Logistic Regression is a statistical and machine learning method used for classification.
Instead of predicting a continuous value (like in linear regression), it predicts the probability that an observation belongs to a particular class (e.g., Yes/No, 0/1, Pass/Fail).

 Key Idea:
We want to predict an outcome that is binary (0 or 1).
If we used linear regression, we could get values outside the range [0, 1], which doesn't make sense for probabilities.
Logistic regression fixes this by applying the logistic (sigmoid) function to map predictions into the range [0, 1].

 Real-Life Examples:

Email spam detection → Spam (1) or Not Spam (0)

Credit scoring → Default (1) or Not Default (0)

Medical diagnosis → Has disease (1) or Healthy (0)

Customer churn → Will leave (1) or Stay (0)

 Extensions:

Multinomial Logistic Regression → Used for multi-class classification (e.g., predicting if a fruit is apple, orange, or banana).

Regularized Logistic Regression → Ridge (L2), Lasso (L1) to avoid overfitting.

Different Equations of line

1 Slope Intercept Form

$$y = mx + b$$

Usefulness: This form is quickly identifying the slope and intercept of line.

2 Point Slope Form

$$y - y_1 = m(x - x_1)$$

Usefulness - This form is helpful when you know a point on the line and its slope.

3 Standard Form

$$Ax + By + C = 0$$

Usefulness - This form is useful for comparing different lines and finding intercepts.

4 Intercept Form

$$\frac{x}{a} + \frac{y}{b} = 1$$

a is the x intercept
b is the y intercept

5 Vertical Line

$$x = a$$

Vertical line with a as the x intercept.

6 Horizontal Line

$$y = b$$

Horizontal Line with b as the y intercept.

Logistic Regression.

We are going to use standard line equation

$$ax + by + c = 0$$

Eq. in term of θ

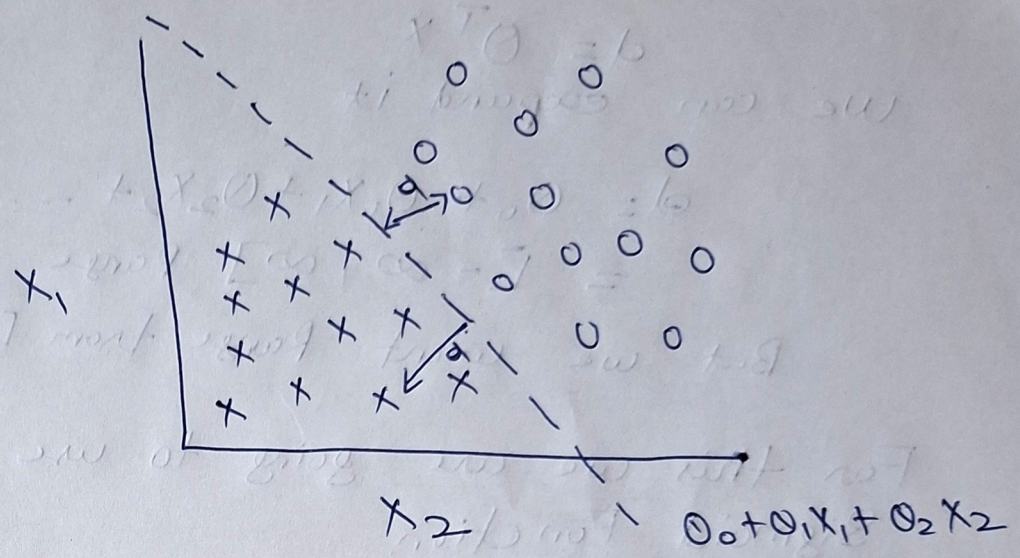
$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

We want a line which is best for seprating data points.

If we have two features then

$$\theta = [\theta_0, \theta_1, \theta_2]$$

i.e $(n+1)$ if we have n features



Distance of a point from a line $d_1 = \dfrac{\theta_0 + \theta_1 X_1 + \theta_2 X_2}{\sqrt{\theta_1^2 + \theta_2^2}}$

$$d_2 = \frac{\theta_0 + \theta_1 X_1' + \theta_2 X_2'}{\sqrt{\theta_1^2 + \theta_2^2}}$$

In both d_1 and d_2 denominator remain same only X_1 and X_2 changes

So

$$d \propto \theta_0 + \theta_1 X_1 \theta_2 X_2$$

d is +ve for one side of line

d is -ve for other side of line.

$$d = \theta^T X$$

We can expand it

$$d = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n$$

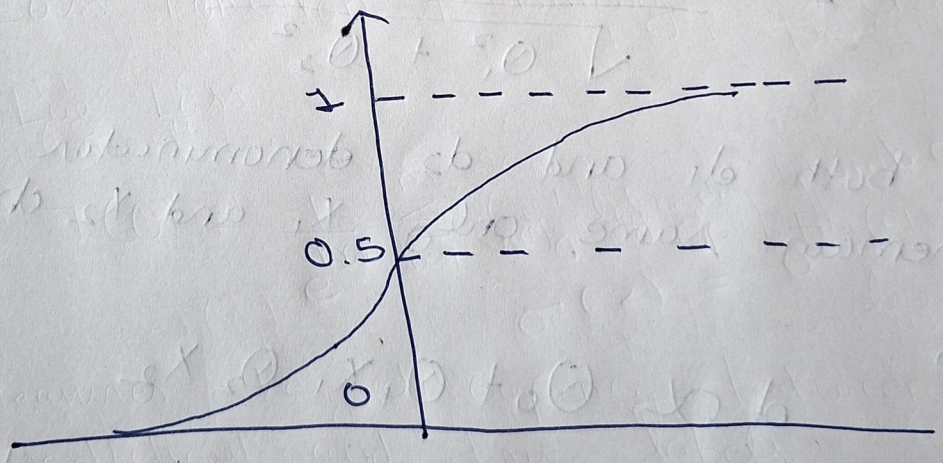
$$= [-\infty, \infty] \text{ Range}$$

But we want Range from $[0, 1]$

For this we are going to use Sigmoid Function.

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

Lets see the graph of sigmoid with different functions.



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

$e^{-\infty}$ is very small approching to zero

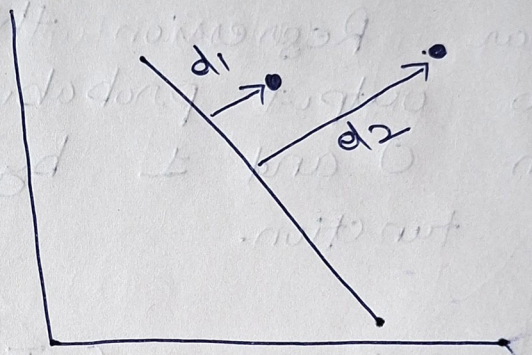
so $\sigma(z) = 1$ when $z = \infty$

When $z = 0$ $\quad \sigma(z) = \frac{1}{1+e^0} = \frac{1}{2} = 0.5$

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

e^{∞} is very large number so

$$\sigma(z) = 0$$



$\sigma(d_2) = 1$

It means farthest is the point from the line we get more confidence in predicting Y.

Smallest the distance less is the confidence

Hypothesis for Logistic Regression

$$y_p = h_\theta(x) = \sigma(\theta^T X)$$

where $X_0 = 1$

$$= \frac{1}{1 + e^{-\theta^T x}}$$

where $\theta^T x = \sum_{i=0}^{n} \theta_i x_i$

Logit Model

The Logit Model is modification of linear Regression that make sure to output probability between 0 and 1 by applying sigmoid function.

$$h_\theta(x) = \sigma(\theta^T x)$$

Loss Function - Binary CrossEntropy Log Loss

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

Measures the discrepency between predictions and true labels.

Case I tve Samples

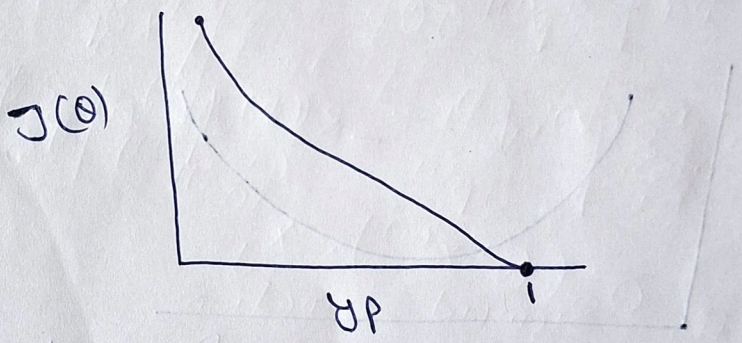
If $y^{(i)} = 1$

second term becomes zero

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log y_p^{(i)}$$

If $y_p = 1$ $\log 1 = 0$

If $y_p = 0$ $\log 0 = -\infty$



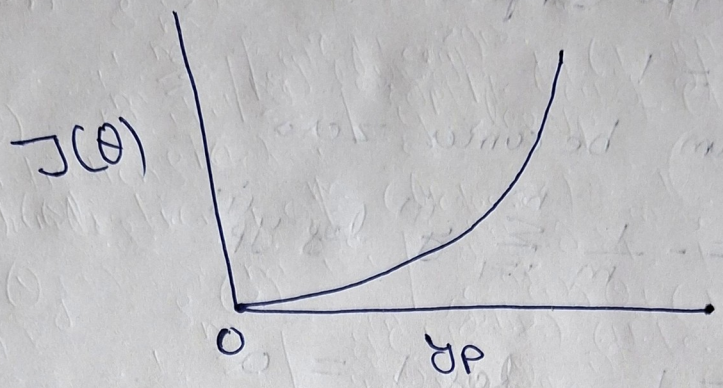
Case II -ve samples

$y_i = 0$ first term becomes zero

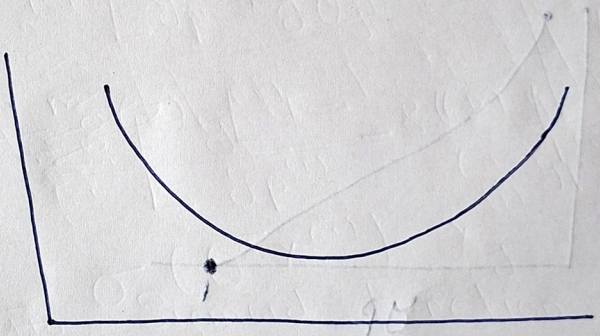
$$J(\theta) = -\frac{1}{m} (1-y^{(i)}) \log (1-y_p^{(i)})$$

If $y_p = 0$ $1 \cdot \log 1 = 0$

If $y_p = 1$ $1 \cdot \log 0 = +\infty$



If we combine two functions and thier graph.



We can use Gradient Descent and it is a convex function (function in which we have one global minima and maxima).

Differentiation of Log Loss Function

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} \log y_p^{(i)} + (1-y^{(i)}) \log (1-y_p^{(i)})\right]$$

Differentiation of $\log x = \frac{1}{x}$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{y_p^{(i)}}\frac{\partial y_p^{(i)}}{\partial \theta_j} + \frac{(1-y^{(i)})(0-\partial y_p^{(i)})}{(1-y_p^{(i)})\partial \theta_j}\right]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{y_p^{(i)}}\frac{\partial y_p^{(i)}}{\partial \theta_j} + \frac{(1-y^{(i)})(-\partial y_p^{(i)})}{(1-y_p^{(i)})\partial \theta_j}\right]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{y_p^{(i)}} - \frac{(1-y^{(i)})}{(1-y_p^{(i)})}\right]\frac{\partial y_p^{(i)}}{\partial \theta_j} \quad (1)$$

Hypothesis Sigmoid Function

$$y_p^{(i)} = h_\theta(x^{(i)})$$
$$= \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n)$$
$$= \sigma(\theta^T x^{(i)})$$

Differentiation of Sigmoid Function

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

$$g'(z) = \frac{\partial g(z)}{\partial z}$$

Differentiation of $\frac{1}{x} = \frac{1}{x^2}$ and $e^x = 1 \cdot e^x$

$$g'(z) = \frac{1}{(1+e^{-z})^2} e^{-z} \times 1$$

$$g'(z) = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$g'(z) = \frac{e^{-z}}{(1+e^{-z})(1+e^{-z})}$$

We can also write it as

$$g'(z) = \frac{1}{1+e^{-z}} \left[1 - \frac{1}{1+e^{-z}} \right]$$

$$\boxed{g'(z) = g(z)\,[1 - g(z)]} \quad -(2)$$

$$\frac{\partial y_p^i}{\partial \theta_j} = \sigma(\theta^T x^{(i)}) \cdot x_j^{(i)}$$

$$y_p^i = \theta^T x^{(i)}$$

$$\boxed{\frac{\partial y_p^i}{\partial \theta_j} = y_p^{(i)} (1 - y_p^{(i)})\, x_j} \quad -(3)$$

Using Eq (1)

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)} - y^{(i)}y_p^{(i)} - y_p^{(i)} + y^{(i)}y_p^{(i)}}{y_p^{(i)}(1-y_p^{(i)})}\right]\frac{\partial y_p^{(i)}}{\partial \theta_j}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)} - y_p^{(i)}}{y_p^{(i)}(1-y_p^{(i)})}\right] y_p^{(i)}(1-y_p^{(i)})x_j$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} - h_\theta(x^{(i)})\right] x_j$$

$$\boxed{\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{i} - y_p^{i}\right] x_j}$$