LECTURE - 47

HOW TO SOLVE/HANDLE OVERFITTING & UNDERFITTING ?

HANDLING OVERFITTING:

↳ Reduce the network's capacity by removing layers or reducing the number of elements in the hidden layers.

↳ Apply regularization, which comes down to adding a cost to the loss function for large weights/values.

↳ Use Dropout layers, which will randomly remove certain features by setting them to zero.

HANDLING UNDERFITTING:

↳ Get more training data.

↳ Increase the size or number of parameters in the model.

↳ Increase the complexity of the model.

↳ Increase the training time, until the cost function is minimised.

We know that,

→ A line eq^n in 2D is $ax_1 + by + c = 0$

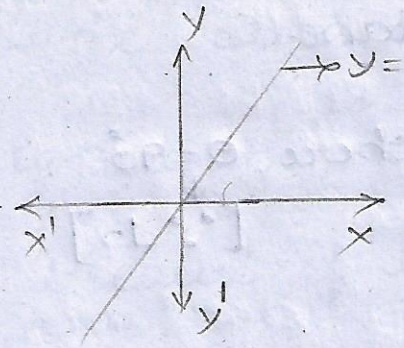
$\Rightarrow w_1x_1 + w_2x_2 + w_0 = 0 \rightarrow ①$

→ A line eq^n in 4D is

$\Rightarrow w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 = 0 \rightarrow ②$

$\Rightarrow$ ② can also be written as

$$\sum_{i=1}^{d} w_i x_i + w_0 = 0$$



→ From the graph we can say that a line is passing through (0,0)

⇒ $y = mx + c$

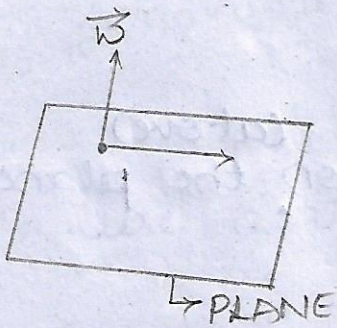
→* c tends to be zero.

⇒ w_0 is zero.

Let us consider a plane is passing through (0,0,0); then

⇒ ② ⇒ $x_2 = \frac{-w_1 x_1}{w_2} - \frac{w_3 x_3}{w_2} - \frac{w_0}{w_2}$ ($w_0 \to 0$)

⇒ $x_2 = \frac{-w_1 x_1}{w_2} - \frac{w_3 x_3}{w_2}$



→ $w_1x_1 + w_2x_2 + w_3x_3 = 0$ can also be written as $\vec{w} \cdot \vec{x} = 0$

where $\vec{w}$ is perperdicular to the plane and also shows the planes facing

$\vec{x}$ is any point on the plane.

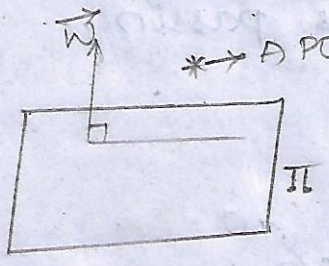
then, $\|w\| \|x\| \cos\theta_{x,w} = 0$ where $\theta = 90°$ [$\because \perp r$]

If we have a eqn of $2x + 3y - z = 4$ then

$\vec{w} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$ where,

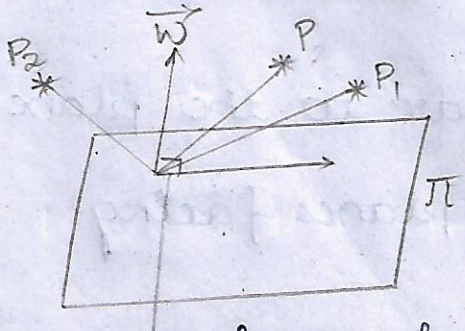
$\vec{w}$ – NORMAL/NORM OF THE PLANE (or)

HYPERPLANE → if it is in higher dimension.



From the above figure, where does the point lie i.e., above the plane or below the plane ?

ANS: The point is present on (above) the plane.



→ $\vec{w} \cdot \vec{p} = \|w\| \|p\| \cos\theta$

↓

the o/p is +ve

→ the angle is between 0° & 90°

↳ the values of w & p never be "-ve".

↳ But the dot product will be "-ve"

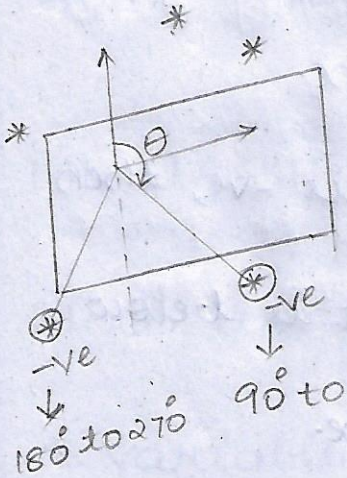
because of the "θ value"

⇒ $\vec{w} \cdot \vec{P_1}$ → the O/p is +ve

⇒ $\vec{w} \cdot \vec{P_2}$ → the O/p is +ve.

NOTE:

→ the normal to the plane and the dot Product is "+ve" if the point is above the plane.



→ In this case, the value of θ lies between 90° to 180°.

→ this clearly states that the dot product is "-ve".

We use a line/hyperplane/plane where

In linear Regression & logistic Regression

there is a bit change.

LINEAR REGRESSION

→ Here,

$y_{pred} = mx + c$

where mx can be written as $w^T \cdot x$.

→ error = $y_{act} - y_{pred}$

→ Best fitting means minimizing the errors

⇒ $min \cdot \Sigma e^2$

LOGISTIC REGRESSION

→ Here,

$y_{Pred} = sign(mx + c)$

where mx can be written as $w^T \cdot x$.

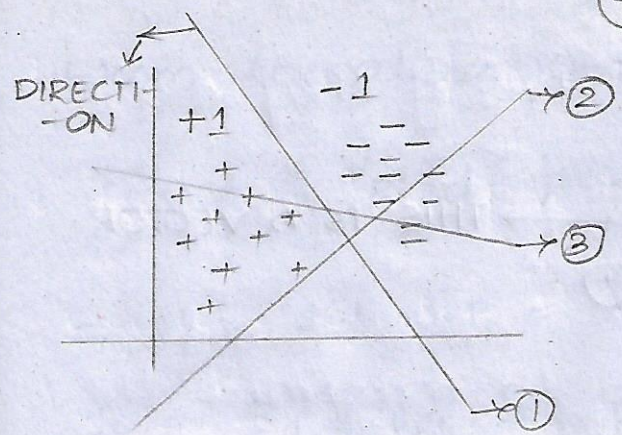
→ sign ↔ '+/-'

→ If it is +ve, then the point is above the plane.

→ If it is -ve, then the point is below the plane.

Over here, we need to understand the misclassification i.e.,

MISCLASSIFICATION = $y_{act} * y_{pred}$



In this case line ① is the best seperator of a hyper plane.

⇒ y_{act} → positive = +1 & negative = -1

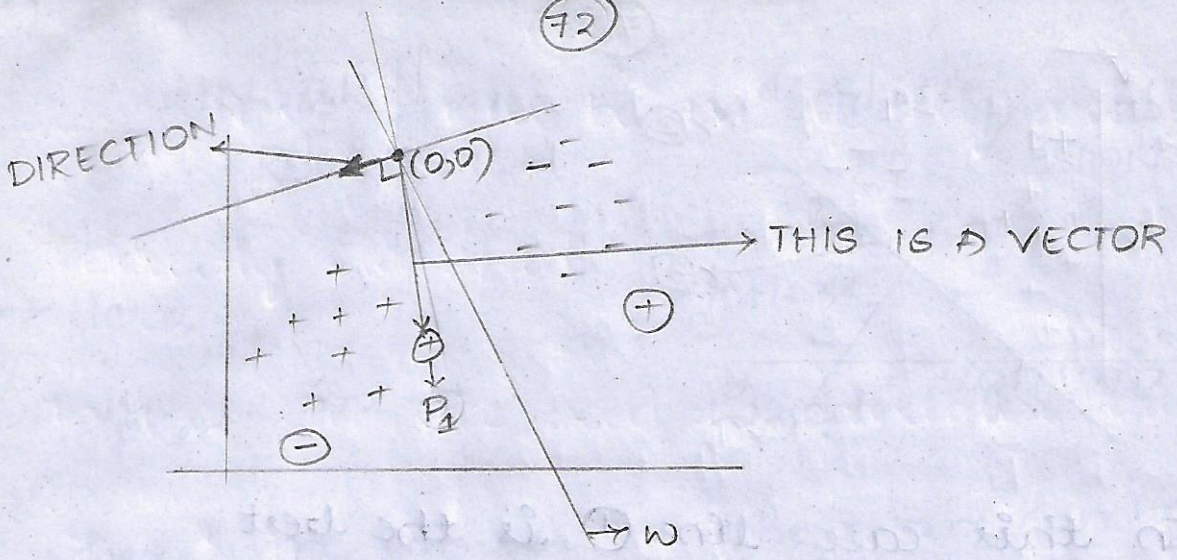
NOTE:

↳ the seperation will never be clean.

MISCLASSIFICATION:

When all classes, groups or categories of a variable have the same error rate or Probability of being misclassified then it is said to be misclassification.

↳ SVM algorithm can be used for the analysis of misclassification.



In this case, there are 2 misclassifiers.

If we talk about the point P_1,

↳ $\vec{w}$ & the plane are at 90°.

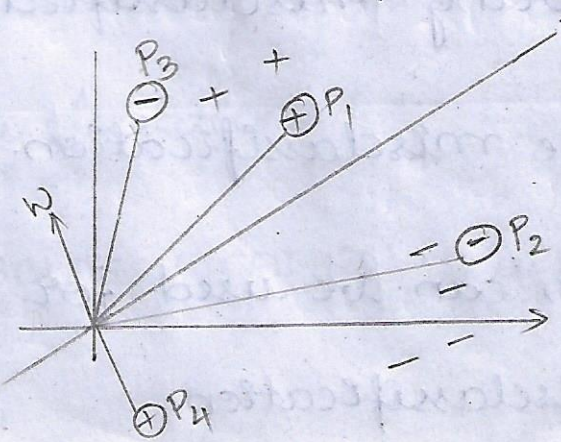
↳ For $P_1 \rightarrow \vec{w} \cdot \vec{P_1} \Rightarrow$ the o/p is +ve. because θ lies b/w 0° & 90°.

$\Rightarrow \|\vec{w}\| \cdot \|\vec{P_1}\| \cdot \cos\theta_{w,P_1}$

Here P_1 is y_{Pred}.

So, we can say that the Norm of the plane and the point has the "+ve" value i.e.,

$\Rightarrow y_{act} = 1$ & $y_{Pred} = \vec{w} \cdot \vec{P_1}$



From the above graph, lets check the $y_{act} * y_{pred}$ for each point.

$\Rightarrow P_1 = 1 * w \cdot P_1 =$ +ve value

$\Rightarrow P_2 = -1 * -w \cdot P_2 =$ +ve value.

In this case, cos θ value is -ve since θ value is $> 90^\circ$ & $< 180^\circ$.

$\Rightarrow P_3 = -1 * w \cdot P_3 =$ -ve value.

In this case, cos θ value is +ve since θ lies between 0° and 90° i.e., $> 0^\circ$ & $< 90^\circ$.

$\Rightarrow P_4 = +1 * -w \cdot P_4 \rightarrow$ -ve value.

In this case, cos θ value is -ve since θ value is $> 180^\circ$ & $< 270^\circ$.

∴ From P_1, P_2, P_3, P_4 we can say that

↳ P_1, P_2 are in correct positions → +ve value

↳ P_3, P_4 are misclassified → -ve value.

→ $y_{act} * y_{pred}$ will be "+ve" if they are correctly classified.

In this case, we need to maximise the $y_{act} * y_{pred}$.

$$\Rightarrow \max \Sigma\, y_{act} * y_{pred}$$

$$\Rightarrow m^*, c^* = \underset{m,c}{\arg} \max \Sigma \left(y_{act} * y_{pred}\right)$$

$$\Rightarrow m^*, c^* = \underset{m,c}{\arg} \max \Sigma \left(y_{act} * \{mx + c\}\right)$$

The above eq^n. is used in classification.

$$\Rightarrow m^*, c^* = \underset{m,c}{\arg} \max \Sigma \left(y_{i\,act} * \{mx_i + c\}\right)$$

can also be written as

$$\Rightarrow \boxed{w^*, w_0^{\to 0} = \underset{w, w_0^{\to 0}}{\arg} \max \Sigma \left(y_{i\,act} * \{w^T \cdot x\}\right)}$$

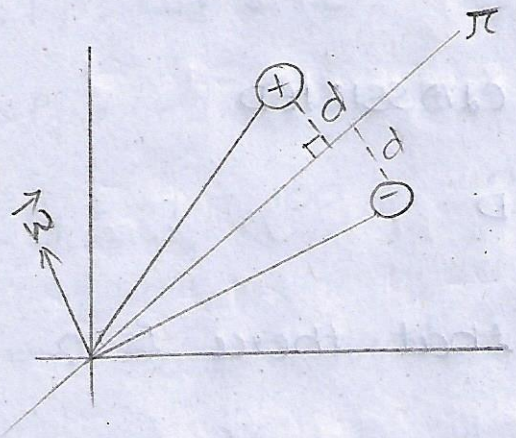
[∵ From line eq^n w_0 tend to be '0' ∴ w_0 is a scalar quantity]

→ w, w_0 will be same i.e., scalar quantity if it is in 2D.

→ If $\hat{w}$ - vector quantity & w_0 - scalar quantity then it is said to be in higher dimensions.

NOTE:

w - vector quantity & w_0 - scalar quantity applies same in LINEAR REGRESSION.



→ In this case,

$y_{pred} = w \cdot p$ i.e.,

the "DISTANCE OF A POINT FROM THE PLANE".

$$\Rightarrow d = \frac{w \cdot p}{\|w\|}$$

where $\|w\| = 1$ → since unit vector.

⇒ $y_{act} * y_{pred}$ is called as the "SIGNED DISTANCE".

SIGNED DISTANCE:

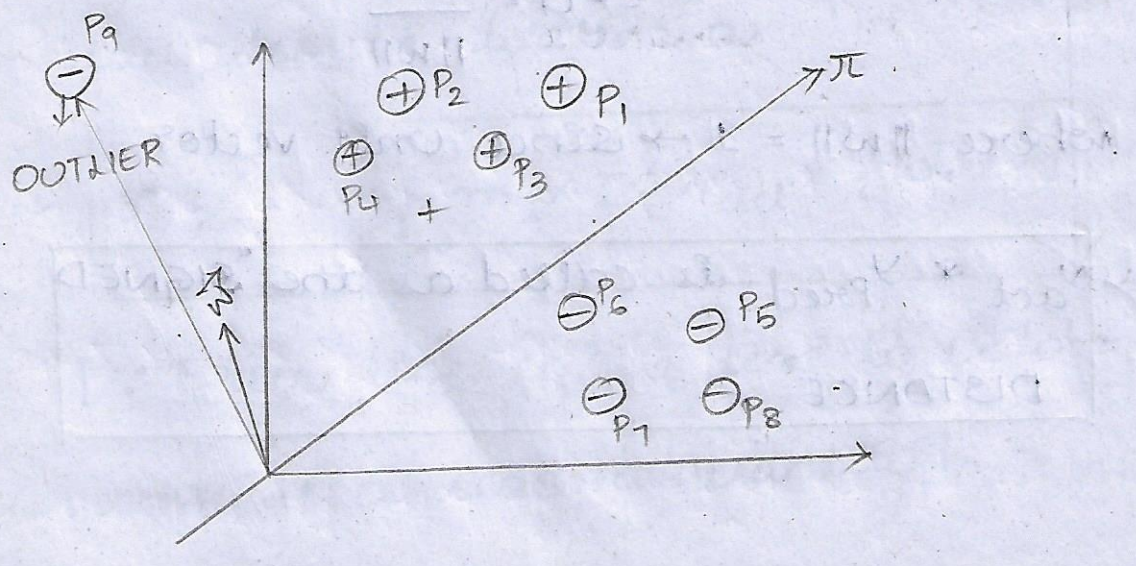
When passed the coordinates of a point in space, return the shortest distance between that point and some surface.

→ the sign of return value indicates whether the point is inside of that surface (or) outside of that surface.

↦ If +ve - CORRECTLY CLASSIFIED

↳ If -ve - MISCLASSIFIED

Now let us consider that there is a misclassified point



From the above graph, lets calculate

$y_{act} * y_{pred}$

$\Rightarrow P_1 = +1 * +2 = +2$

$\Rightarrow P_2 = +1 * +2 = +2$

$\Rightarrow P_3 = +1 * +1 = +1$

$\Rightarrow P_4 = +1 * +1 = +1$

$\Rightarrow P_5 = -1 * -1 = +1$

$\Rightarrow P_6 = -1 * -1 = +1$

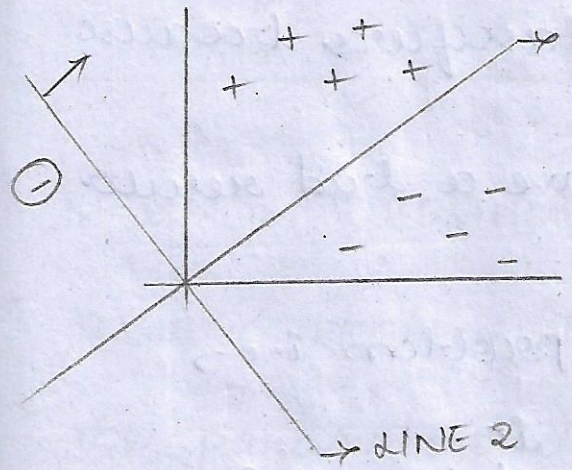
$\Rightarrow P_7 = -1 * -2 = +2$

$\Rightarrow P_8 = -1 * -2 = +2$

$\Rightarrow P_9 = -1 * 100 = -100$

$\Rightarrow \Sigma y_{act} * y_{pred} = -88$

[y_{pred} = HOW FAR IS THE POINT]

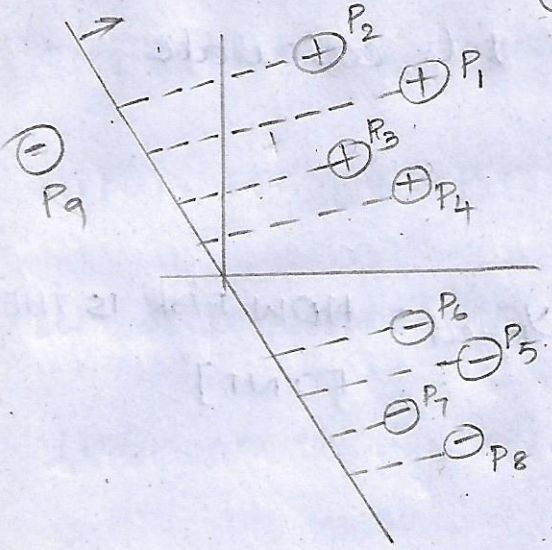


→ In this case,

Misclassification done by

↳ LINE 1 - 1

↳ LINE 2 - 4



In this case,

$y_{act} * y_{pred}$

$\Rightarrow P_1 = +1 * +3 = +3$

$\Rightarrow P_2 = +1 * +2 = +2$

$\Rightarrow P_3 = +1 * +2 = +2$

$\Rightarrow P_4 = +1 * +1 = +1$

$\Rightarrow P_5 = -1 * +3 = -3$

$\Rightarrow P_6 = -1 * +1 = -1$

$\Rightarrow P_7 = -1 * +2 = -2$

$\Rightarrow P_8 = -1 * +4 = -4$

$\Rightarrow P_9 = -1 * -1 = 1$

$\therefore \Rightarrow \Sigma y_{act} * y_{pred} = -1$ which is remaining

∴ From LINE 2 we can say that we have maximum misclassifiers because of one outlier, we have a bad result.

$\Rightarrow \Sigma y_{act} * y_{pred}$ has a problem i.e., because of the outlier in data, this

optimization eqn is giving an incorrect result.

→ The reason is, LINE 1 has 1 misclassifier. i.e., outlier
and LINE 2 has 4 misclassifiers.

↳ In order to solve the $\Sigma y_{act} * y_{pred}$ eqn we use Gradient Descent.

↳ In this case we will minimise the eqn to maximise state.

i.e., in linear Regression → $\min \Sigma e^2$ & also called as CONVEX OPTIMIZATION PROBLEM

↳ In logistic Regression → max Σ signed distance.

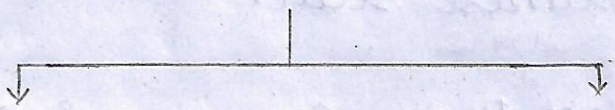
MAXIMUM LIKELIHOOD ESTIMATION (MLE) OF LOGISTIC REGRESSION:

The parameters of a logistic regression model can be estimated by the

Probabilistic framework called maximum likelihood estimation.

↳ the parameters of the model can be estimated by maximizing a likelihood function that predicts the mean of a Bernoulli Distribution for each example.

MAXIMUM LIKELIHOOD ESTIMATION (MLE)



GEOMETRICAL WAY

PROBABILISTIC WAY

As probabilistic way is a bit difficult we go with Geometrical way.

WHY DO WE USE MLE ?

For Logistic regression, least squares estimation is not capable of producing

minimum variance unbiased estimators for the actual parameters.

↳ In its place, maximum likelihood estimation is used to solve for the parameters that best fits the data.

↳ In other words, it is a technique used for estimating the parameters of a given distribution, using some observed data.