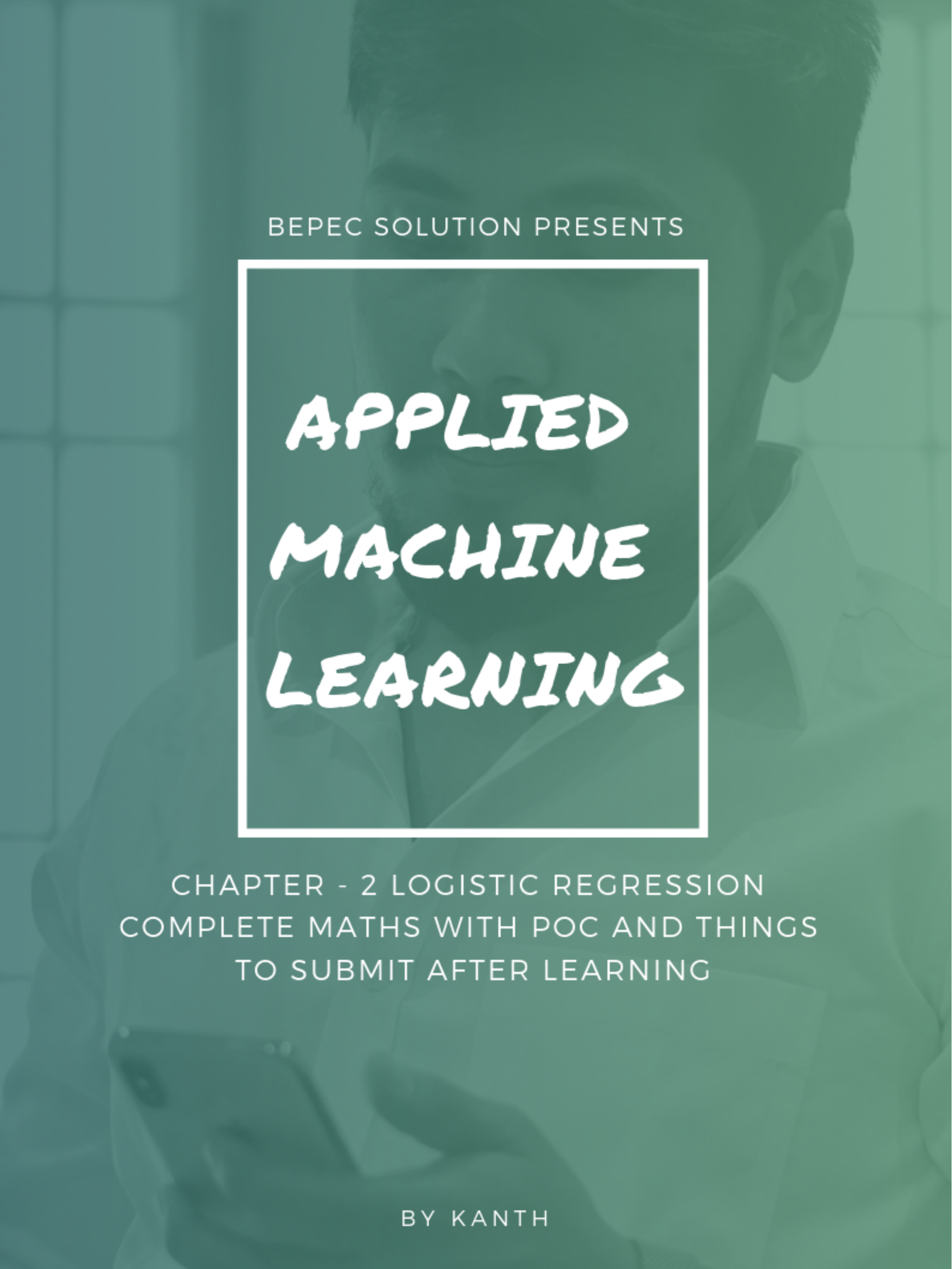


BEPEC SOLUTION PRESENTS

APPLIED MACHINE LEARNING

CHAPTER - 2 LOGISTIC REGRESSION
COMPLETE MATHS WITH POC AND THINGS
TO SUBMIT AFTER LEARNING

BY KANTH

1. Introduction to Logistic Regression
2. Why Sigmoid curve or S-Curve?
3. Why to transform intercepts to non -linear format
4. Sigmoid Equation and Probability Equation
5. Different Families of Logistic Regression
6. Assumptions of Logistic Regression
7. Project on Logistic Regression

LOGISTIC REGRESSION

Regression Equation y = mx+c

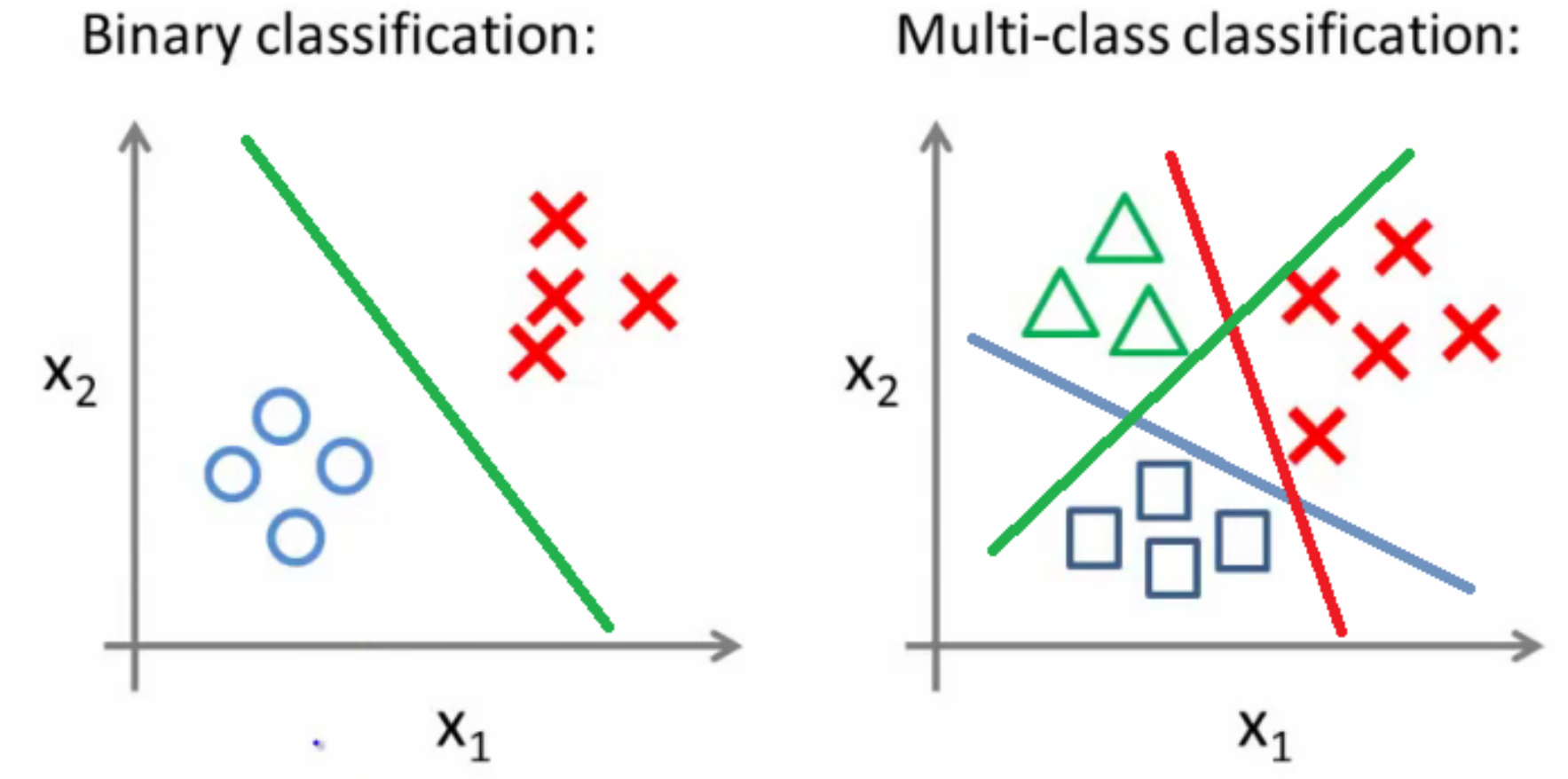
When y is continuous we go with linear regression

When y is discrete we go with logistic regression

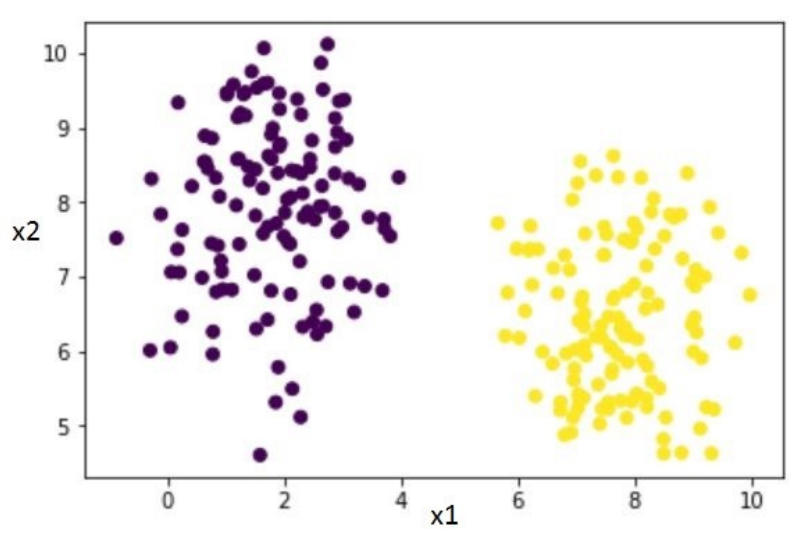
When logistic is doing classification, but still why we call it as regression behalf of classification?

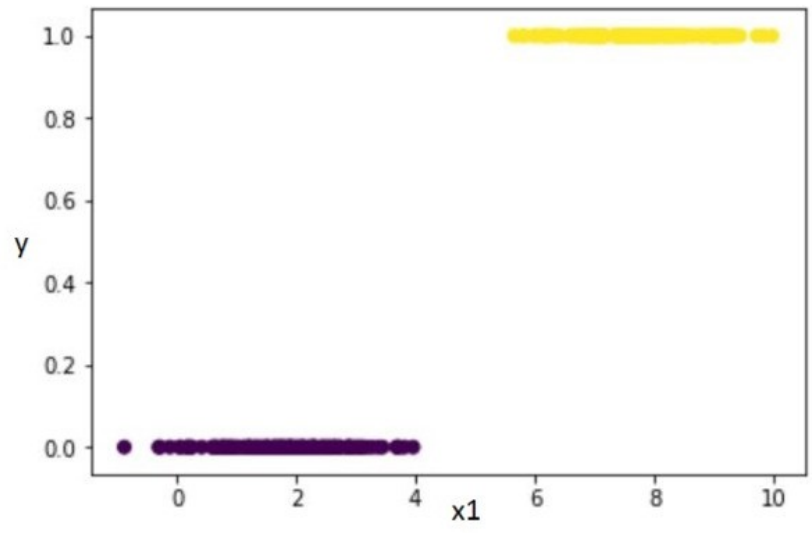
Thinking???

You may be wondering why the name says regression if it is a classification algorithm, well,It uses the regression inside to be the classification algorithm.
Classification : Separates the data from one to another.

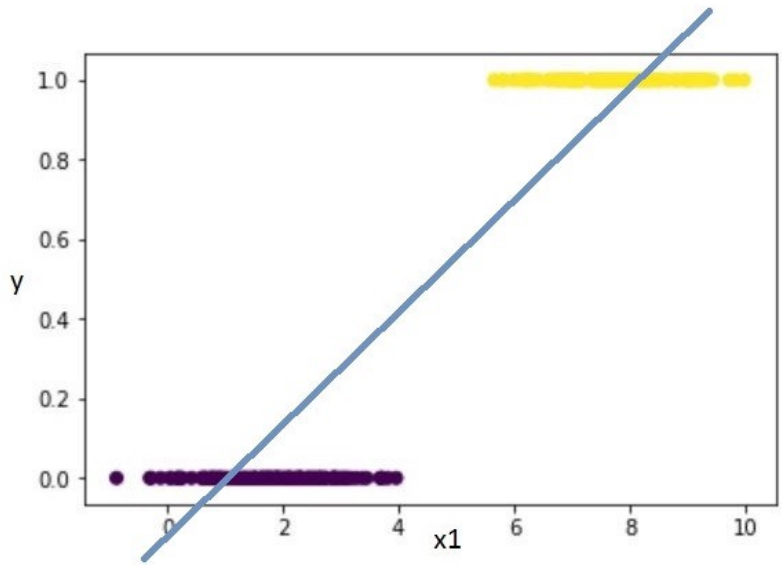


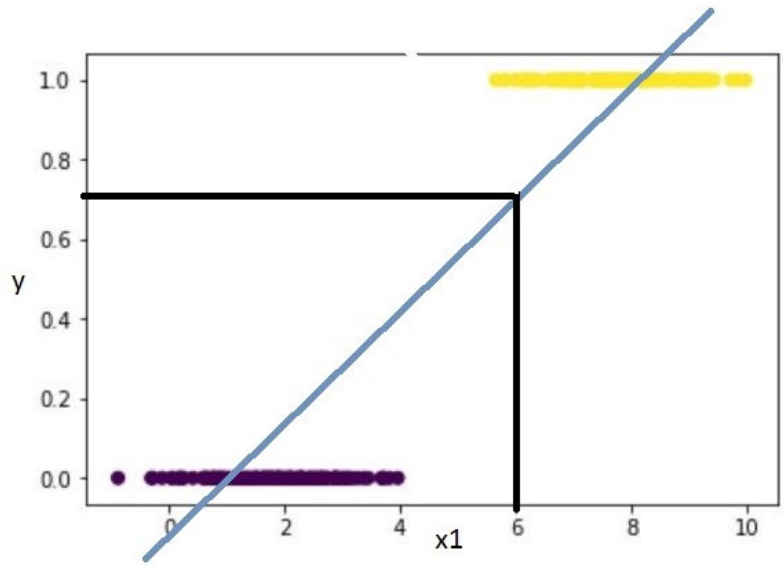
This story we talk about **binary classification** (0 or 1) Here target variable is either 0 or 1
Goal is to find that green straight line (which separates the data at best)
so we use regression for drawing the line , makes sense right?
Lets take a random dataset and see how it works,



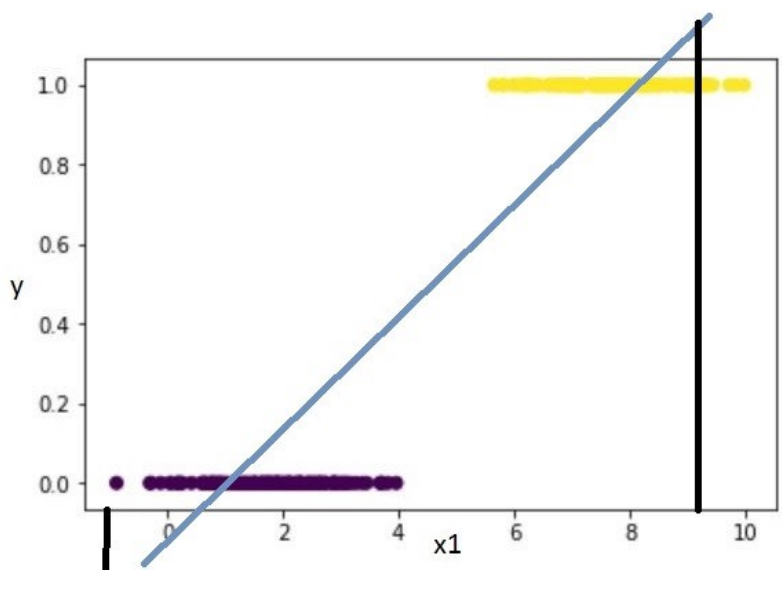


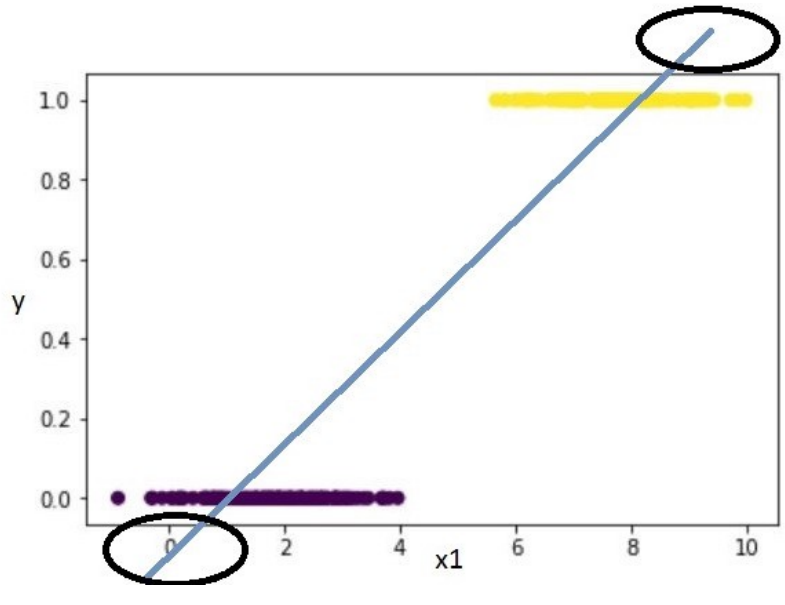
if we observe the *right* picture we have our independent variable (X) and dependent variable(y) so this is the graph we should consider for the classification problem
Given X or (Set of x values) we need to predict whether it's 0 or 1 (Yes/No).
If we apply Linear regression for above data we get something like this,





Given X value 6 we can say y is 0.7 (close to 1), that's cool but wait, What if I give negative X value or greater X value??? The output is this

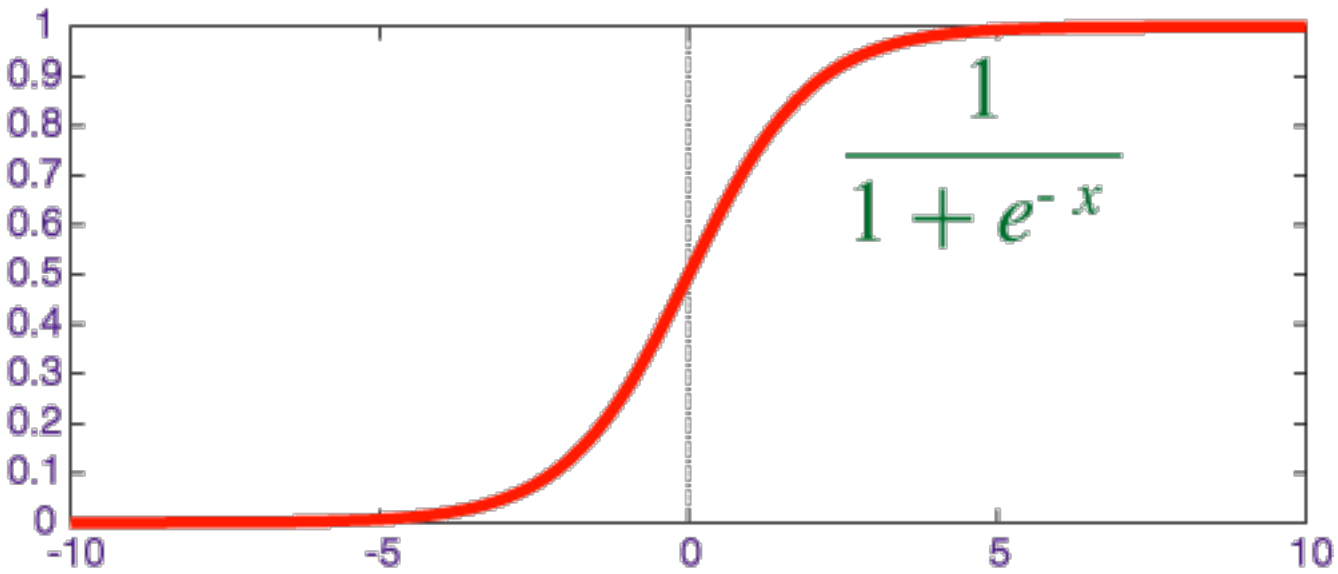




We only accept the values between 0 and 1 (We don't accept other values) to make a decision (Yes/ No)

so how do we proceed further?

There is an awesome function called **Sigmoid** or **Logistic function** , we use to get the values between 0 and 1



This function squashes the value (any value) and gives the value between 0 and 1

How??? and what is **'e'** ???

e here is 'exponential function' the value is **2.71828**

this is how the value is always between 0 and 1.

2.71828^{+x} = Positive value

$2.71828^{-x} = \frac{1}{\text{Positive value}}$ = Value between (0 and 1)

$\frac{1}{1 + \text{Positive value}}$ or $\frac{1}{1 + \frac{1}{\text{Positive value}}}$ = $\frac{1}{1 + e^{-x}}$

Sigmoid Function

So far we know that we first apply the linear equation and apply Sigmoid function for the result so we get the value which is between 0 and 1.
The hypothesis for *Linear regression* is **h(X) = θ0+θ1*X**
The hypothesis for this algorithm is

$$\frac{1}{1 + e^{-(\theta 0+\theta 1*X)}}$$

How does it work??

- First we calculate the **Logit function,** what the heck is that??

logit = θ0+θ1*X (hypothesis of linear regression)
2. We apply the above Sigmoid function (Logistic function) to logit.
3 we calculate the error , Cost function (Maximum log-Likelihood)
Cost function for linear regression is

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Cost function

here it does not work as h(x) hypothesis gives non convex function for J(**θ0,θ1)** so we are not guaranteed that we reach best minimum.

Simple Step by Step Process of Logistic Regression:

Step-1: y = mx+c ——————————Equation -1

Step-2: Slope m and intercept c are linear based intercepts and linear based slope values

Step-3: By using m and c we may get probabilities under negative values. So to range the probabilities values between 0 to 1. We use sigmoid equation.

Step-4: beta = 1/1+e^(-m) and epsilon = 1/1+e^(-c)

Step-5: New equation of logistic regression y = bx+e

Step-6: Substitute y value in probability equation e^y/1+e^y

Different families in Logistic Regression:

1. Binomial - If there are two classes
2. Nominal - If there are more two classes
3. Ordinal - More than two classes but in order.

Logistic Regression Assumptions:

The logistic regression method assumes that:

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
- There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is logit(p) = log(p/(1-p)), where p is the probabilities of the outcome.
- There is no influential values (extreme values or outliers) in the continuous predictors
- There is no high intercorrelations (i.e. multicollinearity) among the predictors.

To improve the accuracy of your model, you should make sure that these assumptions hold true for your data. In the following sections, we'll describe how to diagnostic potential problems in the data.

Things to master in Logistic Regression: