Logistic Regression (Classification problem)

DATASET

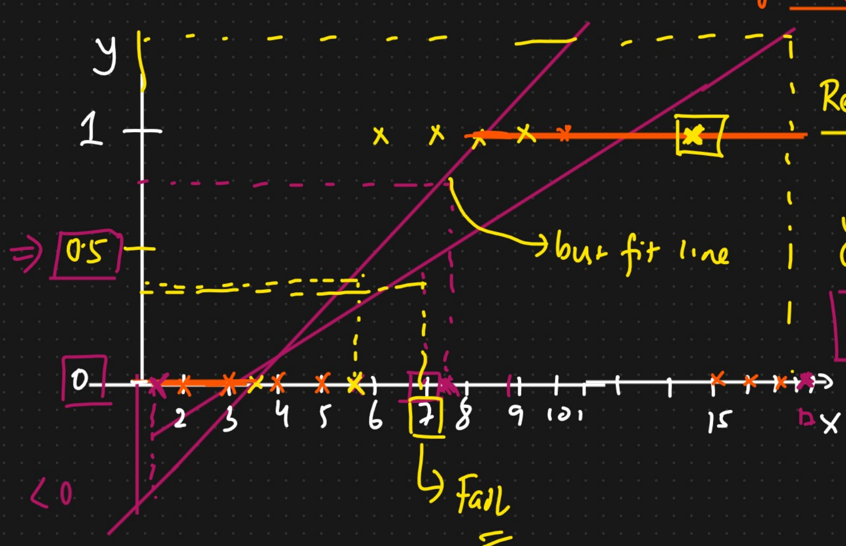
UPSC

Study hours	O/p (Pass/Fail) 1 0
2	FAIL
3	FAIL
4	FAIL
5	FAIL
6	PASS
7	PASS
8	PASS
9	FAIL

① Can we solve this problem using Regression?

>1 & <0

Logistic Regression



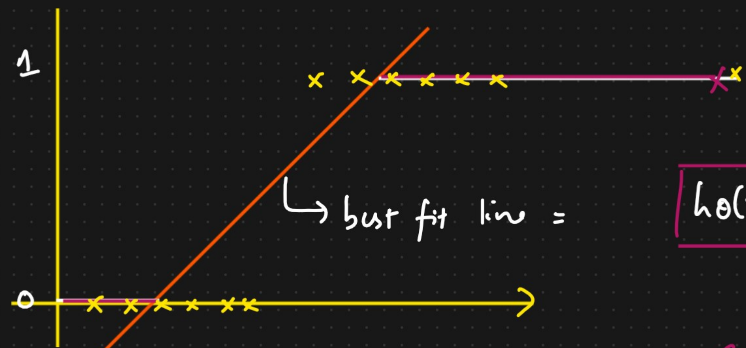
Regression

0.5 ⇒ Threshold

$y \leq 0.5 = 0$

$y > 0.5 = 1$

Sigmoid Activation



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

⇓

Sigmoid Activation ⇒ O/p = 0 to 1

① $z = h_\theta(x) = \theta_0 + \theta_1 x$

② Sigmoid fn $= \boxed{\dfrac{1}{1+e^{-z}}} \Rightarrow$ 0 to 1.

$z = \theta_0 + \theta_1 x$

$\dfrac{1}{1+e^{-z}}$

① Create a Best fit line

② Squashing → Sigmoid function

Linear Regression Cost function

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

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

MSE

$\Downarrow$

1 global Minima. ← $\boxed{\text{Convex function}}$

Logistic Regression Cost function

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

Sig $\quad h_\theta(x) = \text{Sig}(\theta_0 + \theta_1 x)$ ↗ Best fit Line

Let $z = \theta_0 + \theta_1 x$

$h_\theta(x) = \text{Sig}(z)$

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

0.35 ⇒ 0

0.59 ⇒ 1 $\quad$ 0.25 ⇒ 0

0.95 ⇒ 1

0.5 ⇒ 0

$z = \theta_0 + \theta_1 x$

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

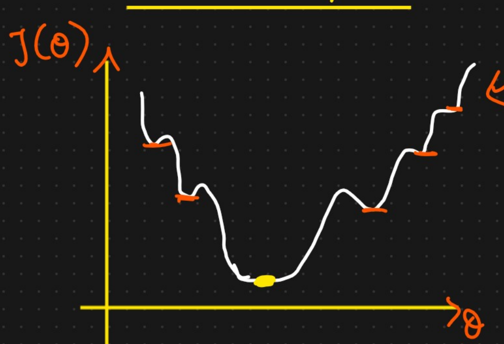
1 ⇐ 0.75 $\quad$ 0 to 1

Threshold

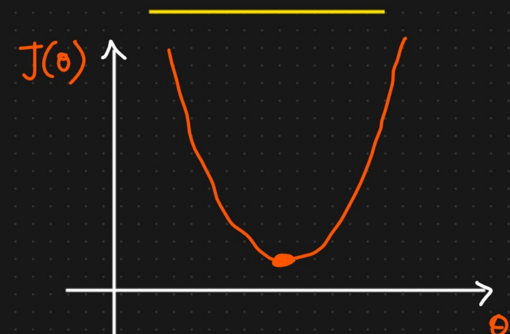
$\leq 0.5 \Rightarrow 0 \Rightarrow$ Fail

$> 0.5 \Rightarrow 1 \Rightarrow$ Pass.

Non Convex function



Convex function

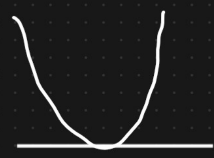


⊛ Log Loss Cost function

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

Truth Value

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



$\Downarrow$ Convex function

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

$\Downarrow$

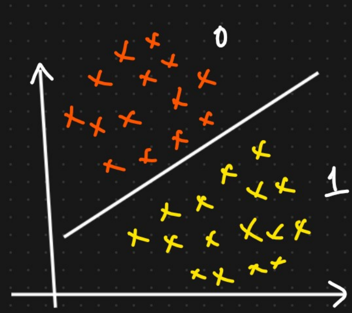
Near Local Minimal

Minimize Cost function $J(\theta_0, \theta_1)$ by changing θ_0, θ_1

Convergence Algorithm.

Repeat Convergence

$$\left\{ \quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad \right\} \qquad j = 0 \text{ and } 1$$



Performance Metrics

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

DATASET

f_1	f_2	y o/p	$\hat{y}$ ⟸ Model Prediction
–	–	0	1
–	–	1	1
–	–	0	0
–	–	1	1
–	–	1	1
–	–	0	1
–	–	1	0

y Actual Record

predicted	1	0
1	3	2
0	1	1

⇓

Confusion Matrix

Confusion Matrix

Actual

Predict	1	0
1	TP	FP
0	FN	TN

$$Acc = \frac{TP+TN}{TP+FP+FN+TN} = \frac{3+1}{3+2+1+1} = 4/7 \approx 57\%$$

(*) DATASET → BINARY CLASSIFICATION

↳ 1000 datapoints → 900 → 1, 100 → 0 } Imbalanced DATASET

Accuracy

Dumb model → 1 ⇒ 90% Accuracy ⇒ → X Sufficient

(*) Precision : $\boxed{\frac{TP}{TP+FP}}$

Actual

Predicted	1	0
1	TP	FP
0	FN	TN

Out of all the actual values how many are correctly predicted

Model → Diabetes or not Diabetes

Problem statement

Mail → Spam or Ham.

(*) Recall : $\boxed{\frac{TP}{TP+FN}}$ ⇒ Out of all the predicted values how many are correctly predicted.

Tommorow the stock market is going to crash.

	1	0
1	TP	FP↓↓
0	FN	TN

⇓

→ Consumer → FN ↓↓

→ Companies → FP ↓↓

① F-Beta Score :

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

① If FP and FN are both important

$\beta = 1$

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

② If FP is more important than FN

$\beta = 0.5$

$$F\text{-}0.5\ Score = (1+0.25)\frac{P*R}{(0.25*P+R)}.$$

③ If FN >> FP

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