logistic Regression

* for classification problem

Example:

* Problem to predict Pass or Fail based on study hors & play hors

IIT JEE

study/hor	play hors	o/p (Pass/fail)
1	8	Fail
2	7	Fail
3	7	fail
6	3	Pass
2	9	Pass → outliers

* we are predicted Pass or Fail
* we can make model to predict

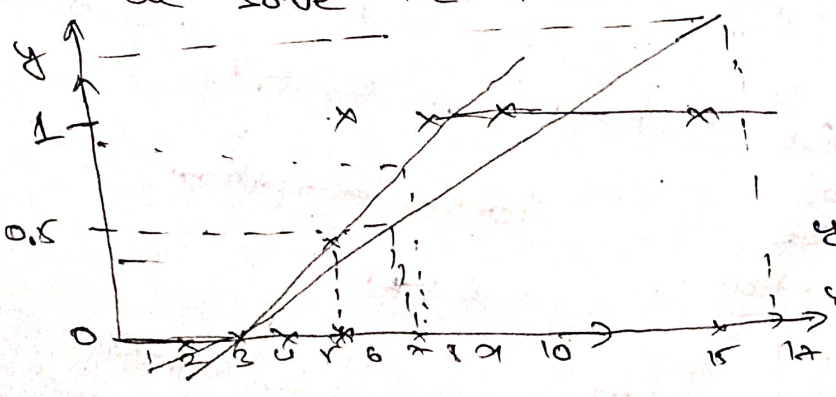
logiregression ⇒ Buy or Not

Dataset (UPSC)

study hours	o/p (Pass/Fail)
2	Fail
3	Fail
4	Fail
6	Pass
7	Pass
8	Pass

1 ⇒ Pass 0 ⇒ Fail

① Can we solve the problem using regression?



regression

* best fit line create

$y \leq 0.5 = 0$

$y > 0.5 = 1$

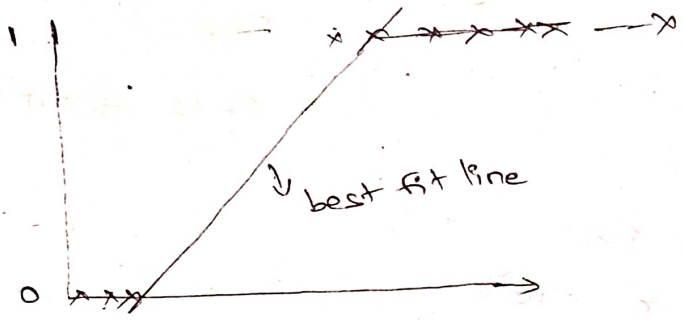
* Threshold is 0.5 ⇒ 0 and 1

* In regression
if outliers are there then the complete best fit line will be change if we see the 7 hours study student will fail it will give corona

* we can't remove outliers (Important)

* we squash the line in between 0 and 1

* find best fit then squash by sigmoid Activation



using
sigmoid Activation
⇓
$h(\theta) = \theta_0 + \theta_1 x$
⇓
o/p ⇒ 0 to 1

Steps logistic

① $z = h\theta(x) = \theta_0 + \theta_1 x$ → best fit line $z = \theta_0 + \theta_1 x$

② sigmoid fn = $\frac{1}{1+e^{-z}}$ ⇒ get in between 0 and 1

① Create a Best fit line

② Squashing → sigmoid function

linear regression cost function

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

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

⇓ MSE ⇓

Convex function ⇐ Global minimum ⇒ IV

Logistic Regression Cost Function :-

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

$z = \theta_0 + \theta_1 x$

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

↓ sigmoid activation

$= \sigma(z)$

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

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

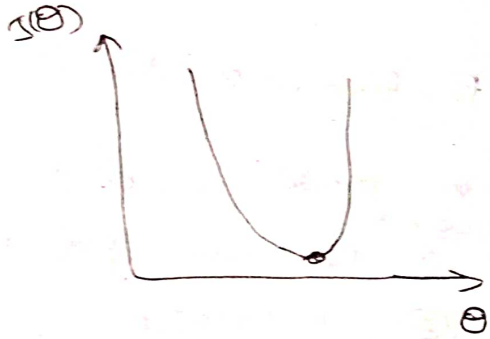
⇓

0 to 1

* Big issue with equation
* when we apply in logistic it give non convex function
* more
* Cost function bad convex or non convex function

Non Convex function

Convex funtion



* lotes of local minima, coeffcgt stuck
* global minima

* linear
* 1 global minima

* to fix Non Convex function we use different fun like log loss Cost function

Cost

* log loss Cost function

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

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

⇓⇓⇓

⇒ 

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

y = truth value

* through the cost function we will not get local minima
* we create convex fun
* inorder to value reduced by using

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

by Convergence Algorithm

Repeat Convergence

{

$J = 0$ and 1

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

}

Performance Metrics :-

* model is perform well or not

using

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

Dataset f_1	f_2	y	model prediction $\hat{y}$
—	—	0	1
—	—	1	1
—	—	0	0
—	—	1	1
—	—	1	1
—	—	0	1
—	—	1	0
—	—		

(1)

Confusion matrix :-

binary classification

$\hat{y}$ Predicted \ y actual	1	0
1	1	1
0	1	1

	1	0
1	TP	FP
0	FN	TN

1 ⇒ Positive

0 ⇒ Negative

(Accuracy)

② $Acc = \frac{TP+TN}{TP+FP+FN+TN}$

$\Rightarrow \frac{3+1}{3+2+1+1} = \frac{4}{7} \simeq 57\%$

* Dataset → binary classification

↳ 1000 data points → 900 → 1 } imbalance data set
→ 100 → 0

Dumb model → 1 ⇒ 90% Accuracy → × sufficient not

*Solve this problems

③ * Precision :- $\frac{TP}{TP+FP}$

	Actual 1	Actual 0
Predict	TP	FP
	FN	TN

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

* We aim reduce False Positive

Problem statement — FP, and FN is error

mail → Spam or Ham

* Focus on FP and FN reduce

Proble statement

Diabetes or not Diabetes.

* Focus on FN ↓↓

④ Use

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

Tomorrow stock market going to crash

→ Consumer → FN ↓↓

→ Companies → FP ↓↓

	1	0
1	TP	FP
0	FN	TN

use

* F-Beta Score :

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

(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

$\beta = 0.5$

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

(3) If FN >> FP

$$F2\ score = \frac{(1+4)\,P*R}{(4*P+R)}$$

* If imbalanced data set is there then we focus

FP ⇒ precision ↓↓

FN ⇒ recall ↓↓