16/10/22.

Logistic Regression : it is used to solve classification problem.

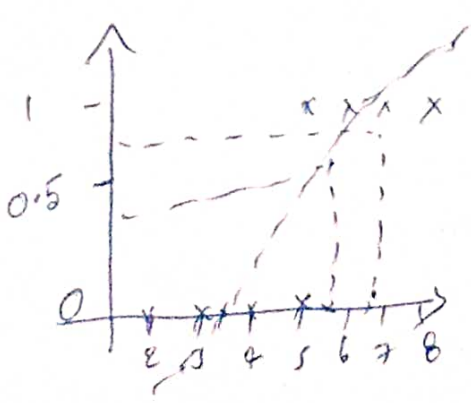
JIT JEE

Study hrs	Play hrs	O/p pass/fail
1	8	Fail
2	7	Fail
3	7	fail
6	3	Pass

Dataset : UPSC

Study hrs	O/p (pass/fail)
2	F
3	F
4	P
5	F
6	P
7	P
8	P
9	F

1. Can we solve this problem statement using Regression?



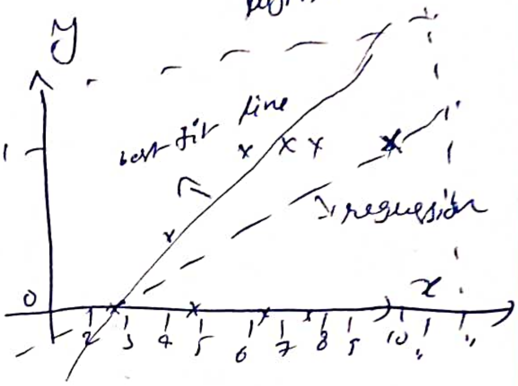
Regression → Best fit line.

$y \leq 0.5 = 0$

$y > 0.5 = 1$

Here 0.5 is Threshold.

Why linear regression will not work for classification problem
↓
because of outliers.



only possible bcz of logistic regression.

Squash ⇒ cut

→ if one outlier comes best-fit line changes

Sigmoid Activation :-

best-fit line

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

↓

Sigmoid Activation ⇒ O/p : 0 to 1.

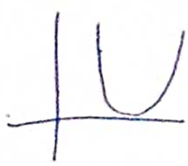
1. $z = h_\theta(x) = \theta_0 + \theta_1 x$ → best-fit line
2. Sigmoid fn $= \frac{1}{1+e^{-z}}$, here $z = \theta_0 + \theta_1 x$

∴ create a best fit line
∴ squashing → sigmoid fn.

Linear regression cost fn :-

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

↓ $h_\theta(x) = \theta_0 + \theta_1 x$ ⇒ MSE ⇓ convex fn.



here Threshold $\leq 0.5 = 0$ → Fail
$> 0.5 = 1$ → Pass

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

Logistic regression cost fn.

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

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

↓ assume $z = \theta_0 + \theta_1 x$

sigmoid Activation fn.

$= \sigma(z)$

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

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

if $0.35 \Rightarrow 0$

$0.25 \Rightarrow 0$

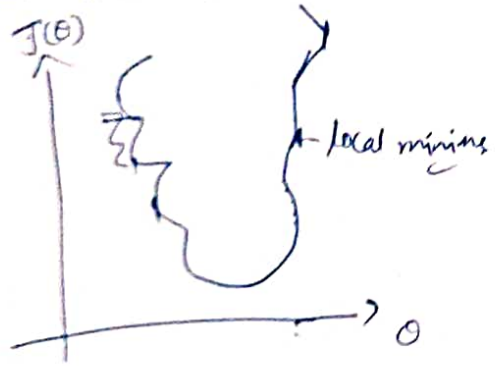
$0.95 \Rightarrow 1$

$0.5 \Rightarrow 0$ because it'll creates a non-convex fn.

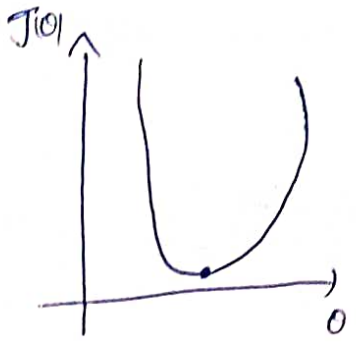
$0.59 \Rightarrow 1$

This creates a non-convex function.

Non Convex fn



Convex fn



How do we fix this:- To simply change the cost fn.

" log loss

Actually creates convex function

Log loss cost fn :-

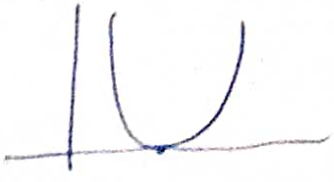
$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log h_\theta(x) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

↑ actual value

creates a convex fn.

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

through this we never get a local minima.



minimise the cost fun $J(\theta_0, \theta_1)$ by changing θ_0, θ_1

Convergence Algorithm.

Repeate until convergence

$J = 0, \frac{1}{2}$

threshold → ROC AUC curve

{

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

}

Performance Metrics :-

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

Dataset

f_1	f_2	y (Op)	$\hat{y}$ (model predictions)
-	-	0	1
-	-	1	1
-	-	0	0
-	-	1	1
-	-	1	1
-	-	0	1
-	-	1	0

Actual records (y)

Predicted Value ($\hat{y}$)	1	0
1	1+1+1 = 3	1+1 = 2
0	1	1

Confusion matrix

- Confusion matrix : it's a 2×2 matrix.

	1	0
1	1+1+1	1+1
0	1	1

y Actual value

$\hat{y}$: predicted value

Predicted (ŷ) \ Actual (y)	1	0
1	TP	FP
0	FN	TN

TP - True Positive
TN - True Negative
FP - False Positive
FN - False Negative.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} = \frac{3+1}{3+2+1+1} = \frac{4}{7} = \simeq 57\%$$

* Dataset ⇒ Binary Classification
  ↳ 1000 data points → 900 Points → 1 } Imbalanced dataset.
  → 100 Po → 0

Dumb model → 1 ⇒ 90% accuracy

- Precision :- $\frac{TP}{TP+FP}$

Predicted (ŷ) \ Actual	1	0
1	TP	FP
0	FN	TN

Out of all the values how many are correctly predicted.

Try to focus on FP
wanted to reduce the FP

Problem statement

Mail → Spam (or) Ham

Predicts a model → Diabetes
or not diabetes.

	1	0
1	TP	FP
0	FN	TN

here we need to focus on FN

Recall : $\frac{TP}{TP+FN}$

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

Tomorrow the stock market is going to crash.

→ Consumes → FN ↑↓

→ Companies → FP ↑↓

Companies

	1	0
1	TP	FP ↑↓
0		

↑ to Consumes

	1	0
1	TP	FP
0	FN	TN

- F - Beta Score :-

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

Conditions

1. If FP, FN are both important

$\beta = 1$

$$F1 \text{ score} = \frac{2 (P \times R)}{P + R}$$

2. If FP is more important than FN

$\beta = 0.5$

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

3. If FN >> FP

$$= \frac{(1+4)\ P \times R}{(4 * P + R)}$$

F2 score