Logistic Regression

Logistic Regression is a supervised machine learning algorithm used to solve classification problems.

It is used for predicting a categorical dependent variable using a given set of independent variables (features) that can be either continous or categorical.

The outcome of logistic regression, therefore can be either Yes or No, 0 or 1, true or false etc, but instead of giving the exact value as 0 or 1, the algorithm gives probabilistic values which lie between 0 and 1.

The algorithm defines the probability of either 0 or 1 using the concept of threshold value ie, the values above the threshold value tends to 1 and values below the threshold value tends to 0.

How does logistic regression work?

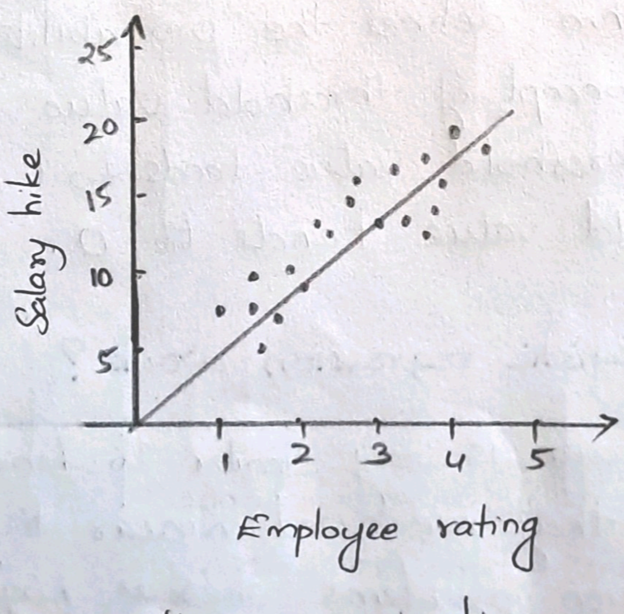
Logistic regression is very similar to Linear Regression except how they are used. Linear Regression is used for solving regression problems whereas Logistic Regression is used for solving classification problems.

Logistic Regression converts the straight best fit line in linear regression to an S-curve using the sigmoid function which always gives values between 0 and 1.

Example 1:

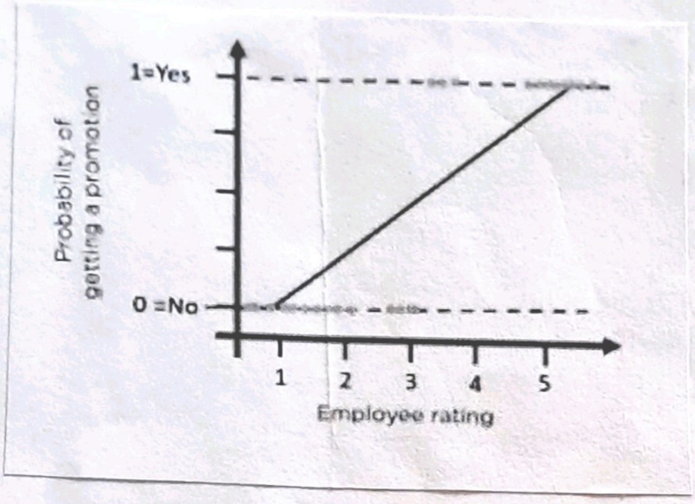
Consider the following example: An organization wants to determine an employees' salary increase based on their performance.

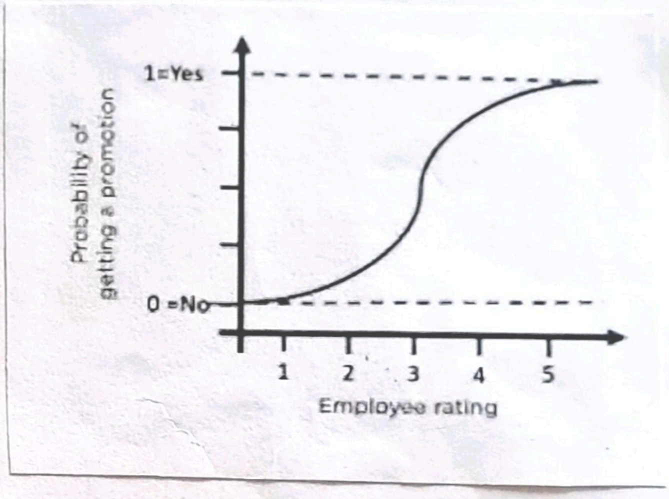
For this purpose, a linear regression algorithm will help them decide. Plotting a regression line by considering the employee's performance as the independent variable and the salary increase as the dependent variable will make the task easier.



Now suppose the organization wants to know whether an employee would get a promotion or not based on their performance.

The above linear graph won't be suitable in this case.
Let us clip the line at zero and one, and convert it into a sigmoid curve (S curve).





So by setting a threshold value, the organization can decide whether an employee will get a salary increase or not.

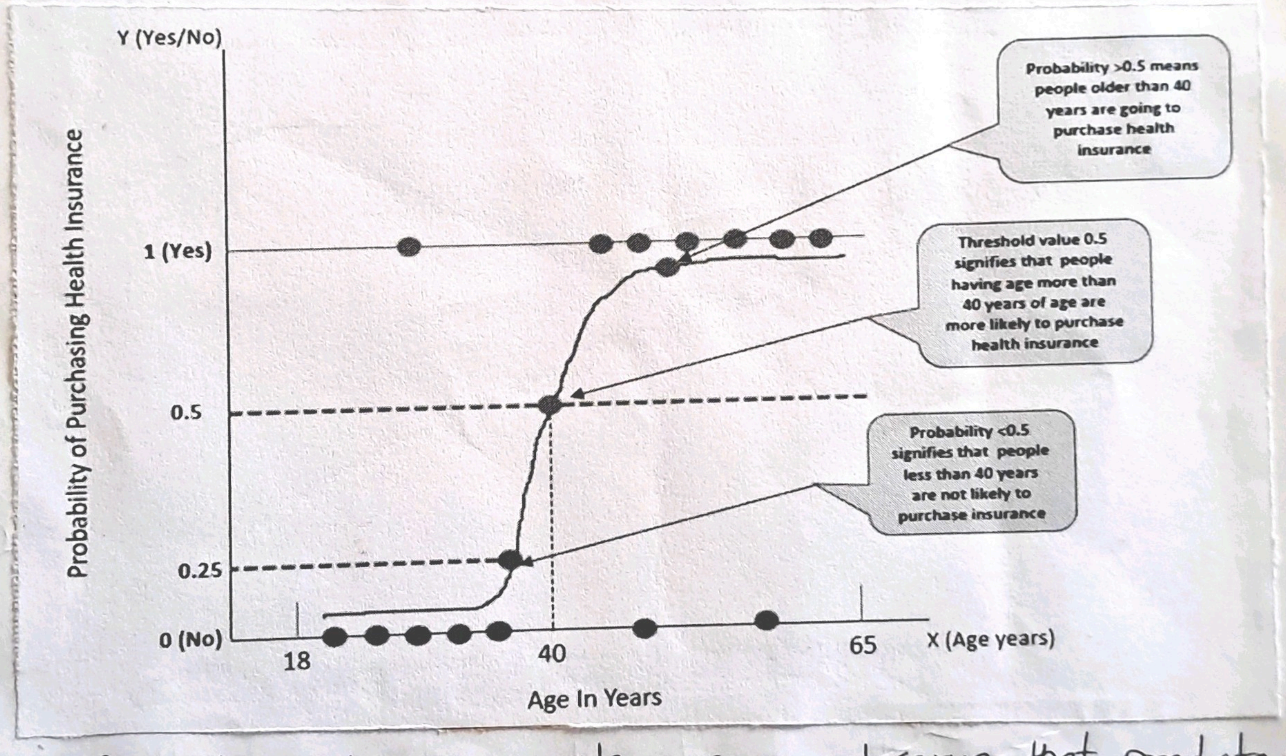
Example 2 :

Consider the problem of purchasing health insurance based on the age of the people. Here purchase is the target variable to be predicted based on the age.

Let us set a threshold value of 0.5. The problem can be generalized by drawing an 'S' shaped curve also known as Sigmoid curve.

People aged 40 years have a probability of 0.5 means, they are likely to purchase health insurance, whereas people having probability >0.5 are definitely going to purchase health insurance.

On the other hand, probability <0.5 signifies there are very less chances or no chances of purchasing health insurance.



So from the above example, we can observe that predicted values can be mapped to probabilities using a mathematical function known as the sigmoid function.

∴ Any real value can be mapped into a value within a range of 0 and 1 which signifies that the value of the logistic regression must be between 0 and 1.

How does logit / sigmoid function squeeze the o/p of linear regression b/w 0 and 1 ?

For a classic logistic regression problem, y is a binary variable with 2 possible values such as win/loss, good/bad etc. Since y is binary, we often label the classes as either 0 or 1, with 1 being the desired class of prediction.

The probability given the i/p variable x is written as

$$P(y=1/x) = P$$

How does probability link to a classification problem?

The higher the value of p, the more likely the new observation belongs to class y=1, instead of y=0.

Odds and log odds.

Odds are the ratio of the probability of something happening to the probability of something not happening.

ie, it is a metric representing the likelihood of the event occurring.

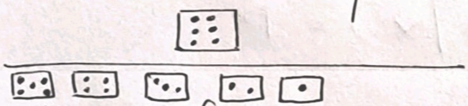
$$\text{Odds} = \frac{P}{1-P} \left(\frac{\text{probability of something happening}}{\text{probability of something not happening}} \right)$$

Odds $= \frac{P}{1-P}$: What does it imply?

When the odds are b/w 0 and 1, the odds are against the observation belonging to $y=1$.

When the odds are greater than 1, the odds are for the observation belonging to $y=1$.

For eg. Let us bet that a six will come up for a toss of a fair six sided die. The probability of it happening is $1/6$.



So the odds in favor of us winning are $(1/6)/(\frac{5}{6}) = \frac{1}{5}$ or 1:5.

The odds of us losing are $(5/6)/(\frac{1}{6}) = 5:1$

Clearly the odds are against us winning

Log odds ie $\log\left(\frac{P}{1-P}\right)$ is merely taking the logarithm of odds

Log (odds) is also called the logit function.

Why do we want to take log of odds?

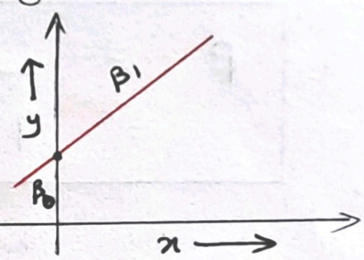
Because we want to adapt the well studied Linear Regression algorithm to classification problems.

Classification tasks have o/p being a probability p ranging from 0 to 1. We can apply linear regression on the transformation if we can transform p to a range from $-\infty$ to $+\infty$

p ranges from 0 to 1.
Odds ($p/1-p$) range from 0 to $+\infty$.
After log transformation log (odds) ranges from $-\infty$ to $+\infty$

Consider the equation of a straight line

$$y = \beta_0 + \beta_1 x.$$



To predict the odds of success, we use the following formula

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x.$$

$$\log\left(\frac{p}{1-p}\right) = \log(\text{odds})$$

Taking exponent on both sides

$$\frac{p}{1-p} = e^{\log(\text{odds})}$$

$$p = (1-p)\, e^{\log(\text{odds})}$$

$$p = e^{\log(\text{odds})} - p \cdot e^{\log(\text{odds})}$$

$$p + p e^{\log(\text{odds})} = e^{\log(\text{odds})}$$

$$\therefore p = \frac{e^{\log(odds)}}{1 + e^{\log(odds)}}$$

ie, $$p = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

$$\boxed{p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}}$$

This is the sigmoid function / logit function.

Assumptions of Logistic Regression:

1. Linearity: The relationship b/w the independent variables and dependent variables should be linear.
2. Independence: The independent variables should not be dependent on themselves.
3. Absence of multicollinearity: The independent variables are not highly correlated with each other.
4. Binary dependent variable: The dependent variable is binary ie 0 or 1.

Types of Logistic Regression:

Logistic regression may be classified into 3 major types as follows:

1. Binary Logistic Regression: When the dependent variable has only 2 categories. For eg. we have to predict whether a patient is suffering from diabetes or not based on a certain medical condition.

2. Multinomial Logistic Regression: When the dependent variable has more than two categories. For eg. we have to predict whether a banking transaction is safe, fraudulent or doubtful based on account level and transaction level details.

3. Ordinal Logistic Regression: When the dependent variable has ordered categories. For eg. we have a situation for diagnosing a patient having high mild or severe viral infection based on a certain medical condition.

Numerical on Logistic Regression:

5 student's dataset of pass/fail in an exam is given. If we use logistic regression as a classifier and assume the model suggested optimizer is given by the odds function of passing the exam

$$\log(\text{odds}) = (-64) + 2 * \text{hours study}.$$

Hour study	Result
29	0
15	0
33	1
28	1
39	1

a. Calculate the probability of pass for a student who studies 33 hours.

b. How many hours a student should study to ensure probability of passing is 95% or more?

Sol:- a) Given $y = -64 + 2 * \text{hours study}$

$$= -64 + 2 * 33$$

$$= 2$$

$$p = \frac{1}{1+e^{-y}} = \frac{1}{1+e^{-2}} = \frac{1}{1+2.8^{-2}} = 0.88$$

$\therefore$ The probability of pass for a student who studies 33 hours is 88 %.

b) $.95 = \dfrac{1}{1+e^{-y}}$

$\therefore 0.95 + 0.95e^{-y} = 1$

$e^{-y} = \dfrac{0.05}{0.95} = 0.0526$

Taking log, $\log(e^{-y}) = \log(0.0526)$

$-y = -2.94$

$y = 2.94$

$y = 64 + 2 * \text{hours study}$

$2.94 = 64 + 2 * \text{hours study}$

$\therefore \text{hours study} = \dfrac{2.94 + 64}{2}$

$= \underline{\underline{33.5 \text{ hours}}}$