Logistic Regression

Geometry
Probability

Eq requirement.

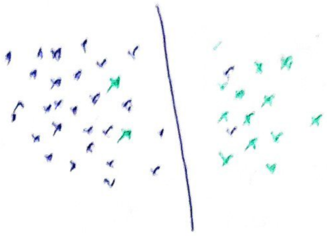
Perceptron Trick

Random $A=1, B=1, C=0$

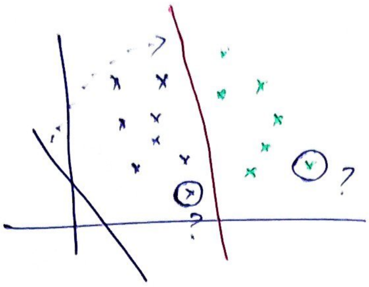
- Ask a point this line ok.
- No.
- line shift towards that point.
- it happens until.

• 2D $Ax_1 + Bx_2 + C = 0$

• 3D
$Ax_1 + Bx_2 + Cx_3 + d = 0$

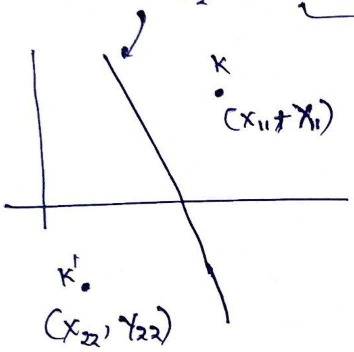


- Data can be separate easily.

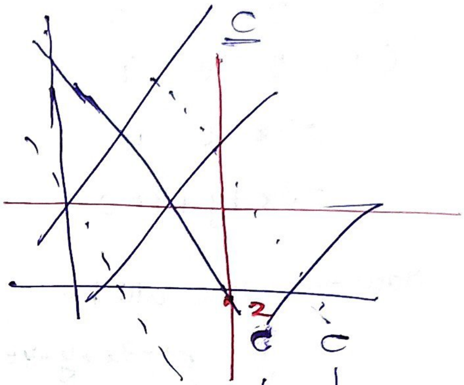


How to label region?

$Ax_1 + Bx_2 + C = 0$ ⟶ put both poin K, K'



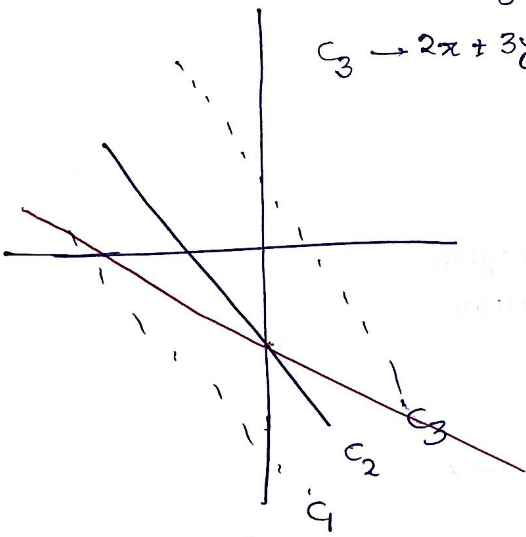
if
• $Ax_{11} + By_{11} + C < 0$
then +ve region
• > 0
-ve region



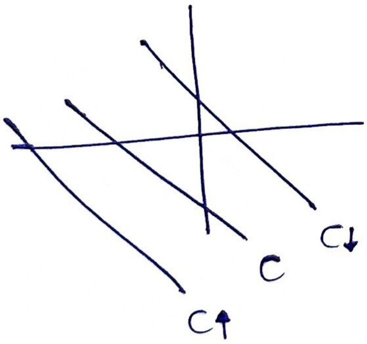
$C_1 - 2x + 3y + 10 = 0$

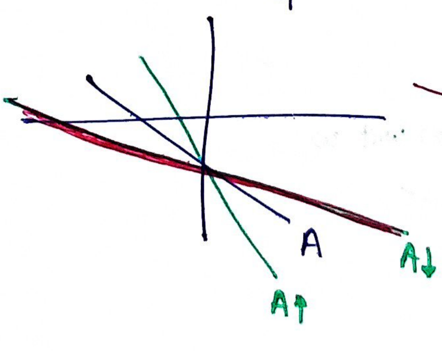
$C_2 \rightarrow 2x + 3y + 5 = 0$

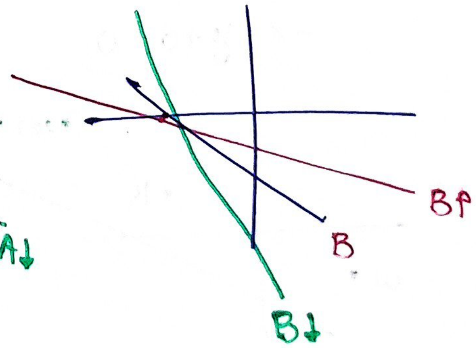
$C_3 \rightarrow 2x + 3y + 0 = 0$



make changes in such a way so that A, B, C change

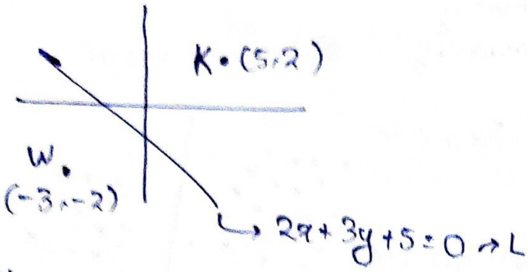






Transformations

$coef_{new} = coef_{old} - \eta \text{ coordinate}$
↳ learning rate



Check for K,

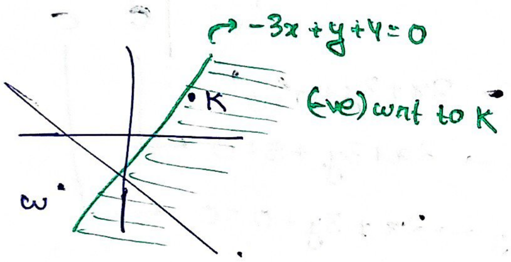
add	5	2	1
(−)	2	3	5
	−3		

1. add 1
2. If check for +ve, region point in do (−)

L =	2	3	5
K = (−)	5	2	1
	$-3\cancel{x}$	1	4

$-3x + y + 4 = 0$

Now our line will be

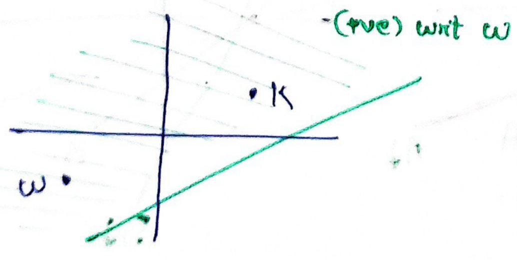


Check for w,

1. add 1.
2. check for point in −ve region do (+) addition

L =	2	3	5
w	−3	−2	1
	−1	1	6

$-x + y + 6 = 0$



Algorithm

	cgpa	iq	placement
x_0	x_1	x_2	y
1	7.6	81	1
1	8	109	1
1	9	110	0

Add a 1 column.

$Ax + By + C = 0$

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

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

$$\begin{bmatrix} W_0 & W_1 & W_2 \end{bmatrix} \begin{bmatrix} X_0 \\ X_1 \\ X_2 \end{bmatrix} = 0$$

$$\sum_{i=0}^{n} W_i * X_i = 0$$

How to check Region.

1.- Take Random n.

$W_0 1 + W_1 8 + W_2 109 > 0 \rightarrow 1$

$< 0 \rightarrow 0$

epoch = 1000 $\eta = 0.01$

for i in range(epoch):

Randomly select a row X_i

if $X_i \in N$ (no placement) and $\sum_{i=0}^{n} W_i X_i \geq 0$ (+ve region)

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

if $X_i \in P$ (yes placement) and $\sum_{i=0}^{n} W_i X_i < 0$ (-ve)

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

Y_i	$\hat{Y}$	$Y_i - \hat{Y}$	$W_{new} = W_{old} + \eta(Y_i - \hat{Y}) X_i$ (1 eqn)
1	1	0	$W_n = W_0$
0	0	0	$W_n = W_0$
1	0	1	$W_n = W_0 + \eta X_i$
0	1	-1	$W_n = W_0 - \eta X_i$

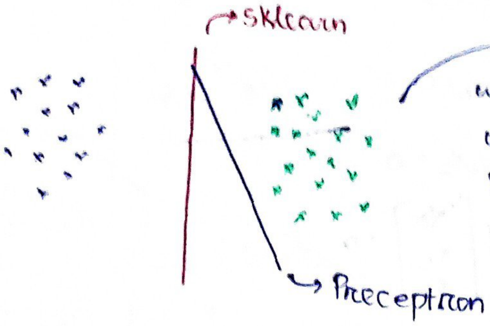
$Ax + By + C = 0$

$y = mx + c$

$m = -\frac{A}{B}$

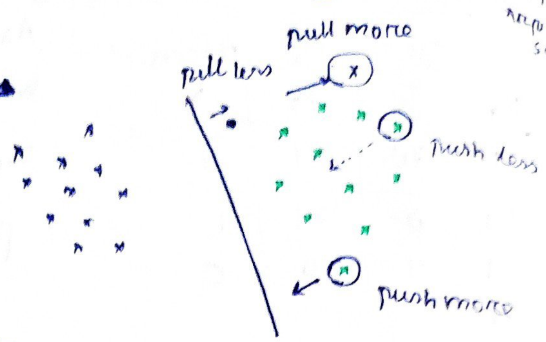
$c = -\frac{C}{B}$

309 pu
Aug 27
sat



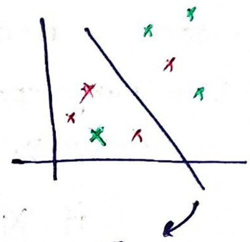
update

why is best?
why?
How?



Ask point
- miss classifed → pull
- correctly → push
- pull/push depends upon mangitude of distance

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$W_n = W_0 + \eta(y_i - \hat{y}_i)x_i$
1	1	0	$W_n = W_0$
0	0	0	$W_n = W_0$
1	0	1	$W_n = W_0 + \eta(y_i - \hat{y}_i)x_i$
0	1	-1	$W_n = W_0 + \eta(y_i - \hat{y}_i)x_i$



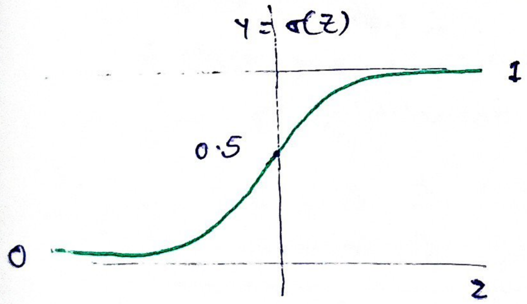
$$\sum_{i=0}^{n} W_i X_i = 0$$

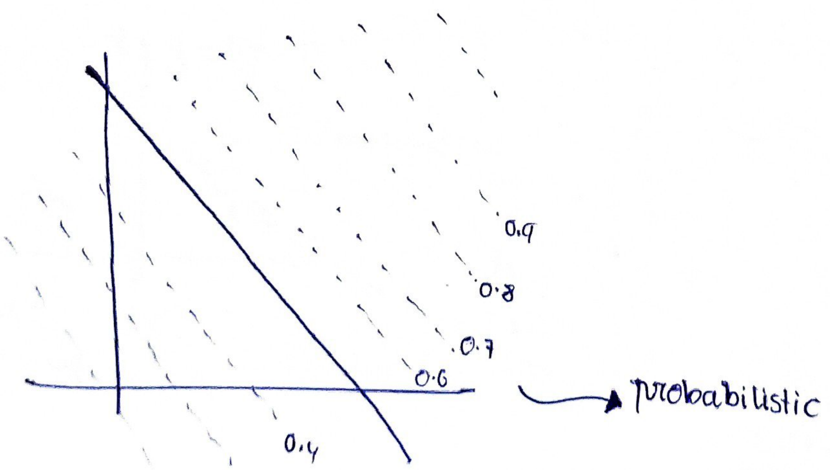
Somehow we need to stop
$(y_i - \hat{y}_i)$ becomes zero

data we can't change

predict model func → step func used either 0 or 1.

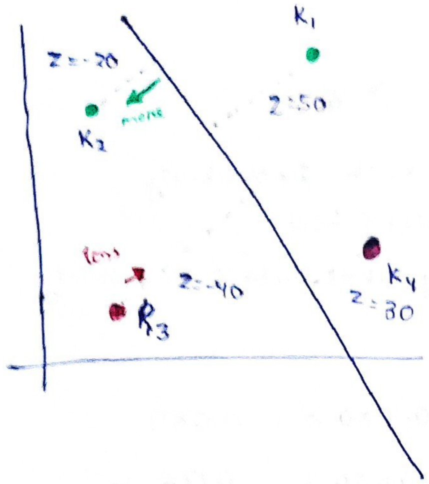
Sigmoid $\quad \sigma(z) = \frac{1}{1-e^{-z}}$





→ probabilistic

Impact



Point	mangitude $z = \sum w_i x_i$	Actual Data y_i	Prediction w/ sigmoid $\hat{y}_i = \sigma(z)$	$y_i - \hat{y}_i$
K_1	50 ~~$\sigma(50)=0.8$~~	1	0.8	0.2
K_2	-20	1	0.3	0.7
K_3	-40	0	0.15	-0.65
K_4	30	0	0.65	-0.15

↳ No placement

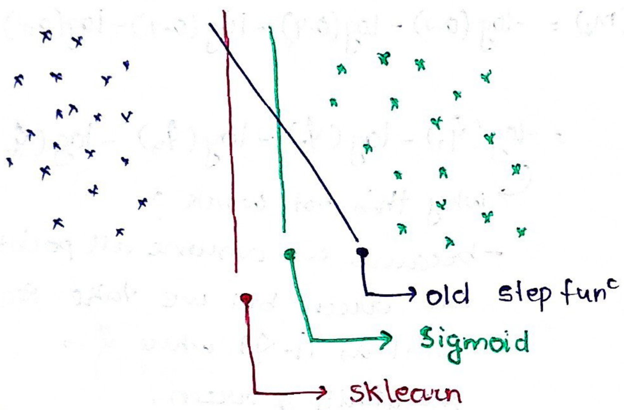
For K_2,

$$w_n = w_0 - \eta \cdot 0.7 \times X_i \rightarrow \text{pull}$$

For K_3,

$$w_n = w_0 + \eta \cdot 0.15 \times X_i \rightarrow \text{push}$$

P4 347 pm



?

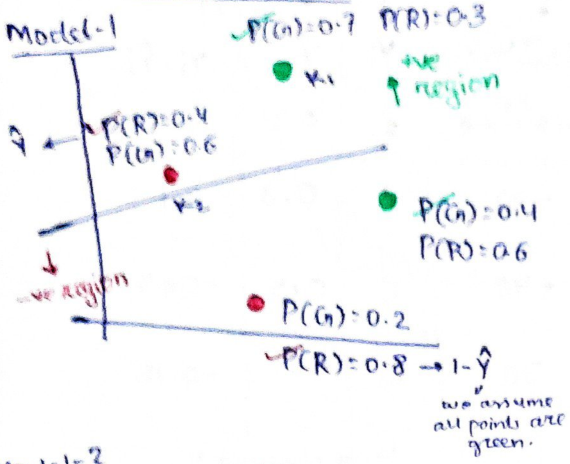
Now where we do mistake?

- if the data will change then sigmoid give another line
- How we/MC knew this is optimal line.

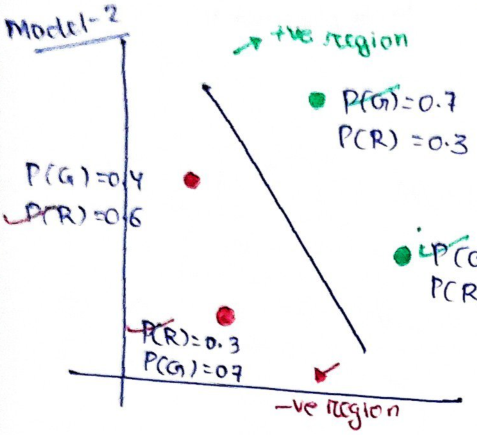
- Yes!
- we need a loss func
- Calculate the loss func to find error then compare

Maximum Likehood

Model-1



Model-2



$\hat{y} = \sigma(z)$

$\hookrightarrow z = \sum W_i x_i$

1. Calculate both probability of Green & Red
2. Take only label data of probability.

for M1

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

$M_2 = 0.7 \times 0.6 \times 0.6 \times 0.3 = 0.176$

best model

But if we multiple then no. become smaller & smaller.

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

CRoss Entropy

- The summation of -ve log of maximum likehood.
- minimize

$\log(m_1) = -\log(0.7) - \log(0.4) - \log(0.4) - \log(0.8)$

$= -\log(\hat{y}_1) - \log(\hat{y}_2) - \log(\hat{y}_3) - \log(\hat{y}_4)$

why this not work?

- because we assume all points are Green. but we take Red point as $(1-\hat{y})$ where $\hat{y}$ is probability of Green.

$= -y_i \log(\hat{y}_i) - (1-y_i)(\log(1-\hat{y}_i)$

for x_1 point

it is green so $\hat{y}_1 = 0.7$ & $y_i = 1$

first row

$= -1.\log(0.7) - \underbrace{(1-1)\log(1-0)}_{0}$

$= -\log(0.7)$

For x_2 point

this is red so $y_2 = 0$ $\hat{y}_R = 1 - \hat{y}_G$

why because red $= 1-0.7$

$= 0.4$

$= \underbrace{-0.\log(0.4)}_{0} - (1-0.7)(\log(1-0.7))$

$= -\log 0.4$

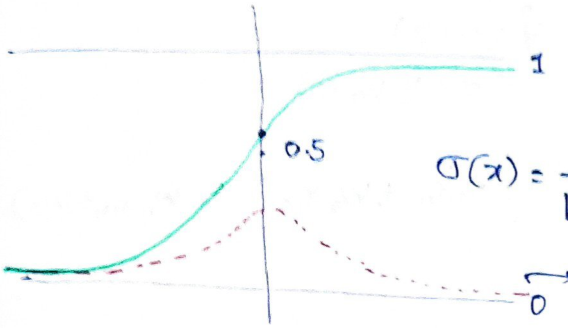
$$J = \sum_{i=1}^{n} -y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)$$

avg error

$$J = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

Binary Cross Entropy

Derivative of Sigmoid



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

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

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

$$= -x^{-2} = -\frac{1}{x^2}$$

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

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

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

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

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

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

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

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

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

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

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

$$= \sigma(x)[1-\sigma(x)]$$

Logistic Regression in Gradient Descent

$$\begin{matrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} & y_1 \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2n} & y_2 \\ \vdots & & & & & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mn} & y_m \end{matrix}$$

$$\hat{y} = \sigma(z)$$
$$= \sigma(\Sigma w_i x_i)$$

$$\hat{y}_1 = \sigma(w_1 x_{11} + w_2 x_{12} \cdots + w_n x_{1n} + w_0)$$

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

$$= \sigma\left(\begin{bmatrix} w_0 + w_1 x_{11} + \cdots w_n x_{1n} \\ w_0 + w_1 x_{21} + \cdots w_n x_{2n} \\ \vdots \\ w_0 + w_1 x_{m1} + \cdots w_n x_{mn} \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}\right)$$

$$\hat{Y} = \sigma(XW)$$

$$J = -\frac{1}{m} \sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)$$

$$= -\frac{1}{m}\left[\Sigma y_i \log(\hat{y}_i) + \Sigma (1-y_i)\log(1-\hat{y}_i)\right]$$

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

$$= -\frac{1}{m}\left[y \log(\sigma(xw)) + (1-y)\log(\sigma(1-xw))\right]$$

$$\sum_{i=1}^{m} y_i \log(\hat{y}_i) = y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 + \cdots\ y_m \log \hat{y}_m$$

$$= \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix} \begin{bmatrix} \log \hat{y}_1 \\ \log \hat{y}_2 \\ \vdots \\ \log \hat{y}_m \end{bmatrix}$$

$$= \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix} \cdot \log \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix}$$

$$= y \log \hat{y}$$

$$= y \log(\sigma(xw))$$

$$\frac{d}{dw}(y \log \hat{y}) = y \frac{d}{dw} \log \hat{y}$$

$$= \frac{y}{\hat{y}} \frac{d}{dw} \hat{y}$$

$$= \frac{y}{\hat{y}} \frac{d}{dw} \sigma(xw)$$

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

$$= \frac{y}{\cancel{\hat{y}}} \cdot \cancel{\hat{y}}[1-\hat{y}] \cdot x$$

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

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

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

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

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

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

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

$$\frac{dJ}{dw} = -\frac{1}{m}\left[Y(1-\hat{Y})X - \hat{Y}(1-Y)X\right]$$

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

$$= -\frac{1}{m}\left[Y - \cancel{Y\hat{Y}} - \hat{Y} + \cancel{Y\hat{Y}}\right]X$$

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

Acc. to gradient descent

$$W_{new} = W_{old} - \eta \frac{\Delta J}{dw}$$

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

$$\hat{Y} = \sigma(XW)$$

$$W = \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_n \end{bmatrix}_{n+1,1} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1n} \\ 1 & X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & & & & \\ 1 & X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}_{(m,n+1)} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{bmatrix}_{(m,1)} \quad \hat{Y} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_m \end{bmatrix}_{(m,1)}$$

<u>Softmax Regression</u> or
Multinomial

P6

<u>Softmax func</u>, no of K = class

$$\sigma(Z)_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}}$$

yes → 1
No → 2
Opt → 3.

$$\sigma(Z_1) = \frac{e^{Z_1}}{e^{Z_1} + e^{Z_2} + e^{Z_3}}$$

$$\sigma(Z)_2 = \frac{e^{Z_2}}{e^{Z_1} + e^{Z_2} + e^{Z_3}}$$

$$\sigma(Z)_3 = \frac{e^{Z_3}}{e^{Z_1} + e^{Z_2} + e^{Z_3}}$$

$$\sigma(Z)_1 + \sigma(Z)_2 + \sigma(Z)_3 = 1$$

cgpa	iq	P
		1
		2
		3

$\xrightarrow{\text{One Hot Encoding}}$

cgpa	iq	P_1	P_2	P_3
-	.	1	0	0
-	.	0	1	0
-	-	0	0	1

↓ Take subset of Data (acc to coeff)

cgpa iq P_1	cgpa iq P_2	cgpa iq P_3
↓	↓	↓
m_1	m_2	m_2
↓ yes	↓ NO	↓ optional
$W_1^{(1)}$ $W_2^{(1)}$ $W_0^{(1)}$	$W_1^{(2)}$ $W_2^{(2)}$ $W_0^{(2)}$	$W_1^{(3)}$ $W_2^{(3)}$ $W_0^{(3)}$

Train model for logsitic Regression

if we give a new data

$S_x = \{7, 70\} = ?$ Y/N/op

↓

$$Z_1 = \sum_{i=1}^{2} x_i w_i = 7 \times W_1^{(1)} + 70 W_2^{(1)} + W_0^{(1)}$$

$$Z_2 = 7W_1^{(2)} + 70W_2^{(2)} + W_0^{(2)}$$

$$Z_3 = 7W_1^{(3)} + 70W_2^{(3)} + W_0^{(3)}$$

Then find the Probability w/ softmax func

$$\sigma(Y) = \frac{e^{z_1}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \sigma(N) = \frac{e^{z_2}}{e^{z_1} + e^{z_2} + e^{z_3}} \qquad \sigma(OP) = \frac{e^{z_3}}{e^{z_1} + e^{z_2} + e^{z_3}}$$

let $= 0.40$ $\quad = 0.35$ $\quad = 0.25$

↓

High so S_x student got placement

Loss function

For Sigmoid

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

For softmax

$$J = -\frac{1}{m}\sum_{i=1}^{m}\sum_{K=1}^{K} Y_K^{(i)} \log(\hat{Y}_K^{(i)})$$

X_1	X_2	Y	$Y_{K=1}$	$Y_{K=2}$	$Y_{K=3}$
X_{11}	X_{12}	1	1	0	0
X_{21}	X_{22}	2	0	1	0
X_{31}	X_{32}	3	0	0	1

$$= Y_1^{(1)}\log \hat{Y}_1^{(1)} + Y_2^{(1)} \log(\hat{Y}_2^{(1)}) + Y_3\log(\hat{Y}_3)^{(1)}$$
$$+ Y_1^{(2)}\log(\hat{Y}_1^{(2)}) + Y_2^{(2)}\log(\hat{Y}_2^{(2)}) + Y_3^{(2)}\log(\hat{Y}_3^{(2)})$$
$$+ Y_1^{(3)}\log(\hat{Y}_1^{(3)}) + Y_2\log(\hat{Y}_2^{(3)}) + Y_3\log(\hat{Y}_3^{(3)})$$

$$= Y_1^{(1)}\log(\hat{Y}_1^{(1)}) + Y_2^{(2)}\log(\hat{Y}_2^{(2)}) + Y_3^{(3)}\log(\hat{Y}_3^{(3)})$$

? ? ?

$$\hat{Y}_1^{(1)} = \sigma(W_1^{(1)}X_{11} + W_2^{(1)}X_{12} + W_0^{(1)})$$
$$\hat{Y}_2^{(2)} = \sigma(W_1^{(2)}X_{21} + W_2^{(2)}X_{21} + W_0^{(2)})$$
$$\hat{Y}_3^{(3)} = \sigma(W_1^{(3)}X_{31} + W_2^{(3)}X_{32} + W_0^{(3)})$$