MODULE-3
Chapter-27

LOGISTIC REGRESSION

→ Logistic Regression is simple and elegant.

→ Even Though The name Refers for Regression It could be mainly applied for classification.

→ There are many Interpretations for logistic regression

Logistic Regression
- → Geometric Interpretation/Intution
- → probability Interpretation/Intution
- → loss-function Interpretation/Intution.

✳ Geometric Intuition:-

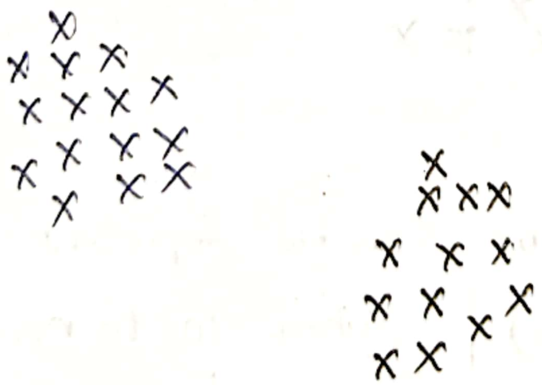
→ To understand this let us consider

X: -Ve class points

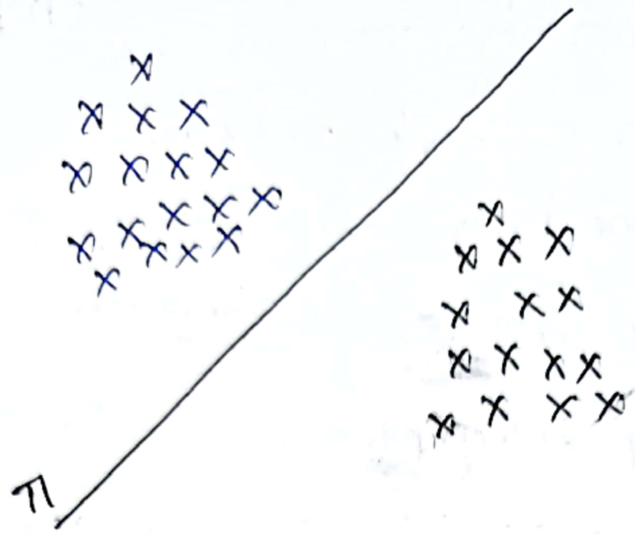
X: +Ve class points.

Logistic Regression is a Statistical method used to predict The outcome of a dependent variable based on previous outcome. It is used to solve binary classification problems

It is classification ~~problem~~ algorithm That predicts a binary outcome based on a series of independent variables

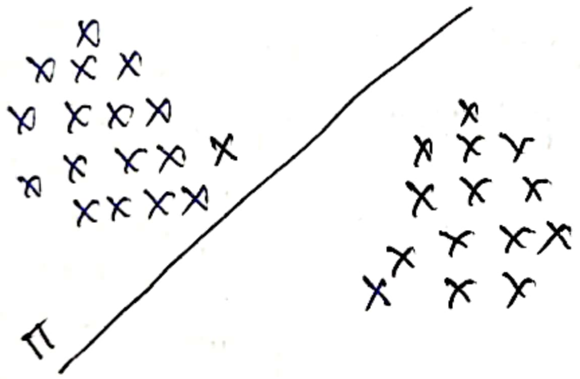


→ lets use a line in 2D (or) a hyperplane:nD To seperate +Ve points from The negative points.



→ If my data is linearly Seperable (which means a line or a plane that seperates the data)

→ There is also a term called almost linearly Seperable like the whole data got seperated except few points



→ The equation of the plane can be represent as

$\boxed{\pi : (\omega, b)}$ where ω to normal to the plane π
b: Intercept

In higher dimensions The equation of plane represented as

$$\boxed{\pi : \omega^T x + b = 0}$$

Scanned with OKEN Scanner

→ if the π passes Through Origin : $b=0$

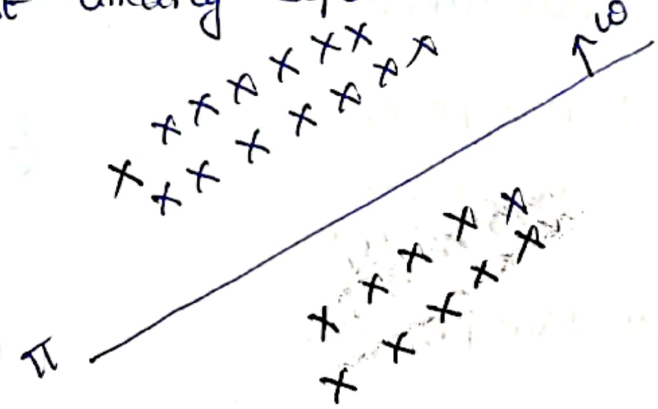
$$\omega^T x = 0$$

π : $\omega^T x + b = 0$

where $x \in \mathbb{R}^d$; $\omega \in \mathbb{R}^d$ ⇒ Vectors

$b \in \mathbb{R}^d$ ⇒ Scalar

→ So the assumption to perform logistic Regression is that Class are Linearly seperable (or) almost linearly Seperable.



→ So from the eqn of the plane

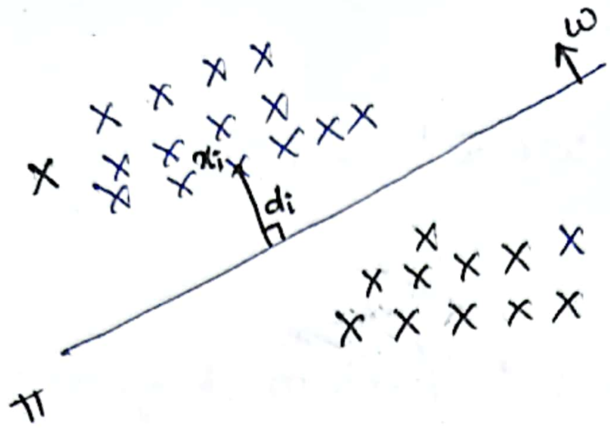
π : $\omega^T x + b$

given : $D_n = \{+ve, -ve\}$

and the Task is to find ω and b

so by finding ω & b we can find the π That best Seperates +ve points from -ve points.

Step 1 :- finding distance of a point from the plane



where:
x_i: the point
d_i: distance of point from the plane.

and let y_i be the class labels that represents

$y_i = +1$: +ve points
$\quad -1$: -ve points

$y_i \in \{-1, +1\}$

Note:- X Not +1 as +ve pts and 0 as -ve points X

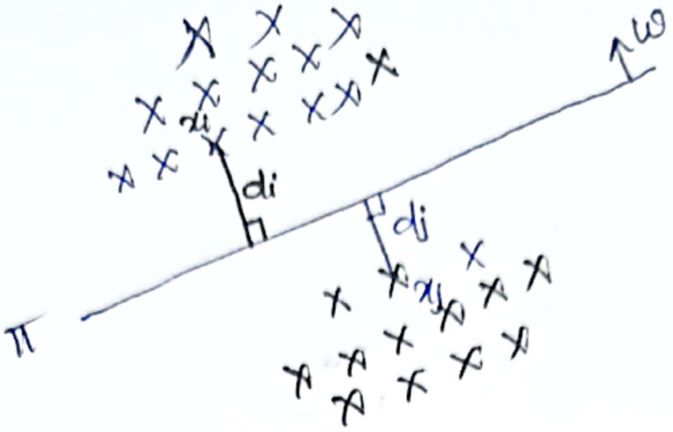
* So from the basics of linear algebra distance d_i

$$d_i = \frac{w^T x_i}{\|w\|}$$

; where w is the normal to the plane
and let $\|w\|$ is a unit vector
So $\|w\| = 1$

Step 2:- finding other side distance of a point from the plane

→ Similarly we will find the distance d_j of other side of the plane



→ as x_i is towards the direction of normal w we can say

$$d_i = w^T x_i > 0$$

~~$d_j = w^T x_j < 0$~~

and lly as x_j is towards opp side of normal we can say

$$d_j = w^T x_j < 0$$

→ And in Logistic Regression the Decision surface will be that line/plane.

→ so our classifier classifies

if $w^T x_i > 0$ then $y_i = +1$

& lly

if $w^T x_i < 0$ then $y_i = -1$

#Note:-

usually we need to consider plane eqn as $w^T x_i + b$ but here we are assuming the plane passed through the origin.

Step3:- Classification of points

Case 1:-

let us Consider the +Ve point

if $y_i = +1$

and $w^T x_i > 0$ $\Rightarrow$ That means if Classifier is saying its +Ve point

So $y_i * w^T x_i > 0$

that means w is correctly classifying the point.

Case2:-

let assume $y_i = -1$: -Ve point

and assume $w^T x_i < 0$ $\Rightarrow$ That means LR is Concluding that x_i is -Ve point.

→ Now if we take

$$\underset{-Ve}{y_i} * \underset{-Ve}{w^T x_i} > 0$$

So we conclude for both +Ve & -Ve points if

$y_i * w^T x_i > 0$ $\Rightarrow$ That means the LR model is Correctly classifying the poin

Case 3
if $y_i = +1$ (+ve points)

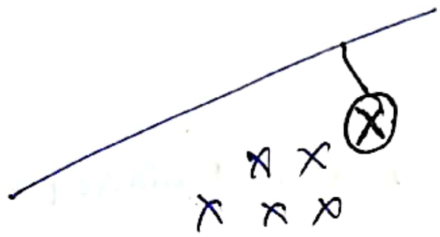
& if $w^T x_i < 0$ ⇒ That means LR is saying x_i is -ve class.

so $y_i * w^T x_i < 0$

That implies

$y_i = +1$

but we got LR: -1 that means misclassified

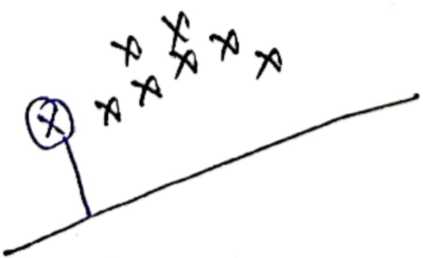


Case 4

if $y_i = -1$

if $w^T x_i > 0$ ⇒ That means LR is saying x_i is +ve class

$y_i * w^T x_i < 0$ ⇒ misclassified



* So @ The end for the classifier to be Very good

① either it has to consider min no of misclassifications

② or max no of correctly classified points.

So we ~~have~~ need as many points as possible to have

$y_i w^T x_i > 0$

Step 3:- Mathematical Optimization.

So as we seen from previous step we need to ~~get~~ have

as many possible to have $y_i w^T x_i > 0$

so that means we need to maximize it.

$$\max_{w} \sum_{i=1}^{n} y_i w^T x_i$$

This means
both x_i & y_i are fixed in our Dataset the only variable is w

→ This represent the term or variable that we need to change / vary to maximize it.

→ In python argmax is used to maximize fn.

$$w^* = \underset{w}{\operatorname{argmax}} \left(\sum_{i=1}^{n} y_i w^T x_i \right)$$

This means optimal w

Variable

And This w^* we need to find in this math optimize

27-2

SIGMOID FUNCTION : SQUASHING :-

as we seen math optimization

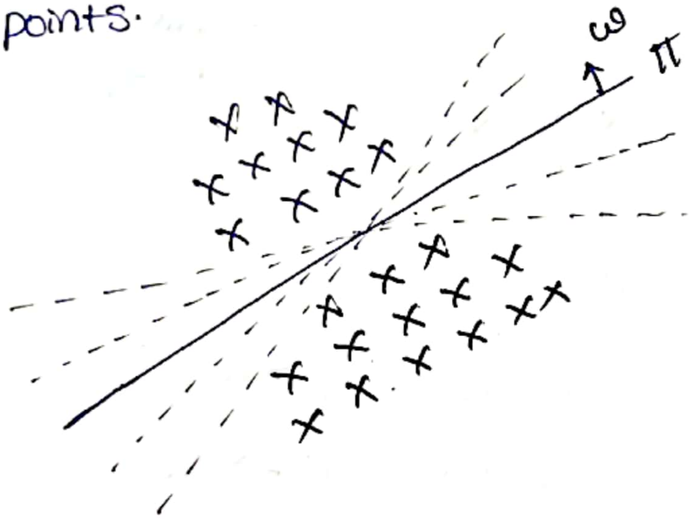
$$\boxed{w^* = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} y_i w^T x_i}$$

↳ optimal w

This means as to seperate +ve points from -ve points

There can be n no of planes could be possible

we need to find best w_i corresponding to the plane π_i that best seperates both +ve and -ve points.



π_i ; w_i

$$* \quad w^* = \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \underbrace{y_i w^T x_i}$$

This portion will be called as signed distance

Signed distance because $w^T x_i$ dist from x_i to π

and

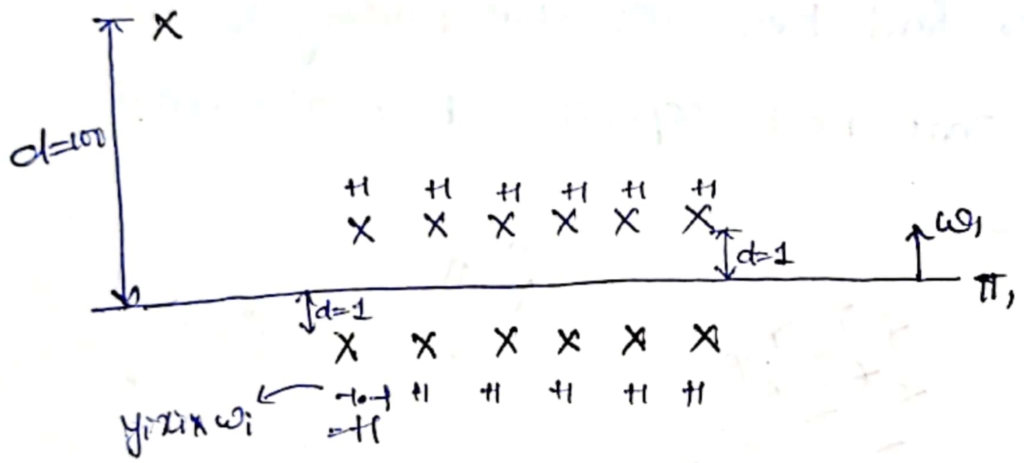
$y_i w^T x_i$: +ve $\Rightarrow$ π as defined by w correctly classifies x_i.

: −ve $\Rightarrow$ incorrectly classifies x_i.

* ~~There are cases where~~

* Ex:- To show where math optimization does not work well

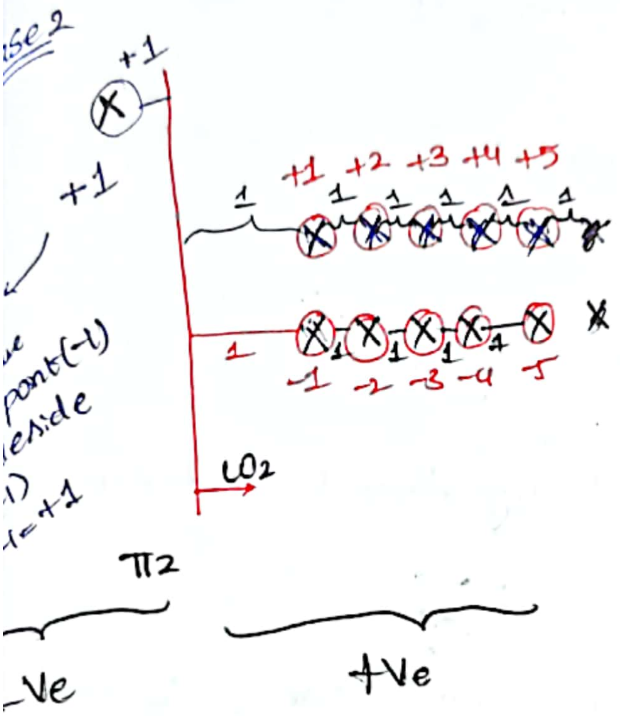
let us assume we have +ve and −ve points



Case 1:- π_1 is my Seperation

$$\sum_{i=1}^{n} y_i w_i^T x_i = 1+1+1+1+1 \rightarrow +ve$$
$$+1+1+1+1+1 \rightarrow -ve$$
$$-100$$

$$= -90$$



5✓

5+✓

5-✗ $\quad \frac{6}{10}$

1-✓

$\cancel{+1}\ \cancel{+2}\ \cancel{+3}\ \cancel{+4}\ \cancel{+5} \rightarrow +ve$

$\cancel{-1}\ \cancel{-2}\ \cancel{-3}\ \cancel{-4}\ \cancel{-5} \rightarrow -ve$

$+1 \rightarrow$ outlier

$\Rightarrow +1$

l as objective is to find ω that maximizes sum of
ned distances.

$+1 > -90$

Can say π_2 as our classifier for this example.

ut as we see Intuitively for π_1 out of 11 points
points are correctly classified whereas if we see π_2
out of 10 points are correctly classified. So if we
ink logically π_1 would be the best compared to π_2
t as we are considering sum of signed distances
ecause of one outlier the model is choosing π_2
nstead of π_1

we can say sum of signed distances is easily
prone to outliers.

→ So one Single / extreme outlier point

→ So to Overcome This we use a technique Called Squashing.

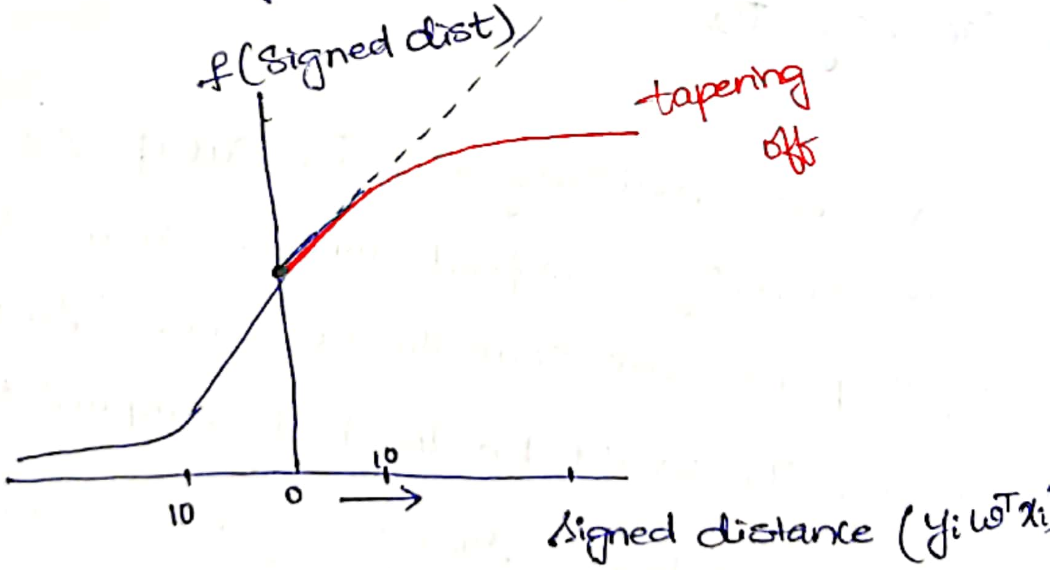
<u>Squashing</u>

* The idea is Instead of using signed distance, we need consider

if signed distance is small use as it is

if signed distance is large: make it a sm value.

* To do squashing



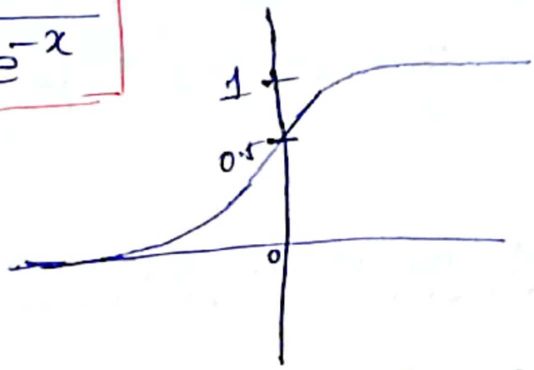
So we will design or Create a function that ever the signed distance gets huge it will be tappered a particular Value.

→ $\underset{w}{\arg\max} \sum_{i=1}^{n} y_i w^T x_i \Rightarrow \underset{w}{\arg\max} \sum_{i=1}^{n} f(\cdot\, y_i w^T x_i)$

and that function is called sigmoid function. $\sigma(x)$

$$\boxed{\sigma(x) = \frac{1}{1+e^{-x}}}$$

The Graph of $\frac{1}{1+e^{-x}}$



→ So by this when our data point is very far away from the plane then $w^T x_i$ is very large. by this fn we get $P(y_i=1)=1$

→ Summary

* As we are optimizing max. sum of signed distance but it got problem with Outlier

↓

So we find function called sigma $\sigma(x)$ which has properties

1. tapering behaviour, linear, a probabilistic interpretation.
2.

↓

At last we get max sum of transformed signed distance.

$$\rightarrow \boxed{w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \sigma(y_i w^T x_i)}$$ $\because \sigma(x) = \frac{1}{1+e^{-x}}$

$$\boxed{w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \frac{1}{1+\exp(-y_i w^T x_i)}}$$

and this is less impacted by outliers.

27.3 Mathematical formulation of Objective fn :-

from Optimization problem

$$\left[w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \frac{1}{1+\exp(-y_i w^T x_i)} \right]$$

we can even ~~Optimize~~ Simply this further

from Monotonic fn's :-

A fn $g(x)$ is said to be monotonic fn

if $x_1 > x_2$ then $g(x_1) > g(x_2)$ then $g(x)$

is said to be monotonically Increasing fn.

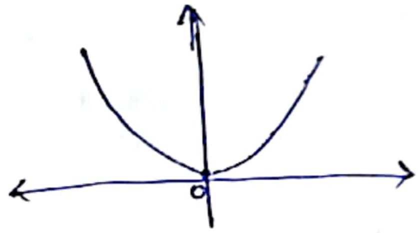
Note

Ex of Optimization problem:

let us consider an example

$$x^* = \arg\min_x x^2$$

by using maxima and minima concepts.
we can say



so ⇒ if $f(x) = x^2$

$$x^* = \arg\min_x x^2 \Rightarrow \arg\min_x f(x) = 0$$

Note

If $g(x)$ is a monotonic~~ally~~ fn.

then
$$\arg\min_x f(x) = \arg\min_x g(f(x))$$

$$\arg\max_x f(x) = \arg\max_x g(f(x))$$

Ex:- let $g(x) = \log(x)$
and let $f(x) = x^2$

$$x^* = \arg\min_x f(x) = \arg\min_x x^2 = 0$$

$$x' = \arg\min \log(x^2) = 0$$

$$w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \frac{1}{1+\exp(-y_i w^T x_i)}$$

(or)

$$w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \sigma(y_i w^T x_i)$$

$$w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \log(\sigma(y_i w^T x_i))$$

$$\boxed{w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} \log \frac{1}{1+\exp(-y_i w^T x_i)}}$$

as $\log(1/x)$

$$\boxed{w^* = \underset{w}{\arg\max} \sum_{i=1}^{n} -\log(1+\exp(-y_i w^T x_i))}$$

* "The point that minimizes a function is the same point that maximizes $-f(x)$"

$\Rightarrow \underset{x}{\arg\max}\ f(x) = \underset{x}{\arg\min}\ -f(x)$ //

$\|$ly $\underset{x}{\arg\max}\ -f(x) = \underset{x}{\arg\min}\ f(x)$

$$\therefore \boxed{w^* = \underset{w}{\arg\min} \sum_{i=1}^{n} \log(1+\exp(-y_i w^T x_i))}$$

and this the Optimization problem of Logistic Regression.

where $y_i : +1$ (or) -1

Note:-

* As we see @at last we could get a Optimization problem by sum of signed distance only just by making up with log and exp

$$\underset{w}{\text{argmin}} \sum_{i=1}^{n} \cancel{\log}\,(\cancel{1+}\,\cancel{\exp}\,(-y_i w^T x_i))$$

$$\underset{w}{\text{argmin}} \sum_{i=1}^{n} (-y_i w^T x_i)$$

$$\downarrow$$

$$\underset{w}{\text{argmax}} \sum_{i=1}^{n} (y_i w^T x_i)$$

→ which is sum of signed distance only

* The above Optimization is of geometric Interpretation. Similarly we can also derive this using probabilistic method.

$$\boxed{w^* = \underset{w}{\text{argmin}} \sum_{i=1}^{n} -\overset{+1 \text{ or } 0}{y_i} \log P_i - (1-y_i) \log (1-P_i) \quad \text{where } P_i = \sigma(w^T x_i)}$$

2.7.4 WEIGHT VECTOR :-

→ As we see from optimization problem

$$w^* = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \log\left(1+\exp\left(-y_i w^T x_i\right)\right)$$

→ and This w^* is called weight Vector which is the best w^T value.

→ and w^* is a d-dimensional Vector

$$w^* = \langle w_1, w_2, w_3, \dots w_d \rangle$$

$$w^* \in \mathbb{R}^d$$

→ If we have a weighted Vector with d dimensions and if There are d features, for every feature There is a corresponding weight associated with it.

$$w = \langle w_1, w_2, w_3 \dots w_d \rangle$$
$$f_1 \quad f_2 \quad f_3 \quad \dots f_d$$

→ If we perform decision

like given a Query point x_q → (we need to find Class label) y_q

if $w^T x_q > 0$ then ~~x_q~~ $y_q = +1$

or if $w^T x_q < 0$ then $y_q = -1$

whereas probabilistic Interpretation. need sigmoid fn

$$\sigma(w^T x_q) = P(y_q = +1)$$

To decide the given Query point $x_q \rightarrow y_q$.

Interpretation of w :-

Case 1

If $w_i = +ve$ for a given feature f_i, and if i^{th} Component of given Query point $x_{qi} \uparrow$ then

$$x_{qi} \uparrow \Rightarrow (w_i x_{qi}) \uparrow$$

$$\Rightarrow \sum_{i=1}^{d} (w_i x_{qi}) \uparrow$$

$$\Rightarrow \sigma(w^T x_q) \uparrow$$

$$P(y_q = +1) \uparrow$$

Case 2 :-

if $w_i = -ve$

then as

$$x_{qi} \uparrow \Rightarrow (w_i x_{qi}) \downarrow$$

$$\Rightarrow \left(\sum_{i=1}^{d} w_i x_{qi} \right) \downarrow$$

$$\Rightarrow \sigma(w^T x_q) \downarrow$$

$$\Rightarrow P(y_q = -1) \uparrow$$

whereas probabilistic Interpretation. need sigmoid fn

$$\sigma(w^T x_q) = P(y_q = +1)$$

To decide the given query point $x_q \rightarrow y_q$.

Interpretation of w:-

Case 1

If $w_i = +ve$ for a given feature f_i, and if i^{th} Component of given Query point $x_{qi} \uparrow$ then

$$x_{qi} \uparrow \Rightarrow (w_i x_{qi}) \uparrow$$

$$\Rightarrow \sum_{i=1}^{d} (w_i x_{qi}) \uparrow$$

$$\Rightarrow \sigma(w^T x_q) \uparrow$$

$$P(y_q = +1) \uparrow$$

Case 2:-

if $w_i = -ve$

then as

$$x_{qi} \uparrow \Rightarrow (w_i x_{qi}) \downarrow$$

$$\Rightarrow \left(\sum_{i=1}^{d} w_i x_{qi}\right) \downarrow$$

$$\Rightarrow \sigma(w^T x_q) \downarrow$$

$$\Rightarrow P(y_q = +1) \downarrow \qquad \Rightarrow P(y_q = -1) \uparrow$$