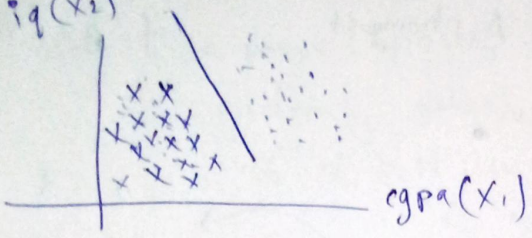
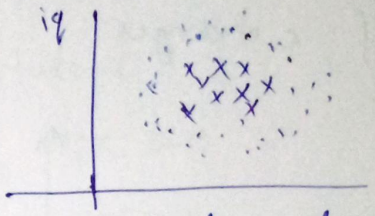
04-Mar-2024

Logistic Regression

- The prerequisits of apply the logistic Regression is that the data is linearly classify.



linearly separable data



non-linearly separable data

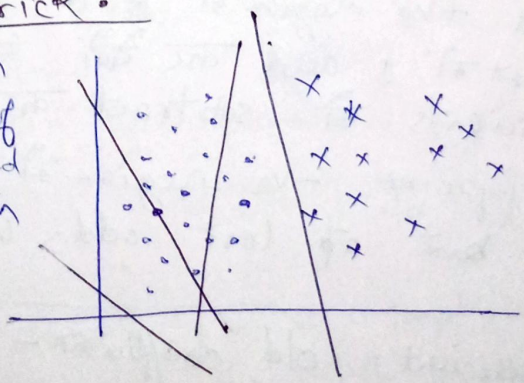
- in logistic Regression we use $AX + By + C = 0$

or $AX_1 + BX_2 + C = 0$ for two-D

$AX_1 + BX_2 + CX_3 + D = 0$ for 3-D

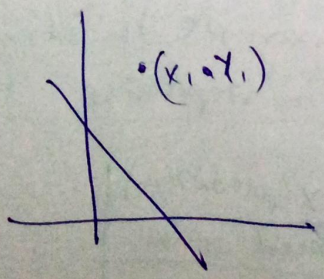
perceptron Trick:

loop start with random values of A, B and C and changes in each loop.



loop continues till convergence (all points are correctly classified.

How to label region?



equation of line $Ax + By + C = 0$

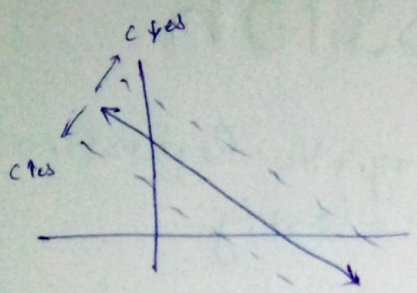
if $Ax_1 + By_1 + C > 0$ +ve region.

if $Ax_1 + By_1 + C < 0$ -ve region

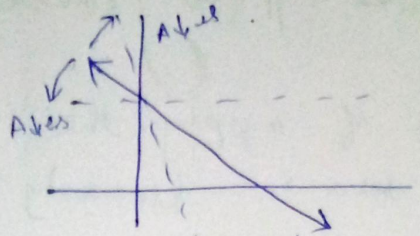
if $Ax_1 + By_1 + C = 0$ on line.

How to transform line?

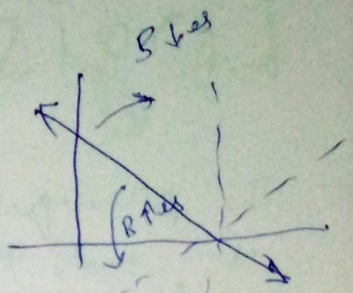
$Ax + By + c = 0$



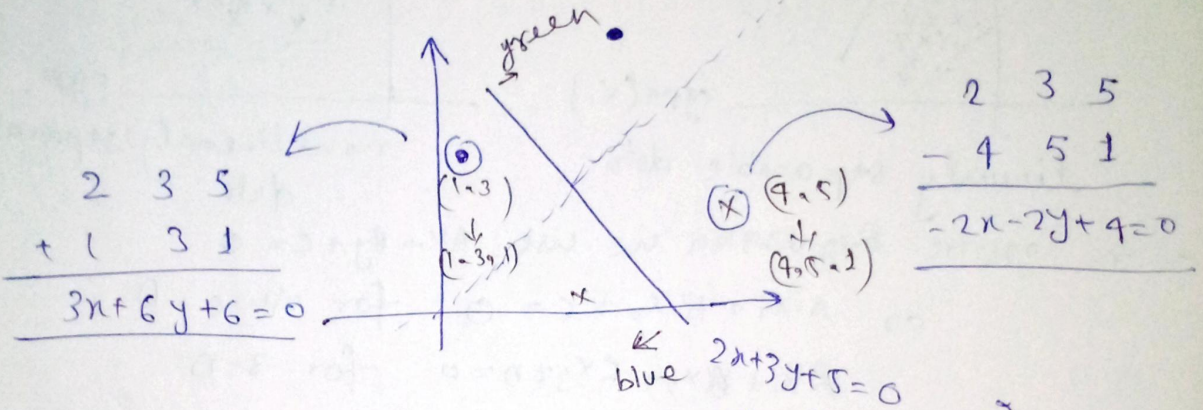
if c changes



if A changes



if B changes



$$\begin{array}{r} 2\ \ 3\ \ 5 \\ +1\ \ 3\ \ 1 \\ \hline 3x+6y+6=0 \end{array}$$

$$\begin{array}{r} 2\ \ 3\ \ 5 \\ -4\ \ 5\ \ 1 \\ \hline -2x-2y+4=0 \end{array}$$

अगर -ve point +ve region में है तो हम उसके coordinates में last में 1 add कर देंगे और फिर line के coefficients से subtract कर देंगे।

और अगर +ve point -ve region में है तो same process as above but at last add. with line's coefficient.

new coefficient = old coefficient − η coordinates.

learning rate (η) = 0.01

in real life The long transformation example in above we can't do that, we do this in stepwise

Algorithm

X_0	X_1	X_2	Y (prediction)
1	cgpa	iq	placed or not
1	7.5	81	1
1	8.9	109	1

$AX + BY + C = 0$

$W_0 X_0 + W_1 X_1 + W_2 X_2 = 0$

$W_0 = C$, $W_1 = A$, $W_2 = B$

if $\sum_{i=0}^{n} W_i X_i \geq 0$ then I

else $\sum_{i=0}^{n} W_i X_i < 0$ then C

epoch = 1000 $\quad \eta = 0.01$

for i in range (epochs):

random ly select a student

if $X_i \in N$ and $\sum_{i=0}^{n} W_i X_i \geq 0$

here if model predic 1 but in data it is zero then here we update the W's

-ve point ≠ ve region ∴ $\frac{\partial}{\partial \epsilon}$

$W_{new} = W_{old} - \eta [Y_0 X_1 X_2]$

$W_{new} = W_{old} - \eta X_i$

or

if $\Sigma_i \in P$ and $\sum_{i=1}^{n} W_i X_i < 0$ +ve point −ve region ∴ $\frac{\partial}{\partial \epsilon}$

$W_{new} = W_{old} + \eta X_i$

$X_i \in N$ and $\Sigma W_i X_i \geq 0$

$W_n = W_0 - \eta X_i$ —(i)

$X_i \in P$ and $\Sigma W_i X_i < 0$

$W_n = W_0 + \eta X_i$ —(ii)

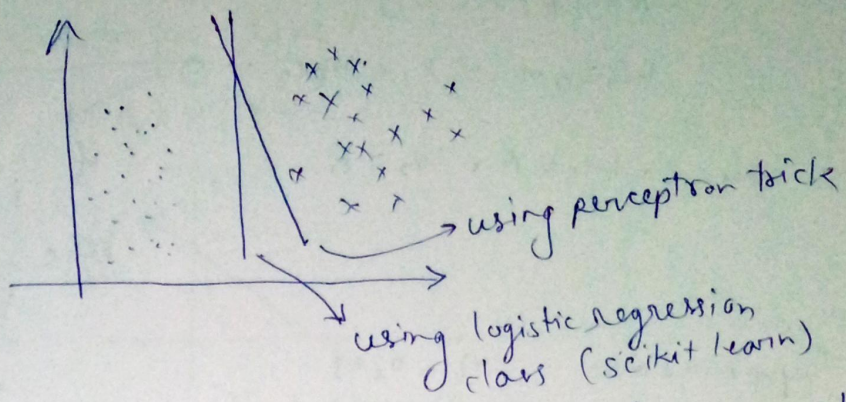
simplified formula

$W_n = W_0 + \eta (Y_i - \hat{Y}_i) X_i$ —(iii)

actual prediction

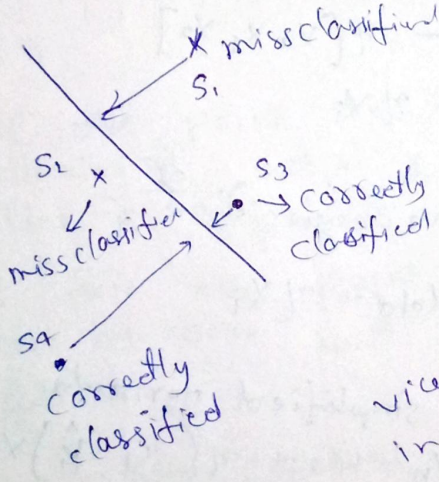
	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
case-1	1	1	0
case-2	0	0	0
case 3	1	0	1
case 4	0	1	-1

in case 1 and 2 no updates in weight but in case 3 and case-4 weight equation becomes eqn (ii) and (i)



If we want the perceptron trick model is work as the logistic regression class then here we also considered the correctly classified ~~class~~ point also ~~Hard~~ correctly classified point line को ~~दूर~~ उपयोग से दूर ~~भेजेंगे~~ Till the ~~eq~~ equilibrium phase.

misclassified — line pull
correctly classified — line push



in misclassified.
a point have less distance (S_2)
less magnitude to pull
a point have large distance (S_1)
large manitude to pull

vice-versa
in correctly classified point

* for all of these above approach our algo (~~perceptron~~ perceptron trick) is work similar as the logistic regression class.

* if we apply the above approach then our plan have to make changes in the equation (iii)
or we can also say that the case 3 and case-4

$$W_n = W_0 + \eta(Y_i - \hat{Y}_i) X_i$$

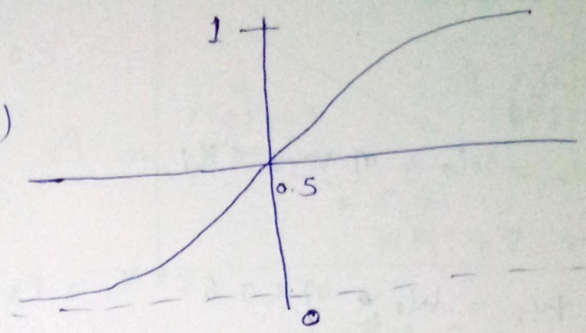
↓
come zero due to this

$Y_i - \hat{Y}$ is come zero due to $\sum w_i x_i = 0$ this function that is step function that output always the 0 or 1 hence now we use the sigmoid function.

sigmoid function

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \text{---(iv)}$$

if $-\infty < z < \infty$
then $0 < y < 1$



$$z = \sum w_i x_i$$

if z +ve then $\sigma(z) > 0.5$

if z -ve then $\sigma(z) < 0.5$

steps:- first calculate z

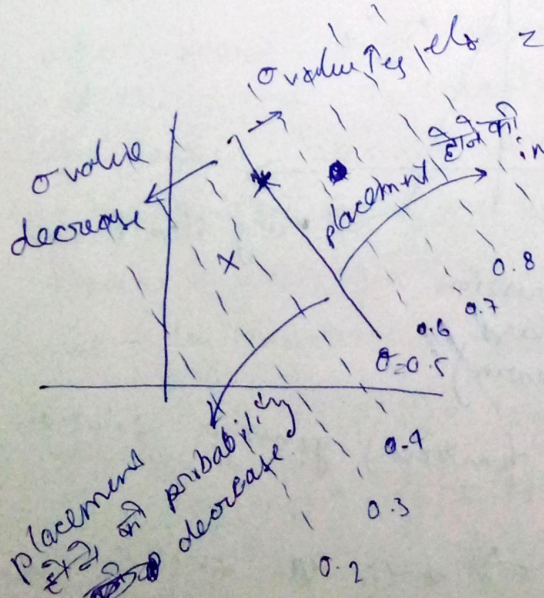
step-2 - put this z in equation (iv) and calculate

step-3 if ~~$\sigma(z) = 0.5$~~

output

z +ve then $\sigma(z) \geq 0.5$ → 1

z -ve then $\sigma(z) < 0.5$ → 0



online $\sum w_i x_i = 0$ $\sigma = 0.5$

for $\sum w_i x_i > 0$ $\sigma > 0.5$

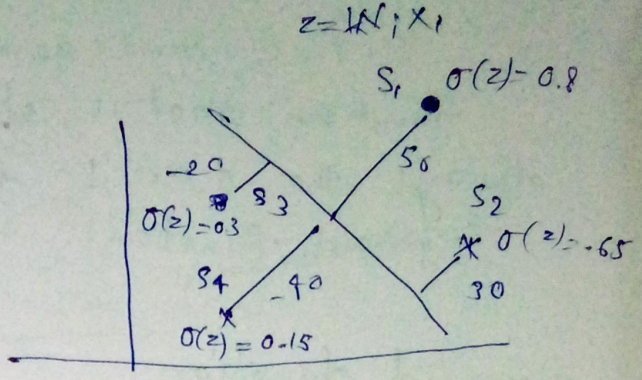
$\sigma = 0.5$ means $P(0.5) =$
it means placement होने की probability is 0.5

impact on Sigmoid

$$W_n = W_0 + \eta(Y_i - \hat{Y}_i) X_i$$

$Y_i = \sigma(z)$

where $z = \sum W_i X_i$



Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	0.8	0.2
0	0.65	-0.65
1	0.3	0.7
0	0.15	-0.15

for S_1

$W_n = W_0 + \eta \times 0.2 \times X_i = P_1$

for S_2

$W_n = W_0 + \eta \times 0.65 \times X_i = P_3$

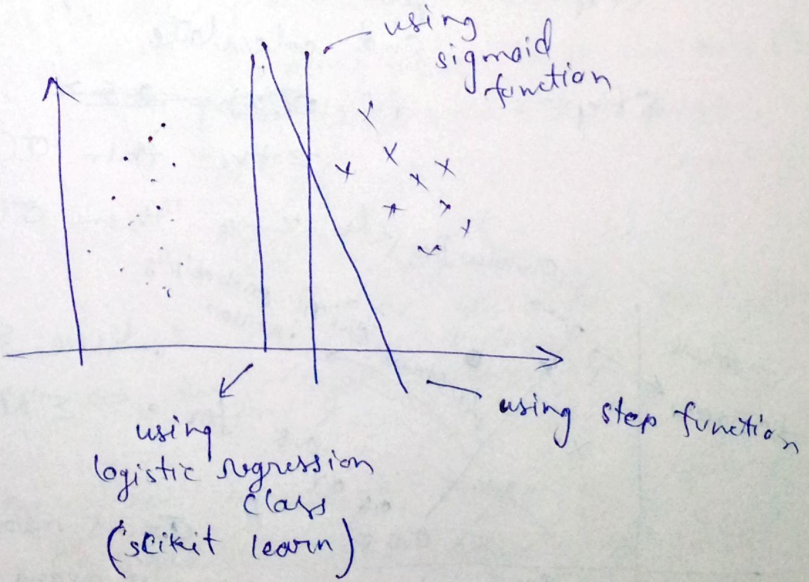
for S_3

$W_n = W_0 + \eta \times 0.7 \times X_i = P_2$

for S_4

$W_n = W_0 - \eta \times 0.15 \times X_i = P_4$

$P_2 > P_1$

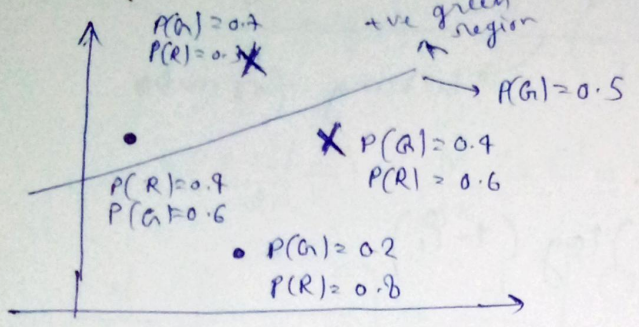


* regression line (using sigmoid function) gives a solution but it is not a optimal.
* इसीलिए हम loss function calculate करते समय यहां करने के लिए कि हमारा model कितना गलती कर रहा है,

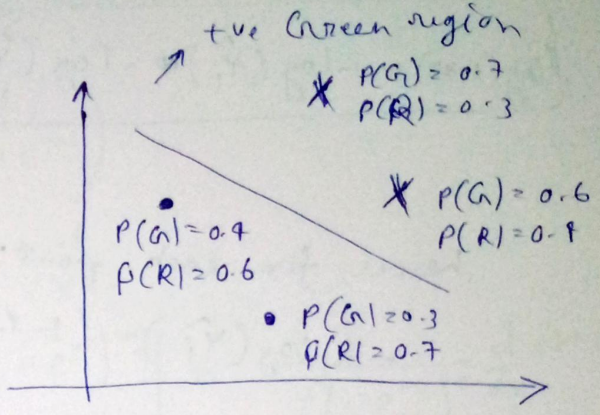
Here we now for deriving the loss function we use the maximum likelihood.

Loss Function Calculation:

maximum likelihood



model-1



model-2

* in the above $\hat{Y}_i$ is calculated as a probability term

model-1 (maximum likelihood)

$0.7 \times 0.4 \times 0.4 \times 0.8 = 0.089$

model-2 (maximum likelihood)

$0.7 \times 0.6 \times 0.6 \times 0.7 = 0.176$

* here the model 2 is better than the model-1
* in the above it is easy to calculate due to less number

what about large dataset?

at the time of calculation of maximum likelihood the values becomes very less and then this comparison (compare b/w two model) is useless

hence we again re-formulate the formula and use log() function

$$\log(a*b) = \log a + b$$

$$\log(\text{max}) = \log(0.7) + \log(0.4) + \log(0.4) + \log(0.8)$$

$$\downarrow$$

$$\log(\text{max}) = -\log(0.7) - \log(0.4) - \log(0.4) - \log(0.8) \rightarrow \text{cross entropy}$$

cross entropy में हम values को minimize करते है, मतलब वह model Best होगा। जिसका cross entropy minimum होगा।

$$\log(\max) = -\log(\hat{Y}_1) * -\log(\hat{Y}_2) - \log(\hat{Y}_3) * -\log(\hat{Y}_4)$$

→ wrong formula

hence for each point:

$$-Y_i \log(\hat{Y}_i) - (1-Y_i)\log(1-\hat{Y}_i)$$

$$\text{Loss function}(L) = \sum_{i=1}^{n} -Y_i \log(\hat{Y}_i) - (1-Y_i)\log(1-\hat{Y}_i)$$

for average

$$L = \frac{-1}{n} \sum_{i=1}^{n} Y_i \log(\hat{Y}_i) + (1-Y_i)\log(1-\hat{Y}_i)$$

the above equation is called log loss error or Binary cross entropy

→ minimize it using and find w_1, w_2, w_n such that is it have minimum value

* for calculating closed form of the above loss function is using gradient descent

---

$$\text{Loss function}(L(w_1, w_2, b)) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(x_i)) + \alpha R(w_1, w_2)$$

↳ Regularization

$$E(W, b) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(x_i)) + \alpha R(w)$$

Here L = loss function and R = Regularization

for perceptron — $L(Y_i, f(x_i)) = \max(0, -Y_i f(x_i))$

Here $f(x_i) = w_1 x_1 + w_2 x_2 + b$; n = rows

Derivative of Sigmoid function:

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

$$\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right)$$

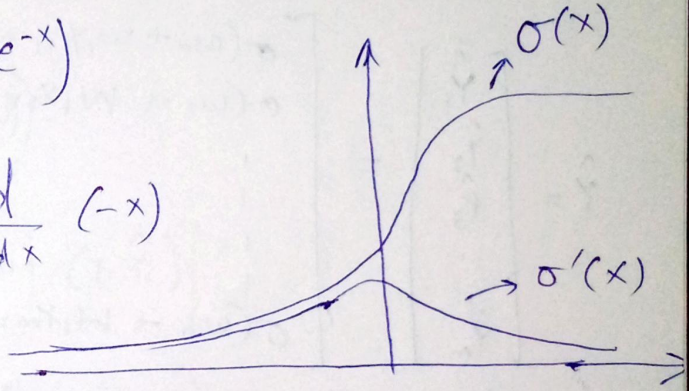
$$\because \frac{d}{dx}\left(\frac{1}{x}\right) = \frac{d}{dx}(x)^{-1} = -\frac{1}{x^2}$$

$$\frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right] = \frac{d}{dx}\left[(1+e^{-x})^{-1}\right] = \frac{-1}{(1+e^{-x})^2}\frac{d}{dx}(1+e^{-x})$$

$$= \frac{-1}{(1+e^{-x})^2}\frac{d}{dx}(e^{-x})$$

$$= \frac{-e^{-x}}{(1+e^{-x})^2}\frac{d}{dx}(-x)$$

$$= \frac{e^{-x}}{(1+e^{-x})^2}$$



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

$$= \sigma(x)\left[\frac{e^{-x}}{1+e^{-x}}\right] = \sigma(x)\left[\frac{1+e^{-x}-1}{1+e^{-x}}\right]$$

$$= \sigma(x)\left[\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right]$$

$$\sigma'(x) = \sigma(x)\left[1-\sigma(x)\right] =$$

Gradient Descent on logistic Regression's Loss Function:

Dataset: Rows = m columns = n

1	2	3	----------	n	Y
x_{11}	x_{12}	x_{13}	----------	x_{1n}	Y_1
x_{21}	x_{22}	x_{23}	----------	x_{2n}	Y_2
⋮	⋮	⋮			
x_{m1}	x_{m2}	x_{m3}	----------	x_{mn}	Y_m

$$\hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \hat{Y}_3 \\ \vdots \\ \hat{Y}_m \end{bmatrix} = \begin{bmatrix} \sigma(w_0 + w_1 x_{11} + w_2 x_{12} + \cdots + w_n x_{1n} \\ \sigma(w_0 + w_1 x_{21} + w_2 x_{22} + \cdots + w_n x_{2n} \\ \vdots \\ \sigma(w_0 + w_1 x_{m1} + w_2 x_{m2} \cdots + w_n x_{mn} \end{bmatrix}$$

$$\hat{Y} = \sigma(X w)$$

$$\hat{Y} = \sigma\left(\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \vdots \\ 1 & x_{m1} & x_{m2} & & x_{mn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \right)$$

$$L = -\frac{1}{m} \sum_{i=1}^{m} Y_i \log(\hat{Y}_i) + (1 - Y_i) \log(1 - \hat{Y}_i)$$

or

$$L = -\frac{1}{m} \left[\sum_{i=1}^{m} Y_i (\log(\hat{Y}_i) + \sum_{i=1}^{m} (1 - Y_i) \log(1 - \hat{Y}_i) \right]$$

for 1st term

$$\sum_{i=1}^{m} Y_i \log(\hat{Y}_i) = Y_1 \log(\hat{Y}_1) + Y_2 \log(\hat{Y}_2) + Y_3 \log(\hat{Y}_3) + \dots Y_m \log(\hat{Y}_m)$$

$$= \begin{bmatrix} Y_1 & Y_2 & Y_3 & \dots & Y_m \end{bmatrix} \cdot \begin{bmatrix} \log \hat{Y}_1 \\ \log \hat{Y}_2 \\ \vdots \\ \log \hat{Y}_m \end{bmatrix}$$

$$= \begin{bmatrix} Y_1 & Y_2 & \dots & Y_m \end{bmatrix} \cdot \log\left(\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_m \end{bmatrix} \right)$$

$$= Y \cdot \log(\hat{Y})$$

$$= Y \cdot \log(\sigma(XW))$$

for 2nd term

$$\sum_{i=1}^{m} (1-Y_i) \log(1-\hat{Y}_i)$$

$$= (1-Y) \log(\sigma(1-XW))$$

then Naru whole formula for loss function

$$L = \frac{-1}{m} \left[Y \log \hat{Y} + (1-Y) \log(1-\hat{Y}) \right]$$

when $\hat{Y} = \sigma(X \cdot W)$

Here the dot product of X and W

loss function in matrix form:

$$\text{loss function} = \frac{-1}{m}\left[Y \log\left(\sigma(WX)\right) + (1-Y)\log\left(1-\sigma(WX)\right)\right]$$

यहाँ पर हमे सारे coefficients (w) का एक ऐसा value निकालना है जहाँ पर loss function is minimum. to solve this problem we use gradient descent.

approach is

- start with random value of $w = [\quad]$
- for i in epochs:

$$W = W - \eta \frac{\partial L}{\partial W} \quad \text{---- (i)}$$

$$\frac{\partial L}{\partial w_0}, \frac{\partial L}{\partial w_1}, \frac{\partial L}{\partial w_2} \cdots \frac{\partial L}{\partial w_n}$$

$$L = \frac{-1}{m}\left[\underbrace{Y \log \hat{Y}}_{1^{st}} + \underbrace{(1-Y)\log(1-\hat{Y})}_{2^{nd}}\right]$$

$$\frac{dL}{dw} = ?$$

1st term

$$\frac{d}{dw} Y \log \hat{Y} = Y \frac{d}{dw} \log \hat{Y}$$

$$= \frac{Y}{\hat{Y}} \frac{d}{dL}(\hat{Y})$$

$$= \frac{Y}{\hat{Y}} \frac{d}{dL}\left(\sigma(WX)\right)$$

$$= \frac{y}{\hat{y}} \cdot \sigma(wx)\left[1-\sigma(wx)\right]\frac{d}{dw}(wx)$$

$$= \frac{y}{\hat{y}} \cdot \hat{y}(1-\hat{y})x$$

$$= y(1-\hat{y})x$$

2nd term $\frac{d}{dw}(1-y)\log(1-\hat{y}) = (1-y)\frac{d}{dw}\log(1-\hat{y})$

$$= \frac{(1-y)}{(1-\hat{y})}\frac{d}{dw}\left[1-\hat{y}\right]$$

$$= -\frac{(1-y)}{(1-\hat{y})}\frac{d}{dw}\sigma(wx)$$

$$= -\frac{(1-y)}{(1-\hat{y})}\left[\sigma(wx)\cdot\left[1-\sigma(wx)\right]\frac{d}{dw}(wx)\right.$$

$$= \frac{-(1-y)}{(1-\hat{y})}\hat{y}(1-\hat{y})x$$

$$= -\hat{y}(1-y)x$$

Now the whole term is

$$\frac{dL}{dw} = \frac{-1}{m}\left[y(1-\hat{y})x - \hat{y}(1-y)x\right]$$

$$= \frac{-1}{m}\left[y(1-\hat{y}) - \hat{y}(1-y)\right]x$$

$$= -\frac{1}{m}\left[y - y\hat{y} - \hat{y} + y\hat{y}\right]x$$

$$\frac{\partial L}{\partial w} = -\frac{1}{m}(Y - \hat{Y})X$$

replace this equation in (i)

$$w = w + \eta \frac{1}{m}(Y - \hat{Y})X$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}_{(n+1,1)} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}_{(m,n+1)} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}_{(m,1)} \quad \hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_m \end{bmatrix}_{(m,1)}$$