* Logistic Regression

- Supervised algorithm
- It is a regression model which tries to predict whether given data point belong to category '1' or '0' using binomial function.

A. Why not use Linear Regression to classify?

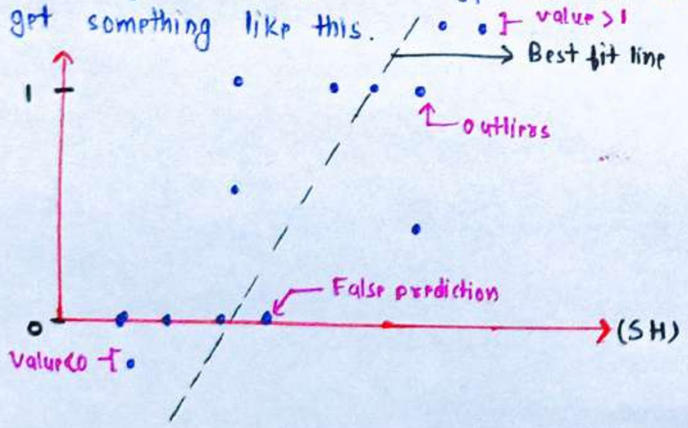
① Linear regression deals with continous value while logistic regression deals with discrete values.

② If we try to classify using threshold on contnous value, it fails to do it properly when new value or outlier added as threshold shifts.

③ As it is binary classification value should either be 1 or 0 but as data is continous it doesnt happen.

Ex-. Say if we want to classify if student pass or fail with criteria: if study hr (SH) > 4 it is pass else fail [1 = pass, 0 = fail]

- When we plot this data and try to get best fit line using linear regression hypothesis line we get something like this.



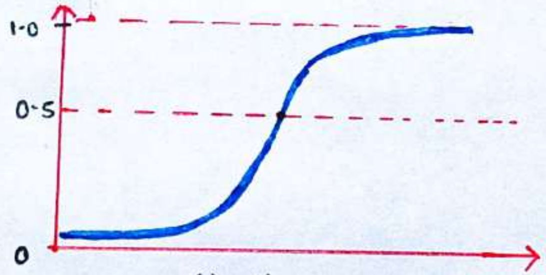
- We can observe that introducing outliers will lead to even worse regression line.
- Also few values even dont belong to the binar classes 1 or 0. (value>1, value<0)
- High chances of false prediction.
- As being type of regression logistic also uses the same hypothesis function as linear regression i.e,

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

B Using Sigmoid function with hypothesis function.

- Applying Sigmoid function (g) over hypothesis function ($h_\theta(x)$) helps to resolve the issue of values going over 1 or going less than 0 not belonging to any classes by limiting the regression line between 1 and 0.

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



Sigmoid function(g)

$\therefore$ $z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

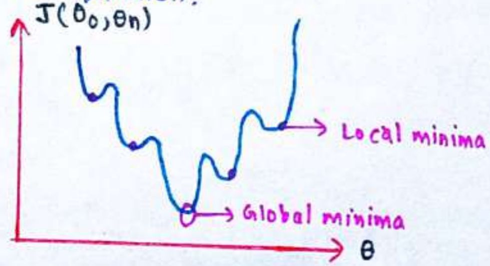
$h_\theta(x) = g(z)$

$$\Rightarrow \boxed{h_\theta(x) = \frac{1}{1+e^{-(\theta_0 + \theta_1 x_1 + \dots \theta_n x_n)}}}$$

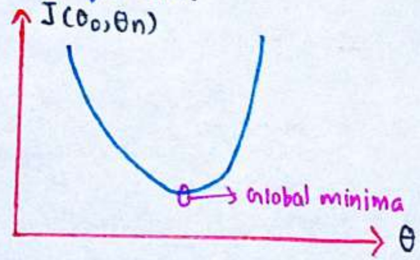
- But Sigmoid function is not a convex function. So when we try to minimize our cost function to reach global minima it fails to do so as of presence of local minima.
- If we use L2 Loss function, and update it with sigmoid applied hypothesis function we get,

$$\boxed{J(\theta_0, \theta_n) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{1+e^{-(\theta_0 + \theta_1 x_1 + \dots \theta_n x_n)}} - y^i \right]^2}$$

and this cost function yields an non convex function.



- But we need a convex function like this one to reach the global minima and have best fit line.



- To achieve this we update our cost function first to problistic function then to log probability function.

c. Upgrading cost function to log likelihood function

- To get convex function in order to reach global minima, first we convert sigmoid function to probalistic function.

- Our cost function with sigmoid applied hypothesis function in probability form is,

$$\begin{cases} P(y_i = 1 \mid x_i : \theta) = h_\theta(x_i) \\ P(y_i = 0 \mid x_i : \theta) = 1 - h_\theta(x_i) \end{cases}$$

→ combining this we get shorter form as,

$$\boxed{P(y_i \mid x_i : \theta) = [h_\theta(x_i)]^{y_i} [1 - h_\theta(x_i)]^{1-y_i}}$$

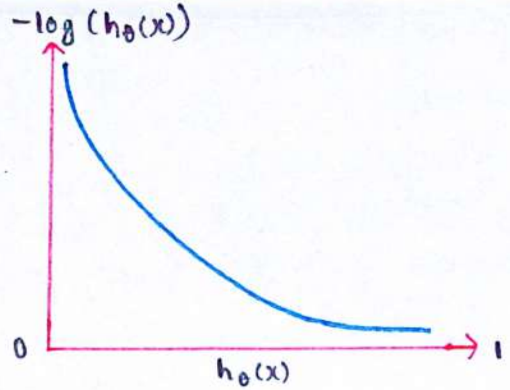
- Now we will further introduce log likelihood on $P(y_i \mid x_i : \theta)$ to get,

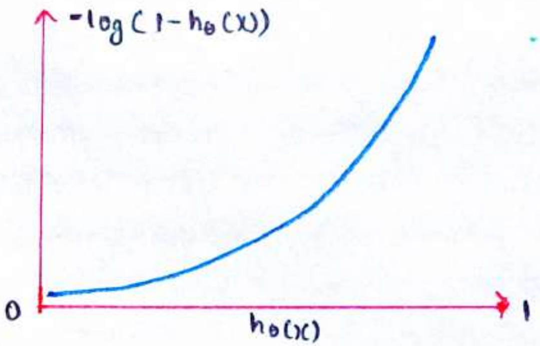
$$\Rightarrow \boxed{L(\theta) = -y \log(h_\theta(x_i)) - (1-y_i) \log(1-h_\theta(x_i))}$$

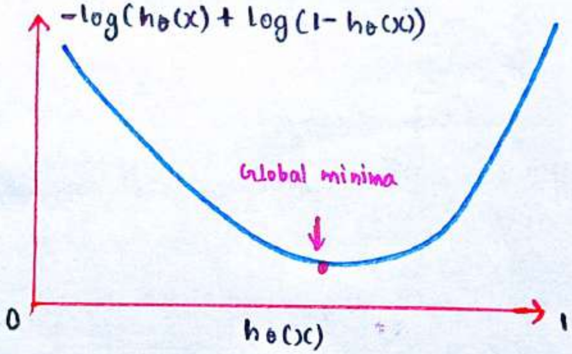
if we put $y=1$ or $y=0$ we can sumarize $L(\theta)$ as,

$$L(\theta) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1, \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

- If we plot both of these cases we will get something like this :







- After combining both we can observe we were able to get a convex function.

- Now substituting this log likelihood hypothesis function ($h_\theta(x)$) in cost function ($J(\theta_0, \theta_n)$) we get our final cost function.

$$\boxed{J(\theta_0,\theta_n) = \frac{1}{n}\left[y^i \log(h_\theta(x^i)) + (1-y^i)\log(1-h_\theta(x^i)) \right]^2}$$

- Now we will apply convergence algorithm on cost function to get the best fit line.

$$\boxed{\theta_n = \theta_n - \alpha \frac{\partial}{\partial \theta_n}(J(\theta_0,\theta_n))}$$

$$\boxed{\theta_0 = \theta_0 - \alpha \frac{\partial}{\partial \theta_0}(J(\theta_0,\theta_n))}$$