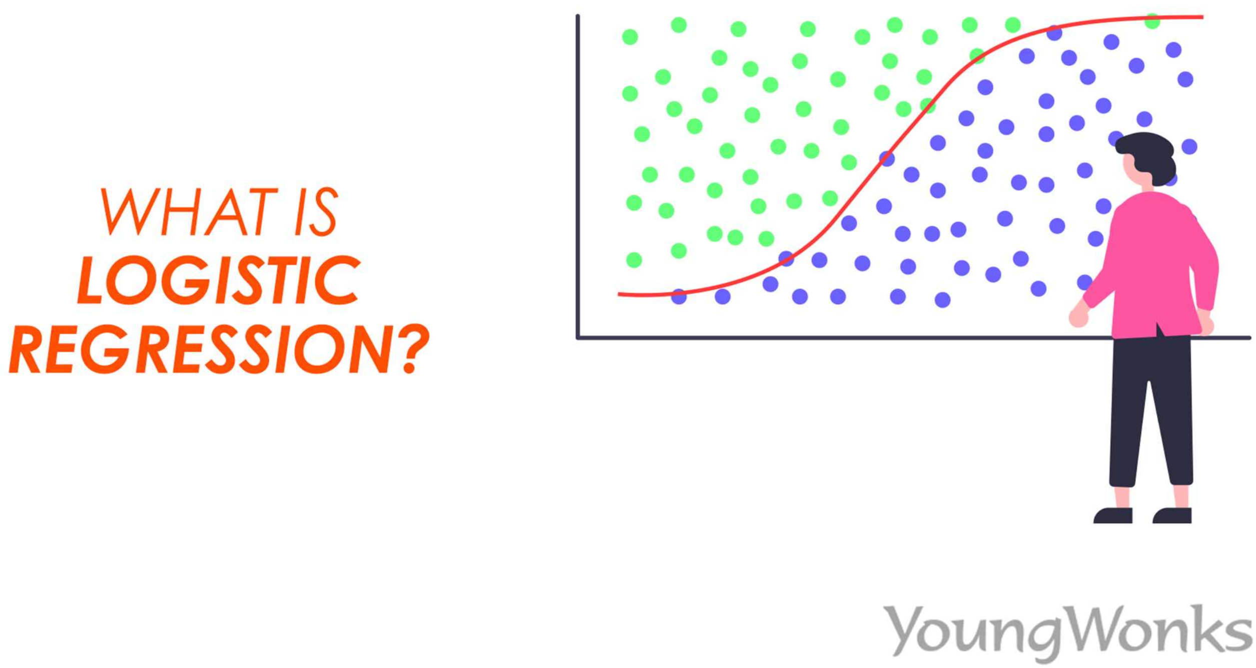
LOGISTIC REGRESSION

In logistic regression, we have a binary output variable Y ,

and we want to model the conditional probability Pr(Y = 1|X = x) as a function of x; any unknown parameters in the function are to be estimated by maximum likelihood

GEOMETRIC INTUITIONS

To apply logistic regression, data should be linearly separable.

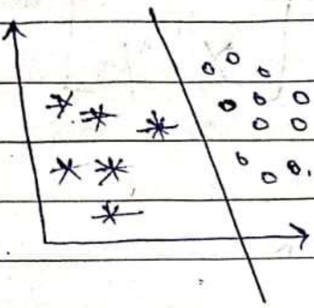


Assume we have the data of students with independent features No. of hours studying and Marks in mock test and we have two labels Pass and Fail.

In logistic regression, we use the general equation of the line rather than the slope-intercept form. Equation for logistic regression for 2D data: (Will look like a line graphically)

$$AX + BY + C = 0$$

* Logistic Regression



eg. for 2D:- $Ax_1 + Bx_2 + C = 0$

3D:- $Ax_1 + Bx_2 + CX_3 + D = 0$

where D is y-intercept

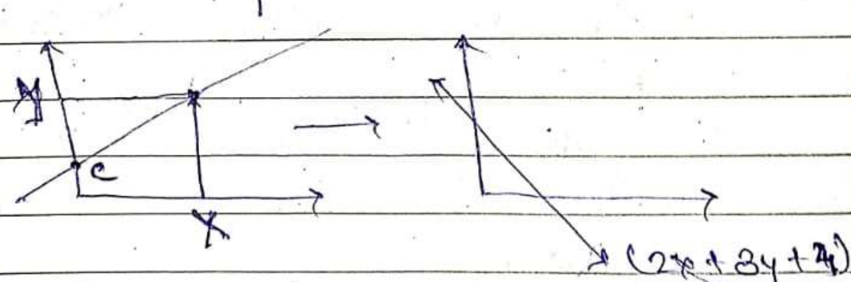
Method I Perceptron Trick

In perceptron trick we started with random value selection & check the prediction is right or worng & according to that we move the line in negative or positive direction.

In logistic Regression the equation of line is :-

$$Ax_1 + By + C = 0.$$

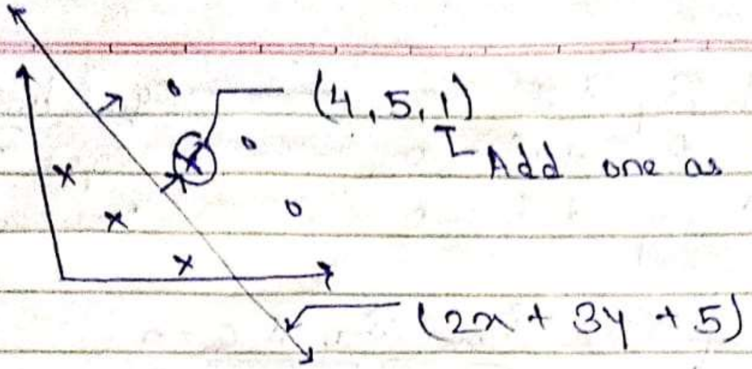
X = x-axis
Y = y-axis
C = y-intercept



The equation is > 0, the point in positive region

equation is < 0, then point in negative region

equation is $= 0$ then point is on line.



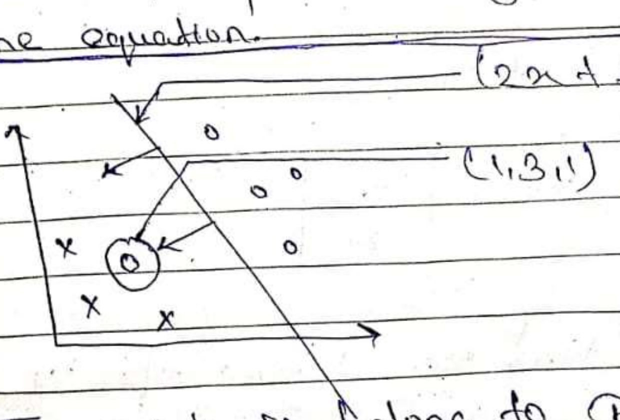
The point is belong to Negative region but the model is predict to positive region then we move line to positive region By Substracting the equation of point with equation of line.

The line move in positive direction.

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

— equation of line
— equation of point
— New equation line

When point belong to Negative region & model predict in positive region then (Substracted)(−) the equation.



The point is belong to positive region but the model predict to negative region the we move line to negative direction by Adding the equation of point with equation of line.

$2x + 3y + 5 = 0$ — eq of line
$+ (1x + 3y + 1) = 0$ — eq of Point
$3x + 6y + 6 = 0$ — New eq of line

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

In logistic regression:-

$$W_0 + W_1X_1 + W_2X_2 = 0$$

where

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

For 3D Data.

$$W_0 + W_1X_1 + W_2X_2 + W_3X_3 = 0$$

W is Weights of line
We need to find value of $(W_0, W_1, W_2 \ldots)$

General formula:-

$$\boxed{\sum W_i X_i}$$

$$\begin{bmatrix} W_0 & W_1 & W_2 \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ X_2 \end{bmatrix}$$

Formula to find New Value of W

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

Where η is learning rate.

Date ______
Page ______

Code-

For in range (epoch):

randomly Select Gradient point

if $X_i \in N$ AND $\sum w_i x_i > 0$

x is actual negative region Model predict in +ve

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

if $X_i \in P$ AND $\sum w_i x_i < 0$

* Common Formula for both codition

e.g Placement DataSet:-

1 - Student Job placement done

0 - Student Job placement is not done

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$
1	1	0
0	0	0
0	1	-1
1	0	1

* For i in range (epoch):

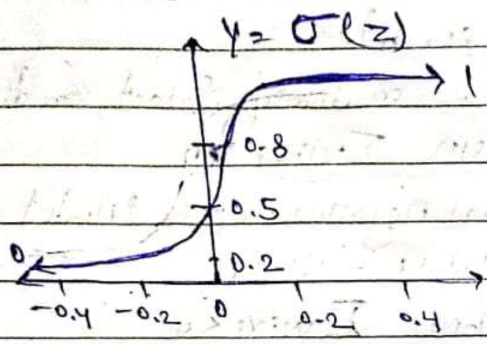
Select random Student

$W_{new} = W_{old} + [\eta (y_i - \hat{y}_i) x_i]$

* This method is not ~~[illegible]~~ completed So we used Sigmoid Function

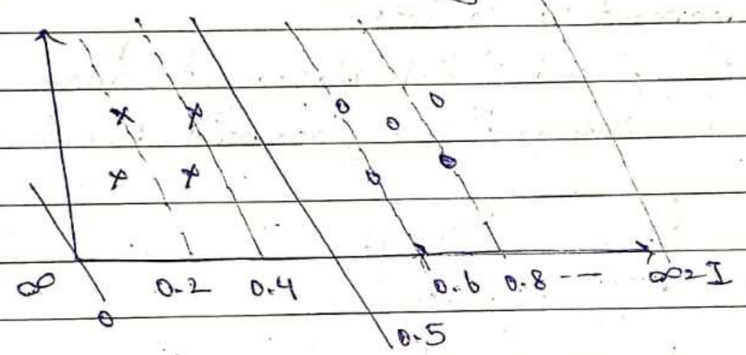
* Sigmoid Function

* Sigmoid Function



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

The Sigmoid function follow the probability function between range (0 to 1)

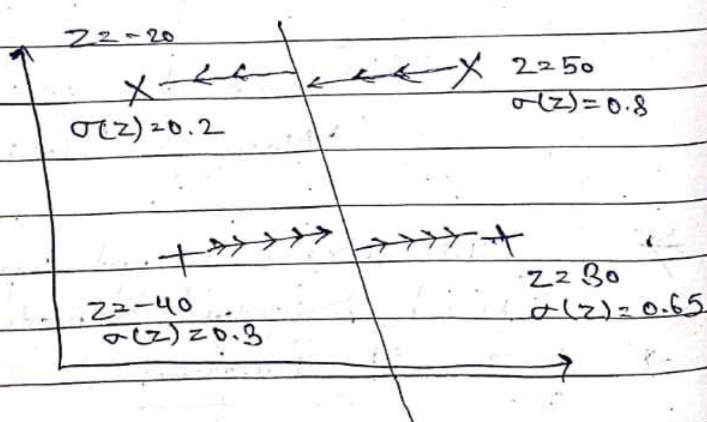


$$W_{new} = W_{old} + \eta (y_i - \hat{y}_i) x_i$$

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

where $z = \sum w_i x_i$

y_i	$\hat{y}$	$y_i - \hat{y}_i$
1	0.8	0.2
0	0.65	-0.65
1	0.3	0.7
0	0.2	-0.2



+ → Negative point

x → positive point

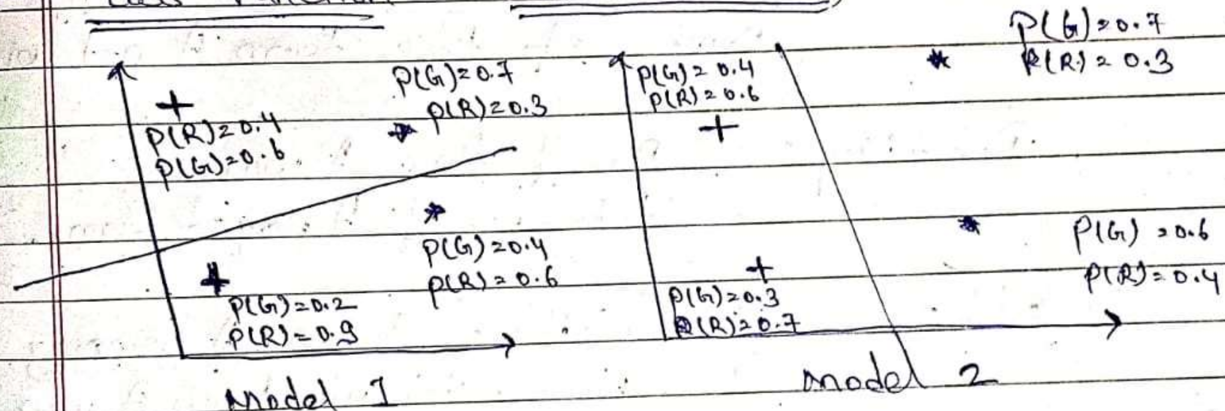
$$W_n = W_o = \eta(y_i - \hat{y})x_i$$

$$W_n = W_o = \eta(y_i - \sigma(z))x_i$$

$\boxed{W_n = W_o + \eta \times 0.2 \times x_i}$ → Push line

$\boxed{W_n = W_o - \eta \times -0.2 \cdot x_i'}$ → pull line.

Loss function - (error function)



The logistic Regression used Maximum likelihood

* Maximum likelihood

In Maximum likelihood is used to know Best-fit line in model we multiply the all probability value of all predicted point.

Note:- We add only the probability of the points is actual is for example.

Red:- point:- Ki Red hone ki probability kitni hai.
Green point ki Green hone ki probability kitni hai.

Model 1 = 0.7 × 0.4 × 0.4 × 0.8 = 0.089
Model 2 = 0.7 × 0.6 × 0.6 × 0.7 = 0.176

Model 2 is Best

$P(y=1 \mid x=x)$

$y = w_0 + w_1x_1 + w_2x_2 \cdots w_nx_n$

$y = w_0 \; w^T x$

$$y = w_0 + [w_1 \; w_2 \cdots w_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\text{odds} = \frac{\text{number of time team A win}}{\text{number of time team A not win}}$$

$$\text{Probability} = \frac{\text{no. of time A win}}{\text{total. no. of A played Game}}$$

$$\text{odds} = \frac{\text{Probability of team A winning}}{1 - \text{probability of team A win}}$$

$$\log\left(\frac{P(x)}{1-P(x)}\right) = w_0 + w_1x$$

Taking exponent on both side.

$$\frac{P(x)}{1-P(x)} = e^{w_0 + w_nx}$$

$$P(x) = e^{w_0 + w_nx}(1 - P(x))$$

multiply across

Date
Page

$$p(x) = e^{w_0 + wx} - e^{w_0 + wx}(p(x))$$

$$p(x) + e^{w_0 + wx}(p(x)) = e^{w_0 + wx}$$

common $p(x)$

$$p(x)(1 + e^{w_0 + wx}) = e^{w_0 + wx}$$

$$p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

$$p(x) = \frac{e^y}{1 + e^y}$$

Divide num/Dom by e^y

$$p(x) = \frac{\frac{e^y}{e^y}}{\frac{1 + e^y}{e^y}}$$

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

$$\text{loss} = P(x_i) * [1 - p(x)]$$

loss function (β) : $\prod p(x_i) \prod (1 - p(x_i))$

But the Problem is when we multiple decimal value (0.001) it's become very small because multiple function.

That why we take log()

As we ~~know~~ know log of value is between 0 to 1 is always negative (-ve)

That we used Cross Entropy

The ~~Su~~ Σ of (-ve) negative log of maximum likelyhood is called Cross entropy

$$\log(\max) = -\log(x) - \log(x) - - -$$

In Cross entropy we check the minimum value

$$\log(0.1) > \log(0.9)$$
$$-(-1) > -(-0.4)$$
$$1 > 0.04$$

Formula:-

$$\log(y) = -y_i \log(\hat{y}) - (1-\hat{y}) \log(1-\hat{y})$$
$$\hat{y} = P(x) = \sigma(x)$$

$$L(B) = -y_i \log(P(x)) - (1-\hat{y}) \log(P(x))$$

To finde the minimum lose we used Gradient decsent

Sigmoid function — derivative

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

$$\sigma(z) = \frac{d}{dx}\left(\frac{1}{1-e^{-z}}\right) \qquad x^{-2} = \frac{-1}{x^2}$$

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

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

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

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

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

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

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

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

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

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

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

Date / / .

Logistic Regression by Gradient Descent

Rows = m   Columns = n

1	2	3	... n	y	$\hat{y}$
x_{11}	x_{12}	x_{13}	x_{1n}	y_1	$\hat{y}_1$
x_{21}	x_{21}	x_{23}	x_{2n}	y_2	$\hat{y}_2$
x_{m1}	x_{m2}	x_{m3}	x_{mn}	y_m	$\hat{y}_m$

To calculated $\hat{y}$

$$\sigma(w_1x_{11} + w_2x_{12} + w_3x_{13} + \ldots w_nx_{1n} + w_0) = \hat{y}_1$$

$$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \sigma(w_0 + w_1x_{12} + w_2x_{12} - - - + w_nx_{1n}) \\ \sigma(w_0 + w_1x_{21} + w_2x_{22} - - - + w_nx_{2n}) \\ \vdots \\ \sigma(w_0 + w_1x_{nm} + w_2x_{nm} - - - + w_nx_{nm}) \end{bmatrix}$$

$$\hat{y} = \sigma \begin{bmatrix} (w_0 + w_1x_{12} + w_2x_{12} - - - + w_nx_{1n}) \\ (w_0 + w_1x_{21} + w_2x_{22} - - - + w_nx_{2n}) \\ (w_0 + w_1x_{nm} + w_2x_{nm} - - - + w_nx_{nm}) \end{bmatrix}$$

$$\hat{y} = \sigma\left(\begin{bmatrix} 1 & x_{11} & x_{12} & - & x_{1n} \\ 1 & x_{21} & x_{22} & - - & x_{2n} \\ 1 & x_{nm} & x_{n2} & - - & x_{nm} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix} \right)$$

$$\boxed{\hat{y} = \sigma(XW)}$$

Write loss Function in metric from

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

Date / /

$$L = \frac{-1}{n}\left[\sum_{i=1}^{n} y_i \log(\hat{y}_i) + \sum_{i=1}^{n} (1-y_i)\log(1-\hat{y}_i)\right]$$

$$\sum_{i=1}^{n} y_i \log(\hat{y}_i) = y_1 \log \hat{y}_1 + y_2 \log \hat{y}_2 \cdots\cdots y_n \log \hat{y}_n$$

$$= \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_m \end{bmatrix} \begin{bmatrix} \log \hat{y}_1 \\ \log \hat{y}_2 \\ \vdots \\ \log \hat{y}_m \end{bmatrix}$$

$$= \begin{bmatrix} y_1 & y_2 & y_3 & \cdots & y_n \end{bmatrix} \log\left(\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} \right)$$

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

If in Second term.

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

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

We apply Gradient descent to find the value of W where our loss function is minimum.

loss function matrix :-

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

Gradient descent :- We initialize the W with random value.

for in epochs:

$$W_n = W_0 - \eta \frac{\Delta L}{\Delta W}$$

$$\frac{\Delta L}{\Delta w} = \left[\frac{dL}{dw_0}, \frac{dL}{dw_1} \frac{dL}{dw_2} - - \frac{dL}{dw_n} \right]$$

loosse Function

$$L = \frac{1}{n} \left[\underbrace{y(\log \hat{y})}_{A} + \underbrace{(1-y)\log(1-\hat{y})}_{B} \right]$$

Derivated the L(w) of A

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

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

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

As we know derivated of $\sigma(x)$ is $\sigma(x)(1-\sigma(x))$

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

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

$$= y(1-\hat{y})x$$ Differension of A term

Now term B

Now term B

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

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

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

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

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

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

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

~~Deveri~~ Derivative of term B

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

Now we consider both condition.

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

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

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

$$\boxed{\frac{\Delta L}{\Delta w} = \frac{-1}{m}(y-\hat{y})x}$$

According to Gradient descent

$$\boxed{w_n = w_0 + \eta \frac{1}{m}(y-\hat{y})X}$$

$$W = \underset{((n+1),1)}{\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}} \quad X = \underset{(m,n+1)}{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & & \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}} \quad Y = \underset{(m,1)}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}} \quad \hat{Y} = \underset{(m,1)}{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_m \end{bmatrix}}$$

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

$$W_{(n+1,1)} = W_{(n+1,1)} + \frac{\eta}{m}(y-\hat{y})_{(1,m)}\,X_{(m,n+1)}$$

(1, n+1)

Result = (n+1, 1)