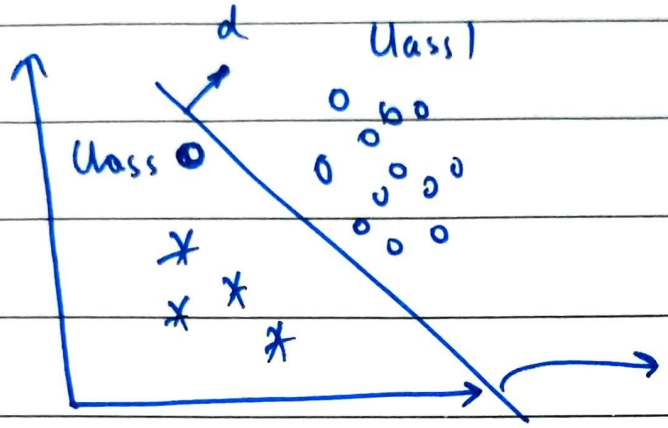
Logistic Regression

* It is a classification technique



This line is Hypothesis.

Or, hypothesis will look like a plane in higher dimensions.

Eqⁿ of line :

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

(This is of form $ax + by + c = 0$)

Now, we are given a point d.

We calculate it's distance from hypothesis.

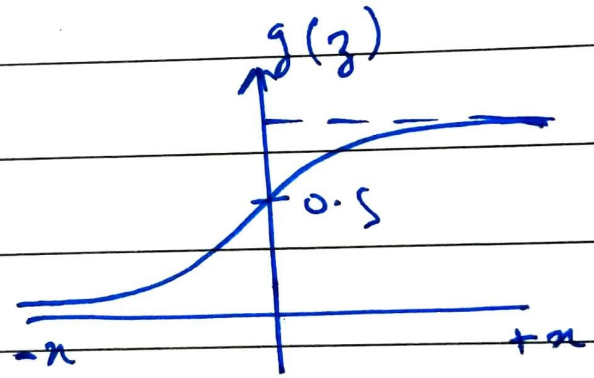
If, $d = \theta_1 x_1 + \theta_2 x_2 + \theta_0 \geq 0$ then point lies RHS of line & class is 1

If $d < 0$ the class is 0.

Higher distance means more confidence.
d can be from $-\infty$ to ∞, we want to map it in range $[0, 1]$. This will be our probability.

Sigmoid function :

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

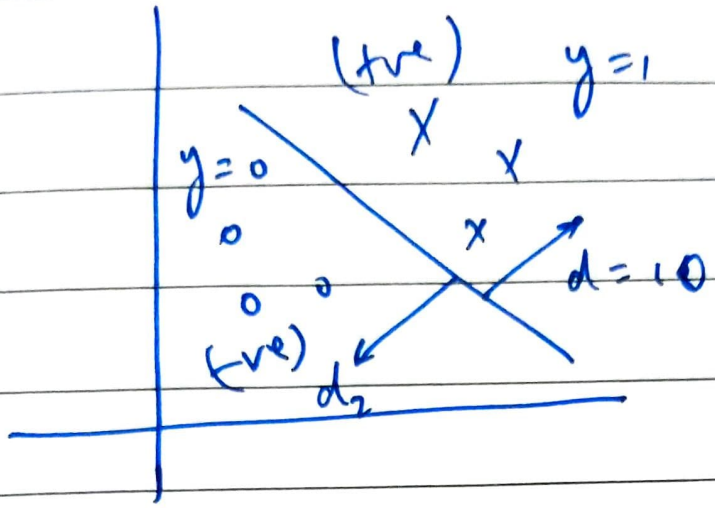


$g(z) \in (0, 1)$

Now, hypothesis is going to be of form :

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

where, $d = \theta^T x$ || distance



$$g(z) = \frac{1}{1+e^{-3}} = 0.8 \text{ (+ve)}$$

Prob $(Y=1 \mid x) = 0.8$ Belonging to class 1

Prob $(Y=0 \mid x) = 0.2$ Belonging to class 2

* As we move away from the line, the probability that point lies in that class increases.

d_2 will be negative.

$$g(\theta^T x) = \frac{1}{1+e^{(-\theta^T x)}} = (0.2)$$

($g(\theta^T x)$ ⇓ -ve; $-\theta^T x$ → This will be +ve now)

$P(Y=1 \mid x) = 0.2$ // Important

$P(Y=0 \mid x) = 0.8$ // Point belongs here.

By assuming, it will create a model like that.

Now, (here we assume that for $y=0$, distance is -ve)

$$P(Y=1 \mid x, \theta) = h_\theta(x)$$

$$P(Y=0 \mid x, \theta) = 1 - h_\theta(x) \rightarrow y_a$$

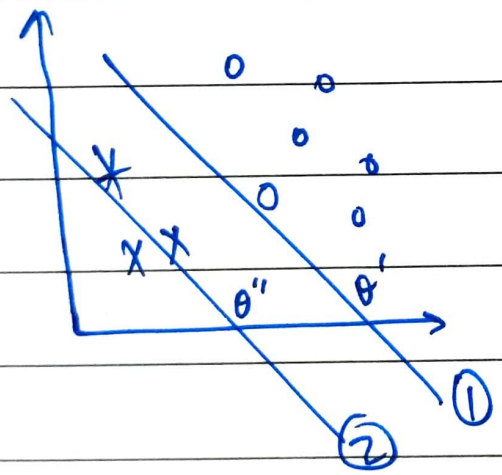
$$P(Y \mid x, \theta) = \left(h_\theta(x)\right)^{y} \left(1-h_\theta(x)\right)^{1-y}$$

$y \rightarrow y_n$

(Here y will be 0 OR 1 only!)

Which is the best line?

Concept : Max likelyhood estimation.



(Obj: line ① has more likelyhood)

$$P(Y \mid x, \theta') > P(Y \mid x, \theta'')$$

Line ① has higher likelyhood than ②

(Basically means all pts. in dataset)

* Each pt. is independent

$P(y_i/x_i)$ → For i^{th} item, the label y_i depends on the features of the pt.

Say you have 2 types of fruits apple & mango in a basket. If you pick apple you get juice, if mango then shake.

$P(\underset{y=1}{\text{Juice}} / \text{Fruit})$	$P(\underset{y=0}{\text{Shake}} / \text{Fruit})$

Here, what drink you will get depends on Fruit you pick. Independant Probabilities.

P(Both Juice & Mango) = Multiply Probabilities

$$\text{Likelyhood} = \prod P(y_i/x_i, \theta)$$ // Product of all Prob.

We maximize this product.

because we want product of all probabilities in training set should be maximised.

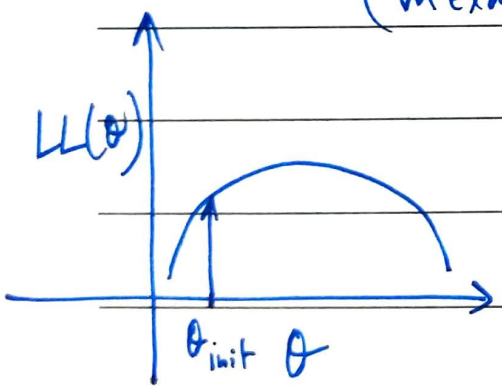
This is $P(Y/X, \theta)$ (on previous page)

* We need to find line such that product of all probabilities (-ve & +ve side both) is maximum.
* This is called maximisize likelyhood.
* After that, we take log likelyhood. LL after taking log on both sides of $P(y/x, \theta)$.

$$LL(\theta) = \sum_{i=1}^{m} y_i \log h_\theta(x_i) + (1-y_i) \log (1 - h_\theta(x_i))$$

(m examples)

AKA cross entropy



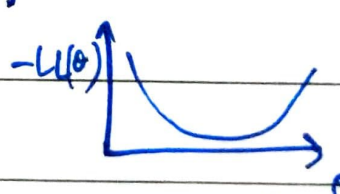
(⇒ Gradient Ascent)

$$\theta = \theta + \eta LL'(\theta)$$

$$\theta_j = \theta_j + \eta \sum_i (y_a - h_\theta(x_j^i)) x_j^i$$

(j^{th} feature) (i examples)

LR



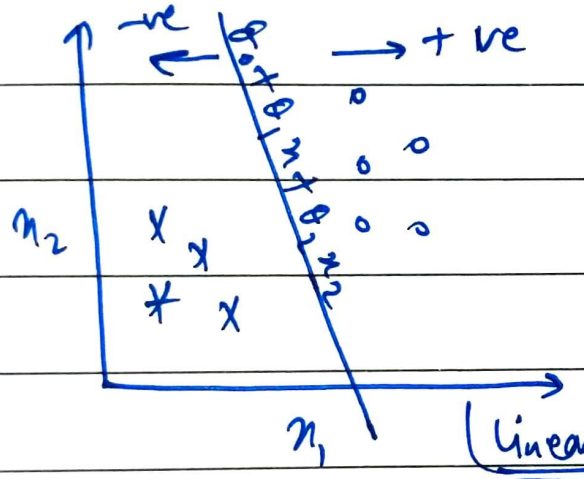
(OR, loss = $-LL(\theta)$ & minize this.)

Prediction :

$h_\theta(x) = g(\theta^T x)$ ∴ Prob * belongs to class 1

$g(\theta^T x) \geq 0.5 \Rightarrow$ Class 1

$< 0.5 \Rightarrow$ Class 0



$\theta^T x^{(i)} \geq 0 \Rightarrow (+)$

$\theta^T x^{(i)} < 0 \Rightarrow (-)$

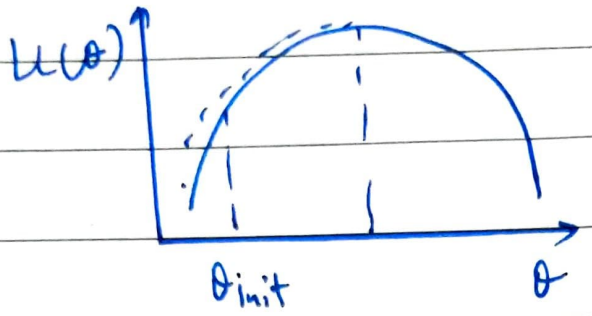
Find, optimal [best set of parameters] θ

$g(\theta^T x)$ — (0, 1), it represents prob.

Update Rule :

$\theta = \theta + \eta \, (\nabla_\theta \cdot LL(\theta))$ Gradient

η → learning Rate



At local maxima, gradient will become 0.

$$\nabla_\theta LL(\theta) = \frac{\partial (LL(\theta))}{\partial \theta}$$

One property, $g(z) = \frac{1}{1+e^{-z}}$

* $g'(z) = g(z)(1 - g(z))$

$$\frac{\partial (LL(\theta))}{\partial \theta_j} = (y - h_\theta(x))\, x_j \quad \text{// Identical to LR}$$

$$\theta_j = \theta_j + \eta \sum_{i=1}^{m} (y_i - h_\theta(x^{(i)}))\, x_j^{(i)}$$

(j^{th} feature of i^{th} example)

// This is our update rule to reach local maxima.

So, we are summing up gradients over all examples for all the features.

In the end, our $X\theta$ is made in such a way that for a point $\in y=0$, distance is -ve & hypothesis $h(x)$ gives $P(y=1/x)$.