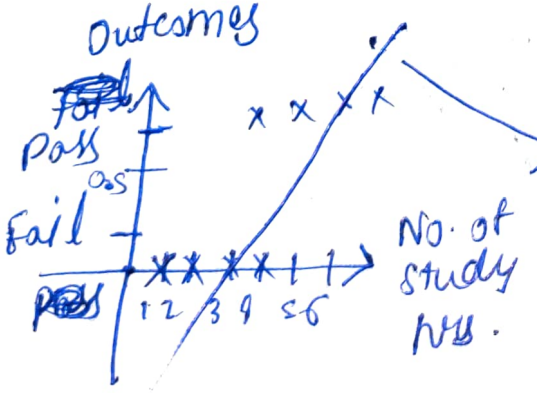
→ It is use to solve

Logistic Regression (First algorithm for classification)
(Binary classification.)

No. of study	No. of play	P/F
—	—	P
—	—	F
—	—	



We can't solve this problem with linear regression

Linear regression

If $h_\theta(x) < 0.5 \Rightarrow 0 \rightarrow$ Fail
$h_\theta(x) \geq 0.5 \Rightarrow 1 \rightarrow$ Pass.

Why we are not using linear regression?

- When we got outliers it change the entire line & give wrong result.
- Greater than 1 & less than 0 then we will get negative no. which will also be problem.

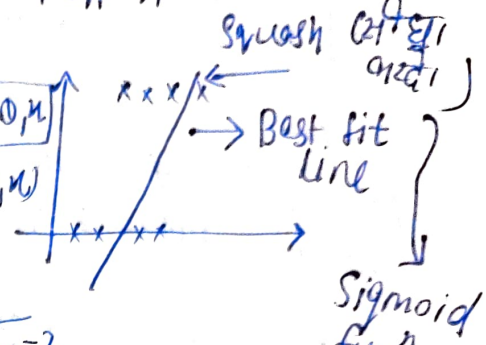
But we are result = 0 & 1.

So sigmoid function is the solution for it.

Decision Boundary logistic Regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

$$h_\theta(x) = \theta^T x$$



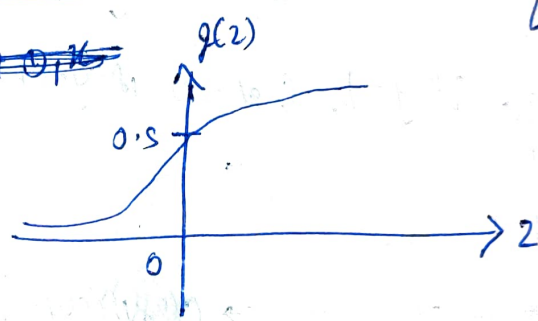
Let $z = \theta_0 + \theta_1 x$

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

$$\boxed{h_\theta(x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}} \rightarrow \text{Sigmoid or Logistic fun}^n.$$

~~Let $z = \theta_0 + \theta_1 x$~~

$g(z) \geq 0.5$ when $z \geq 0$



Training set

$$\{(x^1, y^1), (x^2, y^2), (x^3, y^3), \ldots, (x^n, y^n)\}$$

$y \in \{0, 1\} \rightarrow$ 2 o/p.

Change parameter θ_1 ?

Cost function

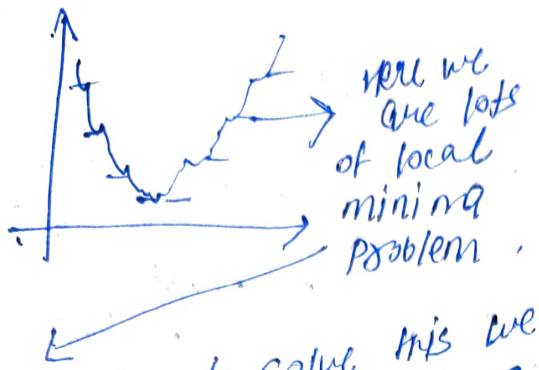
Linear regression $\rightarrow$ $J(\theta_1) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} (h_\theta(x^i) - y^i)^2$

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

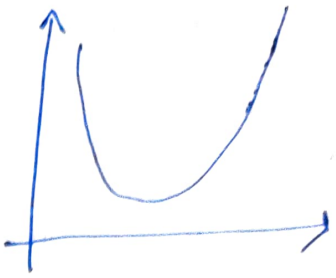
Logistic regression cost function $= \frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2$

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

Non-convex function



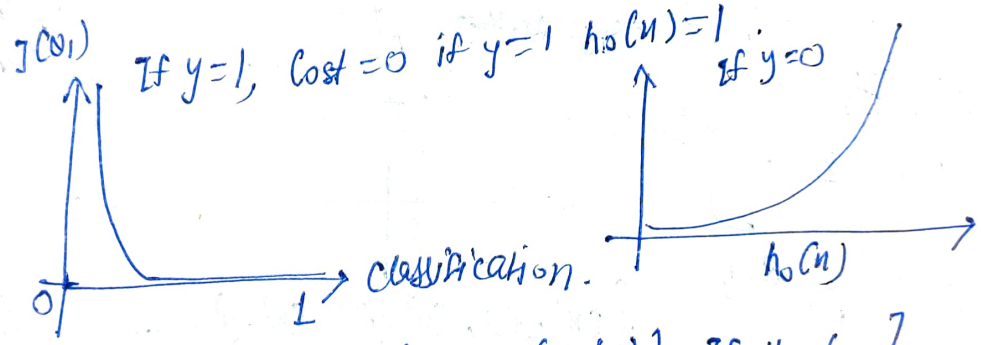
Convex function



In order to solve this we have something called logistic regression cost function -

So, logistic regression actually changes,

$$J(\theta_1) = \begin{cases} -\log(h_\theta(x)) & y=1 \\ -\log(1-h_\theta(x)) & y=0 \end{cases}$$



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

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

↓ Cost fun^n

If $y=1$

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

If $y=0$

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

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

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

Repeat until convergence.

$\{ \theta_j = \theta_j - \alpha \frac{\partial}{\partial j} (J(\theta_1)) \}$

↳ We are changing cost fun^n for non convex function.

Performance metrics {classification problem}

x_1	x_2	y	$\hat{y}$
—	—	0	1
—	—	1	1
—	—	0	0
—	—	1	1
—	—	1	1
—	—	0	1
—	—	1	0

Actual - o/p

Pred

No. of times

Pred \ Actual	1	0
1	3	2
0	1	1

	1	0
1	TP	FP
0	FN	TN

→ Confusion matrix.

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

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

Now, $0 \to 900$, $1 \to 100$ } Imbalanced data

$0 \to 600$, $1 \to 400$ } Balanced Data.

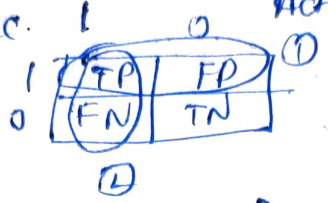
1. ~~Precision~~ ;

2. Recall

Precision | Recall | F-score

$\left\{\frac{TP}{TP+FP}\right\}$ → Reduce Prec.

$\left\{\frac{TP}{TP+FN}\right\}$ → Reduce FN



Spam classification — Precision.

Has cancer or not — Recall

Tom stock market is going to crash — For people / For company — Different for both.

But for both we use F-score.

$$(1+\beta^2)\frac{Prec \times Recall}{\beta^2 \times Prec + Recall}$$

$\beta = 1$ (F_1), $\frac{2(Prec \times Recall)}{Prec + Recall}$ → Harmonic mean.

$\beta = 0.5$ ($F_{0.5}$), $(1+0.5)^2 \frac{P \times R}{0.5^2 P + R}$

$\beta = 2$ (F_2), FN >> FP.

Agenda

1. Practicals.
2. Naive Bayes.
3. KNN algorithms.

. Cross Validation — For doing testing & training of all possible combinations.

→ L_1 & L_2 penalty job is also present in logistic regression.