→ 'Logistic-Regression'

Even though the name here is regression, It is actually a classification technique.

↳ Geometrically, it is very very easy & elegant algorithm.

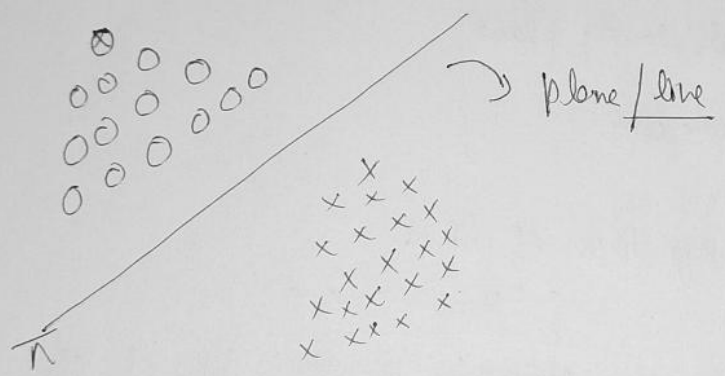
There are multiple interpretations to Logistic Regression (LR), we can interpret it using Geometry, we can interpret it using Probability & even we can interpret it using Loss-functions.

Geometry is much more easy to visualize, so we will be focusing on geometric intuition of Logistic-Regression.

Let's begin with geometric intuition of Logistic Regression.

Imagine we have two classes of points.

X :- -ive labelled points
O :- +ive labelled point

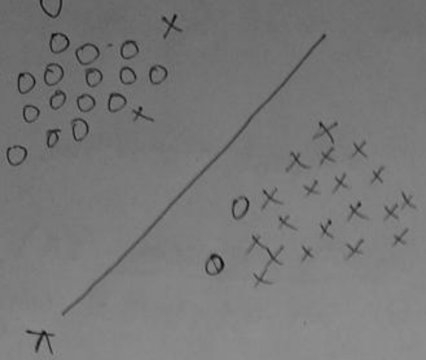


Now to separate both these classes of points, the first thing that we notice, quickly is, we can draw a line / hyperplane (in 2d we can draw a line & in n dimensional space, we can draw a plane)

if I somehow could separate the positive points from negative points using a plane (π). which we need to find.

Note :→ If my data is 'linearly separable', 'linearly separable' means that there is a line or a plane that can separate the '+ive class' from the '-ive class'.

& there is another term called 'almost linearly separable' & it means most of the data is almost linearly separable. (we might have a + ive point along with the -ive class points & vice versa



→ The data is almost "linearly separable data"

Note :→
1) Lines & planes are called linear surfaces.
2) Circles, parabolas & Ellipses are called quadratic surfaces.

we know a plane can be represented by

$\pi : (w, b)$

[where "w" is normal to the plane
&
"b" is the 'y' intercept]

It can be derived using equation of line.

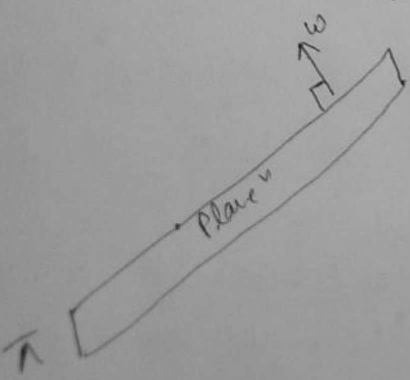
["$y = mx + c$"]

where 'm' is the "slope" & 'c' is the "y-intercept"

Equation of plane is :→ $\boxed{w^T x + b = 0}$ ['w' is normal to the plane]

if the plane passes through the origin. 'b' becomes '0'

$\therefore$ $\boxed{w^T x = 0}$ → [equation of plane passing through origin.]



Generally, a plane can be represented as :-

$\pi: \; w^T x + b = 0.$

where 'w' is $\perp$ to the plane, 'b' is the 'y intercept'

& 'x' is a datapoint belonging to d-dimensional space.

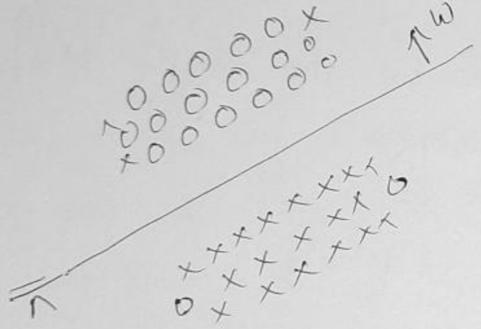
$$\left[\, x \in \mathbb{R}^d \;\; \& \;\; w \in \mathbb{R}^d \,\right]$$

$b \in \mathbb{R}^1$ (It is a scalar)

x & w are d-dimensional vectors

The big assumption that logistic regression makes is, the classes are almost or perfectly "linearly separable".

So, given a bunch of "+ive" & "-ive" points, we will find a plane which almost separates them linearly.



Given these Datapoints

$D_n = \{+ive, -ive\}$

we are supposed to find "w & b" (parameters of a plane)

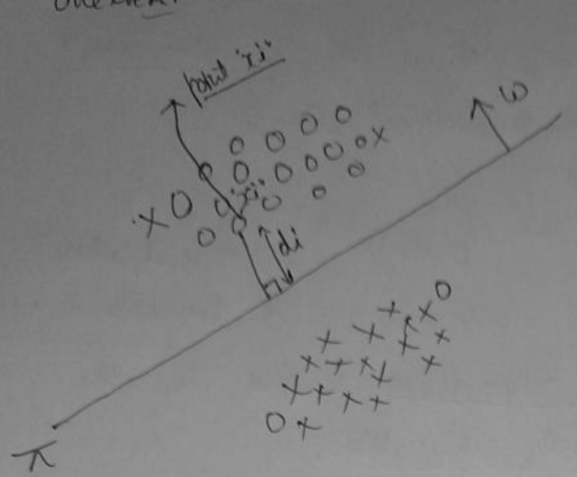
So, the task in logistic regression is to find 'w' & 'b' which corresponds to a plane, such that the plane best separates "-ive points" from the "+ive points".

Fundamental assumption here is classes are almost / perfectly linearly separable.

Let's look at the mathematical part of "logistic Regression":

P.T.O

let's have our positive & negative points along with some outliers.



O → +ive points
x → -ive points
& 'π' is the plane that we need to find out
'ω' is the normal to the plane.

let's assume a point "xi" & let's try to find its distance from the plane 'π', let that distance be 'di'

['di'] is the distance of point ["xi"] from the plane.

Till now we are saying that
1 : +ive pts
0 : -ive pts

Now let's change it slightly

$$\begin{bmatrix} y_i = +1 \rightarrow +\text{ive points} \\ y_i = -1 \rightarrow -\text{ive points} \end{bmatrix}$$

we will see this in a while.

$y_i \in \{-1, +1\}$ → Assumption

& di can be written as

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

Where 'ω' is the normal to the plane.

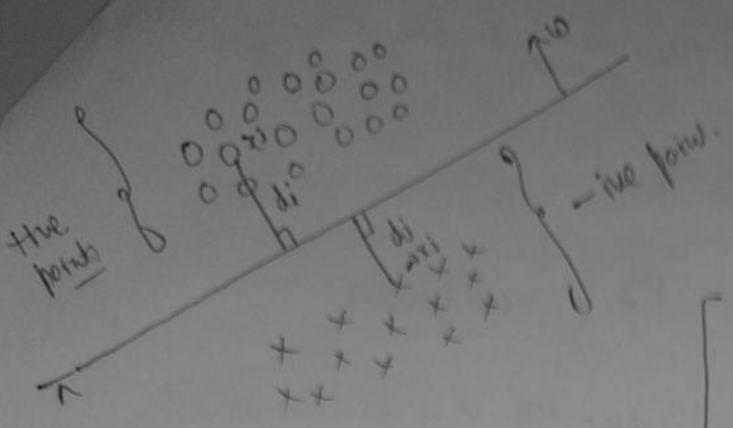
let's assume.

$\|\omega\| = 1$ → (Unit vector)

if 'ω' is a unit vector, we donot have to worry about $\|\omega\|$

So, if 'ω' is a unit vector then $(d_i = \omega^T x_i)$

let's now take another point which is lying in the opposite direction of 'ω'. let that point be "x_j"



$$\begin{bmatrix} d_i = w^T x_i \\ d_j = w^T x_j \end{bmatrix}$$

When we compute "d_i" which is basically "$w^T x_i$", since "w" & "x_i" are on the same side of the plane. we will get the value $\{ d_i = w^T x_i > 0 \}$

Now take $\{ d_j = w^T x_j \}$ since "x_j" is on the opposite side of "w" $\therefore$ $\{ d_j = w^T x_j < 0 \}$

So, we can say that, if we have found a plane, in such a way ~~that~~, if every point ~~that~~ is on the same direction as $\boxed{w}$ are all "positive points" & all the points which are lying in opposite direction of $\boxed{w}$ are all "negative points"

So, this is the final classifier that we have built, & it says

$$\left\{ \begin{array}{l} \text{if } w^T x_i > 0 \text{ then } y_i = +1 \\ \text{if } w^T x_i < 0 \text{ then } y_i = -1 \end{array} \right\} \rightarrow \text{This is our classifier}$$

There will be some mistakes as well, becoz some points which actually belongs to "-ive" class will be classified as "+ive" & vice-versa.

& our "decision-surface" is a line or a plane & it is a "linear surface".

P.T.O

for all +ive points, we said $\boxed{'y_i = +1'}$ & for all -ive points $\boxed{'y_i = -1'}$.

let's now see what happens to

Case ①

$y_i\, w^T x_i$

(for +ive points)

if $y_i = +1 \rightarrow$ +ive points

& $w^T x_i > 0 \Rightarrow$ classifier is saying that it is a +ive point]

then if $y_i * w^T x_i > 0$.

& $y_i = +1$

then plane it correctly classified the point

Case ②

if $y_i = -1 \rightarrow$ -ive points (for -ive points)

& $w^T x_i < 0 \Rightarrow$ Classifier is saying it is a -ive point.]

then $\boxed{y_i * w^T x_i > 0}$

since both of them are -ive, & multiplying two -ive val will become ~~zero~~. greater than zero.

So, for both '+ive' & '-ive' points.

if $y_i\, w^T x_i > 0 \Rightarrow$ [The LR model is correctly classifying the points x_i.]

Case ③

$y_i = +1$ (+ive data point)

& let's assume $w^T x_i < 0$ (It happens in case of an outlier)

This means 'LR' is saying point belongs to ['-ive class']

then $\{ y_i\, w^T x_i < 0 \}$ → This will be less than "Zero"

So when this happens, (our true class label is '+1' but LR is concluding it to be -1) which means it is a misclassified point

Case (4)

{ If $y_i = -1$ → (It is a -ve class point)
If $w^T x_i > 0$ → (LR is saying that x_i is a +ve point) }

{ $y_i w^T x_i < 0$. }

→ This means the point is also misclassified.

Eventually we want our classifiers to be very good.
Which means it should do "minimum number of misclassifications" or "maximum no. of correct classifications".

So, summary of Case 1 & Case 2 is, if the points are correctly classified then

$$\boxed{y_i w^T x_i > 0}$$

& Summary of Case 3 & Case 4 is, if the points are incorrectly classified then

$$\boxed{y_i w^T x_i < 0}$$

So, we want as many points as possible to have $y_i w^T x_i > 0$.

So, our task is to find a plane $\pi \rightarrow (w)$ or find "w" such that we have as many points correctly classified as possible.

Let's try to formulate it

Imagine we have n-data points. So, we want to maximize

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

{ If for every point $y_i w^T x_i$ is +ve, that is great }

Here $(x_i \,\&\, y_i)$'s are fixed becoz they come from our dataset "D_n"

[& we want to maximize it, so we want as many +ve values as possible & as few -ve values as possible.]

So, the only thing that is a variable here is "w^T".
. for maximz. the summation

So task is to vary 'w' (keep changing the plane) to find 'w' which maximizes the $(y_i w^T x_i)$.

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

'optimal w'

variable

[Mathematical Optimization problem that we want to solve.]