SAIF ULLAH **ID:15598** **MS (CS)**

1. Logistic Regression

Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

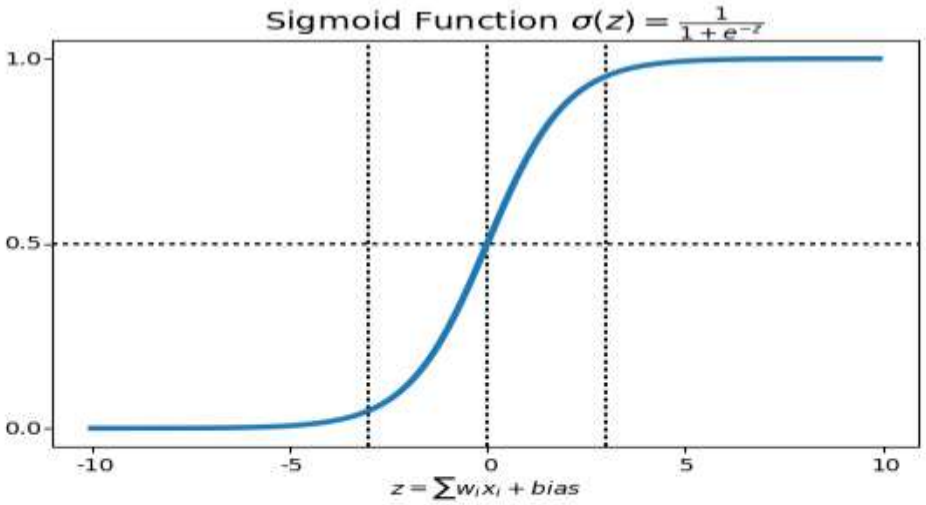
We can call a Logistic Regression a Linear Regression model but the Logistic Regression uses a more complex cost function, this cost function can be defined as the 'Sigmoid function' or also known as the 'logistic function' instead of a linear function.

Basically it is a modeling of an occurring vs event not occurring. This is concept is too much related with probability.

Probability = prob of event occurring vs prop of event not occurring

You can given two types of data it may be category base or numeric base category base mean if we say temperature so it may be hot , cold , humid etc and numeric data mean in the form rang mean 1 to 10 or 10 to 20 and so on so it's a simple term of logistic regression mean event occurring vs non event occurring .

When we talk about



Sigmoid Function Graph

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

logistic regression one thing come with our mind that is binomial distribution .

Binomial distribution:

In simple Binomial distribution mean binary data. Mean data in the form of 0's and 1's. or we can say that it's a classification of binary data or say calculate unknown probability.

P= probability that event will occur

q= probability that event will not occur

p=1-q

Odds Ratio:

Before knowing odd ratio first we have to know what is odds.

Odds: Odds=p(occurring)/p(not occurring)

For example: we have toss 1 coin.

So probability is that we have result head

So probability of head is ½ and not occurring probability mean tail is also ½ .so if you calculate the odds so it will be

Odds= ½ / ½ = 1

Now what is odds ratio

Mean compare the current event with the previous event.

Odd ratio= $p_1/1-p_1/p_0/1-p_0$

$P_{0=}$ previous event occurring

P_1=current event occurring

Logit function:

Logit(p)= l_n(odds)

Logit(p)= l_n(p/1-p)

This is basically your logit function

Now we have to find the inverse of logit function

$$\text{Logit}^{-1}(\alpha) = 1/\ 1+e^{-\alpha}$$

Finally if we have to find regression we have to re arranged the logit function.

$$l_n(p/1-p) = \text{Logit}(p)$$

$$l_n(p/1-p) = \beta_0 + \beta_1 x$$

$$p/1-p = e^{\beta 0 + \beta 1x}$$

$$p = (1-p)\, e^{\beta 0 + \beta 1x}$$

$$p = e^{\beta 0 + \beta 1x} - p \,.\, e^{\beta 0 + \beta 1x}$$

$$p + p \,.\, e^{\beta 0 + \beta 1x} = e^{\beta 0 + \beta 1x}$$

$$p(1 + e^{\beta 0 + \beta 1x}) = e^{\beta 0 + \beta 1x}$$

$$\mathbf{p=)= e^{\beta 0 + \beta 1x}/1+ e^{\beta 0 + \beta 1x}}$$

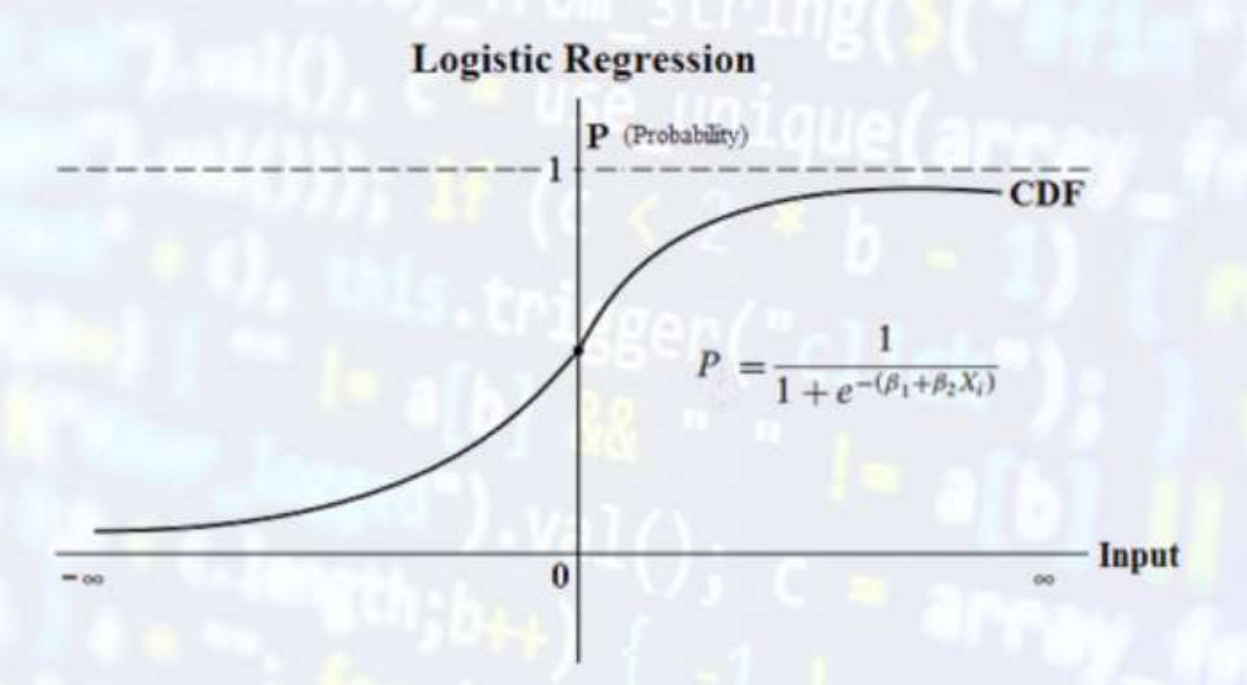
So this is your simple logistic regression. Through this you can calculate your probability and plot a graph from it.

If you follow multiple regression then it will be

$$P = e^{\beta 0 + \beta 1x1 + \beta 1x2 \ldots. + \beta nxn}/1+ e^{\beta 0 + \beta 1x1 + \beta 1x2 \ldots. + \beta nxn}$$

It's your Multiple regression.

So simply if we see logistic regression mean map an event w.r.t event not occurring.



Prepare Data for Logistic Regression:

The assumptions made by logistic regression about the distribution and relationships in your data are much the same as the assumptions made in linear regression.

Much study has gone into defining these assumptions and precise probabilistic and statistical language is used. My advice is to use these as guidelines or rules of thumb and experiment with different data preparation schemes.

Ultimately in predictive modeling machine learning projects you are laser focused on making accurate predictions rather than interpreting the results. As such, you can break some assumptions as long as the model is robust and performs well.

- **Binary Output Variable**: This might be obvious as we have already mentioned it, but logistic regression is intended for binary (two-class) classification problems. It will predict the probability of an instance belonging to the default class, which can be snapped into a 0 or 1 classification.

- **Remove Noise**: Logistic regression assumes no error in the output variable (y), consider removing outliers and possibly misclassified instances from your training data.
- **Gaussian Distribution**: Logistic regression is a linear algorithm (with a non-linear transform on output). It does assume a linear relationship between the input variables with the output. Data transforms of your input variables that better expose this linear relationship can result in a more accurate model. For example, you can use log, root, Box-Cox and other univariate transforms to better expose this relationship.
- **Remove Correlated Inputs**: Like linear regression, the model can overfit if you have multiple highly-correlated inputs. Consider calculating the pairwise correlations between all inputs and removing highly correlated inputs.
- **Fail to Converge**: It is possible for the expected likelihood estimation process that learns the coefficients to fail to converge. This can happen if there are many highly correlated inputs in your data or the data is very sparse (e.g. lots of zeros in your input data).

2. K-NN Algorithm(K Nearest neighbor algoraithm):

It is basically a classification algorithm classification is determining what group something belongs in this could be what specify on inside is it the type of tumer etc. In order for K-NN to work you need some examples with correct things .This is just you predict future group that is in what group the future data is inn. We normally refer to that a reference data that is we need some reference data for K-NN Algorithm. Now for the data record that lead to be classified it compute the distance between data record and all of the difference data record.it looks of the K closest data record in the reference data for example if you say k =5 than it will look at 5 closest record in the reference data. Whatever is the majority class in this group of k data records it is a predicted class of algorithm.

Let say that we want to create a program that will predict if someone like cricket football or neither one of them for this example we need to assumed that a person cannot like both cricket and football.so here we have talk about k-NN algorithm with simple example.

So this is what the data we have using this example we are going to see how K-NN algorithm actually works.

K-NN Algorithm Example

Name	Age	Gender	Sports
Ajay	32	M	Football
Mark	40	M	Neither
Sara	16	F	Cricket
Zaira	34	F	Cricket
Sachin	55	M	Neither
Rahul	40	M	Cricket
Pooja	20	F	Neither
Smith	15	M	Cricket
Laxmi	55	F	Football
Michad	15	M	Football

Now here we seen that we can predict the classification of sport based upon age and gender. So if you look at this data what we have name, age, gender and sports.

Now as we need a data with correct classification so we have assumed that we got this data after survey.

So to proceed with the prediction of K-NN what we can do here is as we have data with correct classification k-NN can make prediction let say that we want to predict what kind of sport for a person name anjaleena.

Name	Age	Gender	Sports
Ajay	32	M	Football
Mark	40	M	Neither
Sara	16	F	Cricket
Zaira	34	F	Cricket
Sachin	55	M	Neither
Rahul	40	M	Cricket
Pooja	20	F	Neither
Smith	15	M	Cricket
Laxmi	55	F	Football
Michad	15	M	Football
Anjaleena	5	F	?

Now we have to find what class of sport should belong to we can efficiently used the K-NN Algorithm to know the sport of the person name angaleena.

As we say earlier that K-NN is a classification Algorithm so we can efficiently used K-NN Algorithm to find out where the person name angaleena with the age of 5 and gender female that which sports belong to her. Whether she belong to the group of people who like cricket or football or none of them.

Let see how this can be done using K-NN algorithm

Let say we assume value of k is 3

Some of the data is numeric like age and other is discrete like gender.

So first of all we need to convert discrete data into numeric data let say that for male we identify it with 0 and for female we use 1

Female = F =1

Male =M = 0

After converting the data into numeric let proceed the K-NN Algorithm so the table will be updated.

Name	**Age**	**Gender**	**Sports**
Ajay	32	0	Football
Mark	40	0	Neither
Sara	16	1	Cricket
Zaira	34	1	Cricket
Sachin	55	0	Neither
Rahul	40	0	Cricket
Pooja	20	1	Neither
Smith	15	0	Cricket
Laxmi	55	1	Football
Michad	15	0	Football
Angaleena	5	1	?

K-NN Algorithm is going to find out the distance of angaleena from each point of the reference data. So we will try to find out the distance of angaleena from ajay mark sara and so on . so how to do this

To find the distance the following distance equation is used

$$(X_1 - X_2)^2 + (Y_1 - Y_2)^2$$

Euclidean distance.

We have two piece of data age and gender

The first person in reference is ajay with the age of 32 and is a male we can compute distance between angaleena and ajay using this distance formula.

Ajay Male=0 Age = 32

$= (5-32)^2+(1-0)^2$

$= 729 + 1$

$= 730$

$= 27.02$

So this is the distance between angaleena and ajay. That is the distance of angaleena from one of the reference point from the reference data.So this is the way you can calculate distance between angaleena and other remaining people so once we got all the distances we get the following particular table.

Name	**Age**	**Gender**	**Distance**	**Class of sport**
Ajay	32	0	27.02	Football
Mark	40	0	35.01	Neither
Sara	16	1	11.00	Cricket
Zaira	34	1	9.00	Cricket
Sachin	55	0	50.01	Neither
Rahul	40	0	35.01	Cricket
Pooja	20	1	15.00	Neither
Smith	15	0	10.00	Cricket

Laxmi	55	1	50.00	Football
Michal	15	0	10.05	football

This is what we got after calculating the distance between angaleena and all the other people.

Now we already assumed the value of k=3. So we look for three (3) closest records to angaleena accordingly to the particular table we have find the 3 closest record i:e

Zaira ,distance is 9.00 smith distance is 10.00 and michal distance is 10.05. Now look at the classification of this three nearest neighbor if you look at these three neighbor for example zaira like cricket smith like cricket and michal like football

Zaira 9.00 cricket

Smith 10.00 cricket

Michal 10.05 football

Now check the most common among the most nearest neighbor which is cricket so according to the K-NN prediction angaleena will be in the class of people who like cricket . that's how K-NN Algorithm works to classified data build upon reference data that already known what group or class of data belong too.

So the table will be.

Name	**Age**	**Gender**	**Sports**
Ajay	32	0	Football
Mark	40	0	Neither
Sara	16	1	Cricket
Zaira	34	1	Cricket
Sachin	55	0	Neither
Rahul	40	0	Cricket

Pooja	20	1	Neither
Smith	15	0	Cricket
Laxmi	55	1	Football
Michad	15	0	Football
Angaleena	5	1	Cricket

3. Naive Bayes algorithm:

Naïve bayes algorithm is the algorithm that learn the probability of an object with certain features belonging to a particular group or class.

For instance, if you are trying to identify a fruit based on its color , shape , and taste, then an orange colored, spherical, and tangy fruit would most likely be orange.

All of these properties individually contribute to the probability that this fruit is an orange and that is why it is known as ‘naïve’.

As for the ‘bayes’ part, it refers to the statistician and philosopher, Thomas bayes and the theorem named after him, bayes theorem, which is the base for naïve bayes algorithm.

The Mathematics of the Naïve Bayes Algorithm:

P(A)(B)=P(B)(A)P(A)/P(B)

P(A)(B): Probability (conditional probability)of occurrence of event A given the event B is true.

P(A) and P(B): Probability of the occurrence of event A and B respectively.

P(B)(A): Probability of the occurrence of event B given the event A is true.

A is called proposition and B is called the evidence.

P(A) is called the prior probability of proposition and P(B) is called the prior probability of evidence.

P(A)(B) is called posterior.

P(B)A) is called likelihood.

Posterior = (Likelihood)(proposition prior probability) / Evidence prior probability

Suppose you have to draw single card from a standard deck of 52 cards. Now the probability that the card is a queen is P(Queen)=4/52=1/13 you are given evidence that the card that you have picked is a face card, the posterior probability p(Queen/Face)can be calculated using Bayes Theorem as follow:

P(Queen/Face)= p(Face/Queen) / p(Face) .p(Queen)

Now p(Face/Queen)=1 because given the card is Queen. It is defiantly a face card. We have already calculated p(Queen). The only value left to calculate is p(Face). Which is equal to 3/13 as there are three face card for every suit in a deck. Therefore

P(Queen/Face)=1/13 . 3/ 13 =1/3

Example:

In a machine learning classification problem, there are multiple features and classes, say. The main aim in the naïve bayes algorithm is to calculate the conditional probability of an object with a feature vector belongs to a particular class.

$$P(C_i|x_1,x_2 \ldots x_n) = P(x_1,x_2 \ldots x_n|C_i).P(C_i) / P(x_1,x_2 \ldots x_n)$$

Step 1:

Create a frequency table for all the features against the different classes.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps 2:

Draw the likelihood table for the features against the classes.

Likelihood Table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Step 3:

Now , use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

Problem:

Players will play if weather is sunny. Is this statement is correct?

We can solved it using above discussed method of probability .

P(Yes/sunny)=P(sunny | yes)*P(yes / P(Sunny)

Here we have

P(sunny | yes) = 3/9 =0.33

P(sunny)= 5/14 =0.36

P(yes)= 9/14=0.64

Now, P(Yes | Sunny)= 0.33 * 0.64/ 0.36= 0.60, which has higher probability.

Naïve bayes uses a similar method to predict the probability of different class based on various attributes. This algorithm is mostly used in text classification and with problem having multiple classes.

THE END