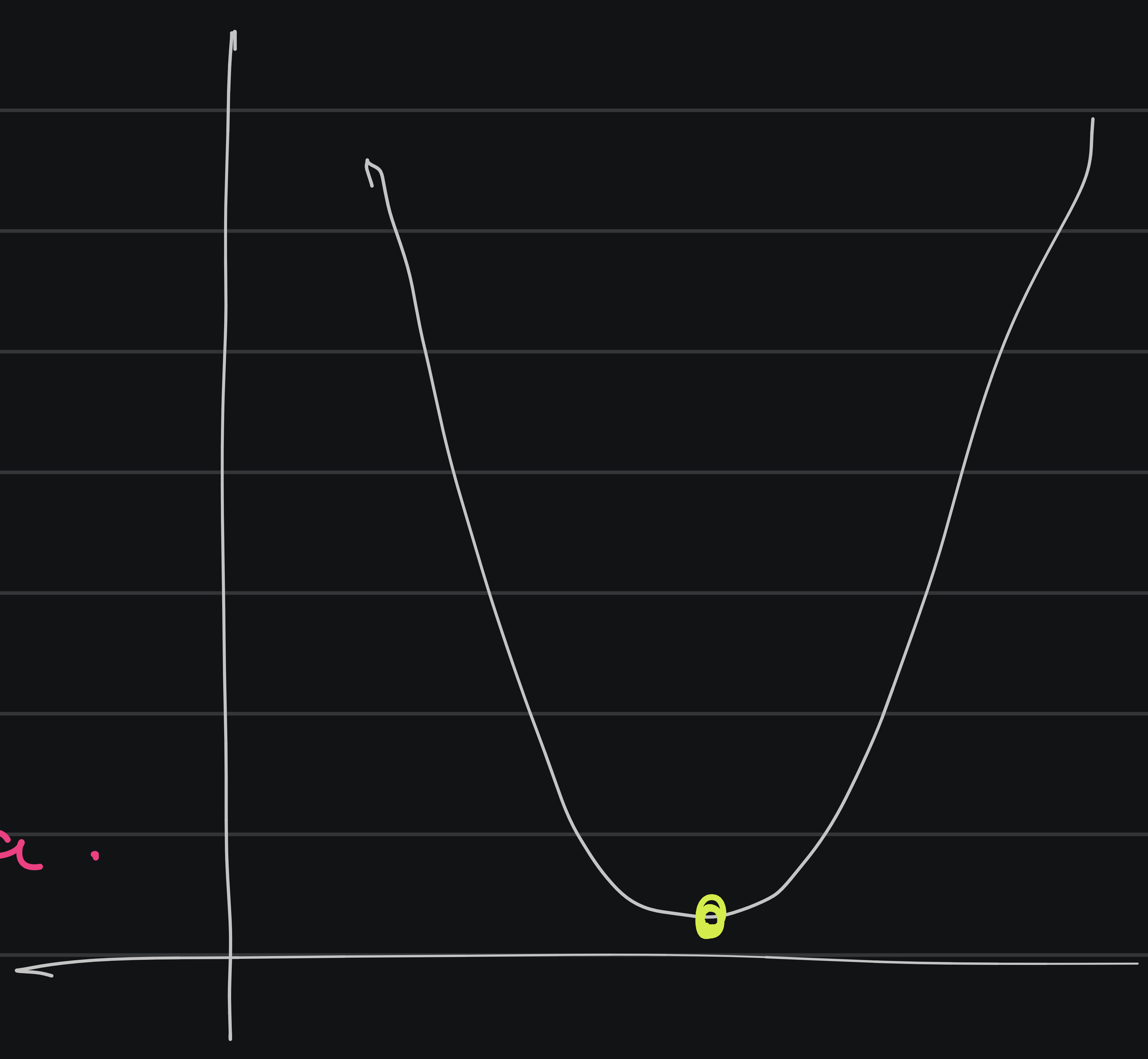
$> 0.5 \Rightarrow 1$

Leads to

Non-convex fn.



Convex fn.



(*) Log loss cost fn.

$$Cost(h_\theta(x)^i, y^{(i)}) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1. \\ -\log(1 - h_\theta(x)) & \text{if } y = 0. \end{cases}$$

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

substitute to get above values.

$$\boxed{Cost(h_\theta(x)^i, y^{(i)}) = -y \log(h_\theta(x)) - (1-y) \times \log(1 - h_\theta(x))}$$

→ No local minima.

→ Convex fn.

→ Single global minima.

⇒ Minimize cost function. $J(\theta_0, \theta_1)$ by changing θ_0, θ_1

Convergence Algorithm.

Repeat until convergence

$\{$ for $j = 0$ and 1.

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

$\}$

Performance metrics.

1. Confusion matrix.
2. Accuracy.
3. Precision.
4. Recall.
5. F - Beta Score.

DATASET.

f_1	f_2	y. output	$\hat{y}$ → model prediction.
-	-	0	1
-	-	1	1
-	-	0.	0
-	-	1	1
-	-	1.	1
-	-	0.	1
-	-	1.	0.

1. Confusion Matrix.

Actual values (y).

Predicted values. $(\hat{y})$

	1	0
1	3 (TP)	2 (FP)
0	1 (FN)	1 (TN)

T = True.
F = False.
P = Positive.
N = Negative.

$$\boxed{\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}}$$

$$= \frac{4}{7} = 57\%$$

DATASET → BINARY CLASSIFICATION.

→ 1000 datapoints. → 900 → 1
→ 100 → 0.

Imbalanced dataset.

Dumb model → only gives 1.

Here., 90% accuracy.

This is why precision, recall and F-Beta score are required.

* $\boxed{\text{Precision} = \dfrac{TP}{TP + FP}}$

Out of all actual values how many are correctly predicted.

Problem -

Mail $\longrightarrow$ Spam or Ham.

↑↑ TP	↓↓↓↓↓ FP
FN ↓	TN.

mail is ham but not recieved.

mail is spam but we recieve it, we can check it.

Problem.

Disease

TP	FP ↓↓
↓↓ FN. ↓↓	TN

Disease goes untreated

$$* \boxed{\text{Recall} = \frac{TP}{TP+FN}}$$

Out of all the predicted values how many are correctly predicted.

e.g. Tomorrow the stock market is going to crash.

- → Consumer → FN ↓↓
- → Companies → FP ↓↓

* F - Beta Score.

$$\boxed{F\beta = \frac{(1+\beta^2)(\text{Precision} * \text{Recall})}{(\beta^2 * \text{Precision} + \text{Recall})}}$$

* if FN and FN are both important :

$\beta = 1$

$$F1 \text{ Score} = 2 \frac{P * R}{P + R}.$$

* if FP is more important than FN :

$\beta = 0.5$.

$$F0.5 \text{ Score} = \frac{(1 + 0.25) P * R}{0.25 P + R}.$$

* if FN >> FP.

$$F2 \text{ Score} = \frac{(1+4) P * R}{4P + R}.$$