Department of Computer Engineering
Academic Term: July-November 2022

Class : B.*E. Computer* **Subject Name :*ML***
Semester : VII **Subject Code : *CSC701***

Practical No:	3
Title:	**To Study and implement Logistic regression**
Date of Performance:	22/08/2022
Roll No:	8953
Name of the Student:	Brendan Lucas

Evaluation:

Indicator	**Very Poor**	**Poor**	**Average**	**Good**	**Excellent**
Timeline (2)	More than three sessions late (0)	More than two sessions late (0.5)	Two sessions late (1)	One session late (1.5)	Early or ontime (2)
Efforts(4)	N/A	N/A	Not Completed (1)	Partially Completed (2)	Completed(4)
Legibility(2)	N/A	N/A	Poor(1)	Good(1.5)	Very Good(2)
Oral Assessment (2)	N/A	N/A	N/A	Partially Understood the concept (1)	Understood the concept(2)

Total Marks :

Signature of the Teacher :

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Academic Year: 2022-2023 | **Semester: VII**

Subject: Machine Learning | **Class / Division: BE/COMP/B**

Name :- Brendan Lucas | Roll Number: 8953

Experiment No.: 3

Aim : To Study and implement Logistic regression.

I-OBJECTIVE

- To understand basic concepts of Logistic Regression
- To implement the logistic regression.

II-THEORY

Introduction to Logistic Regression

Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature.

That means Logistic regression is usually used for Binary classification problems.

Binary Classification refers to predicting the output variable that is discrete in two classes.

A few examples of Binary classification are Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, etc.

Types of Logistic Regression

- Simple Logistic Regression: a single independent is used to predict the output
- Multiple logistic regression: multiple independent variables are used to predict the output

Extensions of Logistic Regression

Although it is said Logistic regression is used for Binary Classification, it can be extended to solve multiclass classification problems.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

Multinomial Logistic Regression: The output variable is discrete in three or more classes with no natural ordering.

Food texture: Crunchy, Mushy, Crispy

Hair colour: Blonde, Brown, Brunette, Red

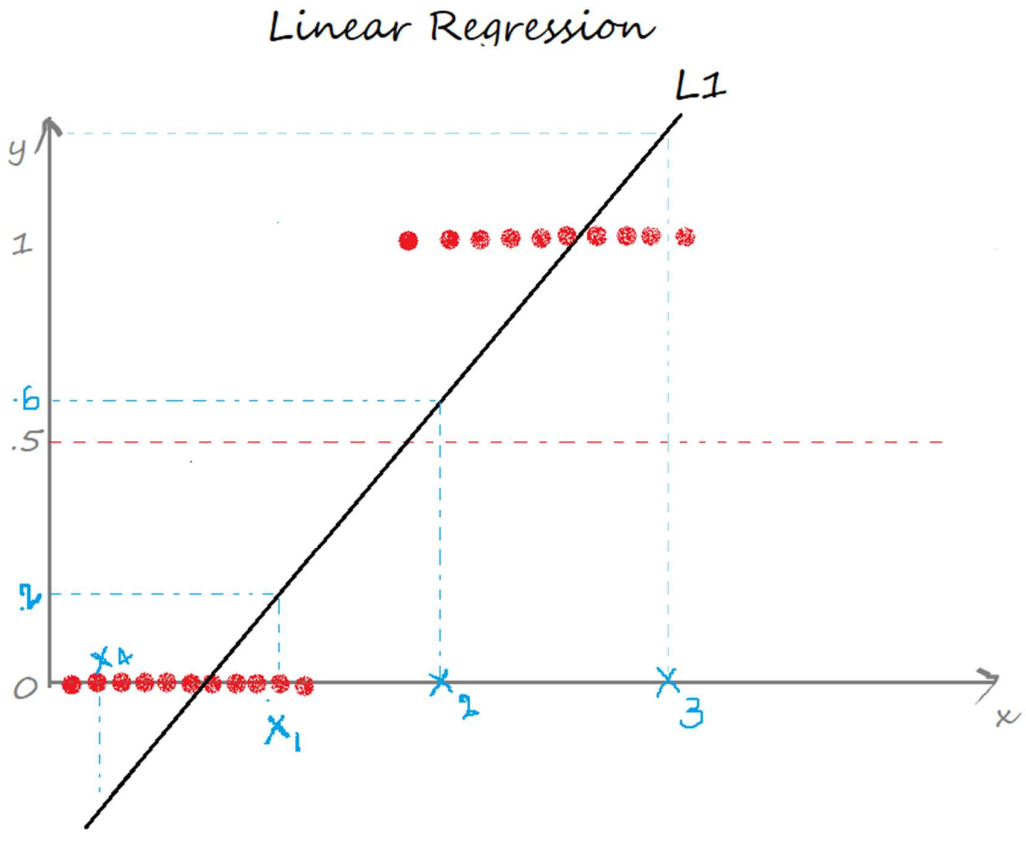
Ordered Logistic Regression: Aka Ordinal regression model. The output variable is discrete in three or more classes with the ordering of the levels.

Customer Rating: extremely dislike, dislike, neutral, like, extremely like

Income level: low income, middle income, high income

Use Linear Regression for classification

Now, let us try if we can use linear regression to solve a binary class classification problem. Assume we have a dataset that is linearly separable and has the output that is discrete in two classes (0, 1).



Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

In Linear regression, we draw a straight line(the best fit line) L1 such that the sum of distances of all the data points to the line is minimal. The equation of the line L1 is y=mx+c, where m is the slope and c is the y-intercept.

We define a threshold T = 0.5, above which the output belongs to class 1 and class 0 otherwise.

$$y = mx + c, \text{ Threshold } T = 0.5$$

$$y = \begin{cases} 1, & mx+c \geq 0.5 \\ 0, & mx+c < 0.5 \end{cases}$$

Case 1: the predicted value for x1 is ≈0.2 which is less than the threshold, so x1 belongs to class 0.

Case 2: the predicted value for the point x2 is ≈0.6 which is greater than the threshold, so x2 belongs to class 1.

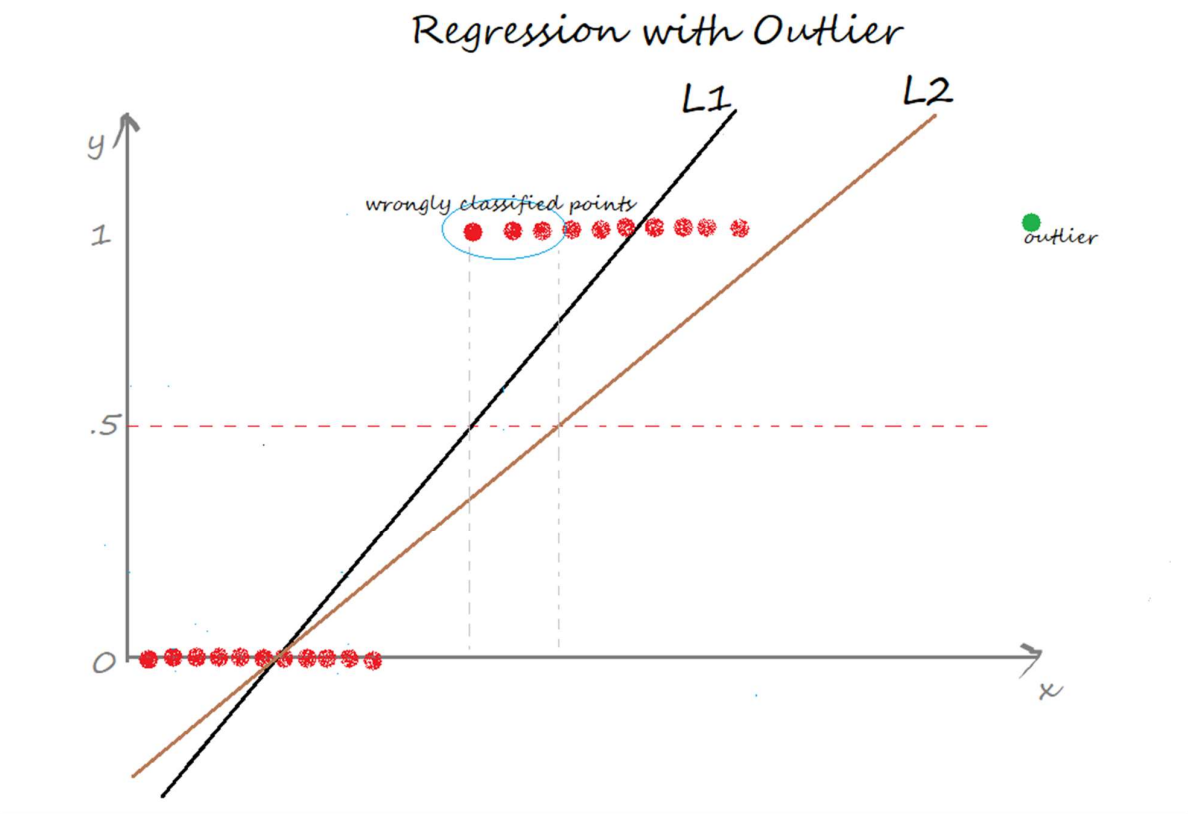
So far so good, yeah!

Case 3: the predicted value for the point x3 is beyond 1.

Case 4: the predicted value for the point x4 is below 0.

The predicted values for the points x3, x4 exceed the range (0,1) which doesn't make sense because the probability values always lie between 0 and 1. And our output can have only two values either 0 or 1. Hence, this is a problem with the linear regression model.

Now, introduce an outlier and see what happens. The regression line gets deviated to keep the distance of all the data points to the line to be minimal.



L2 is the new best-fit line after the addition of an outlier. Seems good till now. But the problem is, if we closely observe, some of the data points are wrongly classified. Certainly, it increases the error term

This again is a problem with the linear regression model.

> **The two limitations of using a linear regression model for classification problems are:**
>
> - **the predicted value may exceed the range (0,1)**
> - **error rate increases if the data has outliers**

There definitely is a need for Logistic regression here.

How does Logistic Regression Work?

The logistic regression equation is quite similar to the linear regression model.

Consider we have a model with one predictor "x" and one Bernoulli response variable "ŷ" and p is the probability of ŷ=1. The linear equation can be written as:

$$p = b_0 + b_1 x \quad \text{--------> eq 1}$$

The right-hand side of the equation (b_0+b_1x) is a linear equation and can hold values that exceed the range (0,1). But we know probability will always be in the range of (0,1).

To overcome that, we predict odds instead of probability.

Odds: The ratio of the probability of an event occurring to the probability of an event not occurring.

Odds = p/(1-p)

The equation 1 can be re-written as:

$$p/(1-p) = b_0+b_1x \quad \text{--------> eq 2}$$

Odds can only be a positive value, to tackle the negative numbers, we predict the logarithm of odds.

Log of odds = ln(p/(1-p))

The equation 2 can be re-written as:

$$\ln(p/(1-p)) = b_0+b_1x \quad \text{--------> eq 3}$$

To recover p from equation 3, we apply exponential on both sides.

$$\exp(\ln(p/(1-p))) = \exp(b_0+b_1x)$$
$$e^{\ln(p/(1-p))} = e^{(b0+b1x)}$$

From the inverse rule of logarithms,

$$p/(1-p) = e^{(b0+b1x)}$$

Simple algebraic manipulations

$$p = (1-p) * e^{(b0+b1x)}$$
$$p = e^{(b0+b1x)} - p * e^{(b0+b1x)}$$

Taking p as common on the right-hand side

$$p = p * ((e^{(b0+b1x)})/p - e^{(b0+b1x)})$$
$$p = e^{(b0+b1x)} / (1 + e^{(b0+b1x)})$$

Dividing numerator and denominator by $e^{(b0+b1x)}$ on the right-hand side

$$p = 1 / (1 + e^{-(b0+b1x)})$$

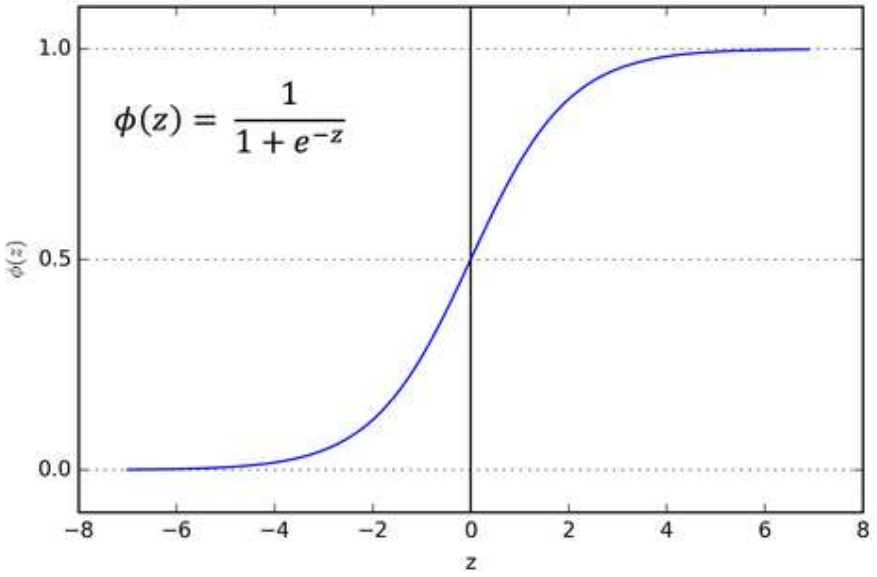
Similarly, the equation for a logistic model with 'n' predictors is as below:

$$p = 1/ (1 + e^{-(b0+b_1x1+b2x2+b3x3+----+bnxn)})$$

The right side part looks familiar, isn't it? Yes, it is the sigmoid function. It helps to squeeze the output to be in the range between 0 and 1.

Sigmoid Function:

The sigmoid function is useful to map any predicted values of probabilities into another value between 0 and 1.



We started with a linear equation and ended up with a logistic regression model with the help of a sigmoid function.

Linear model: $\hat{y} = b_0+b_1x$

Sigmoid function: $\sigma(z) = 1/(1+e^{-z})$

Logistic regression model: $\hat{y} = \sigma(b_0+b_1x) = 1/(1+e^{-(b_0+b_1x)})$

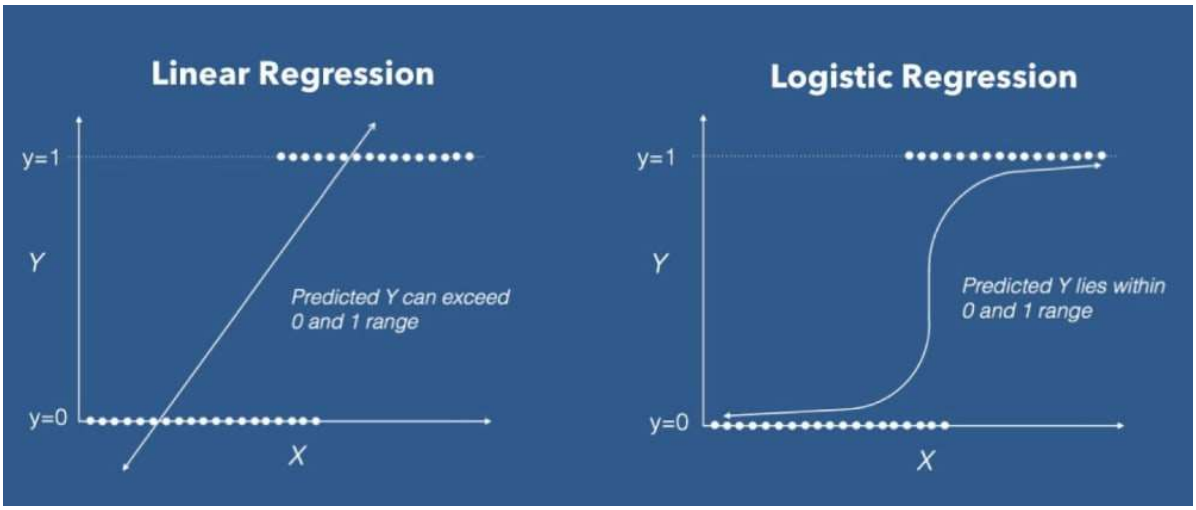
So, unlike linear regression, we get an 'S' shaped curve in logistic regression.

Fr. Conceicao Rodrigues College of Engineering

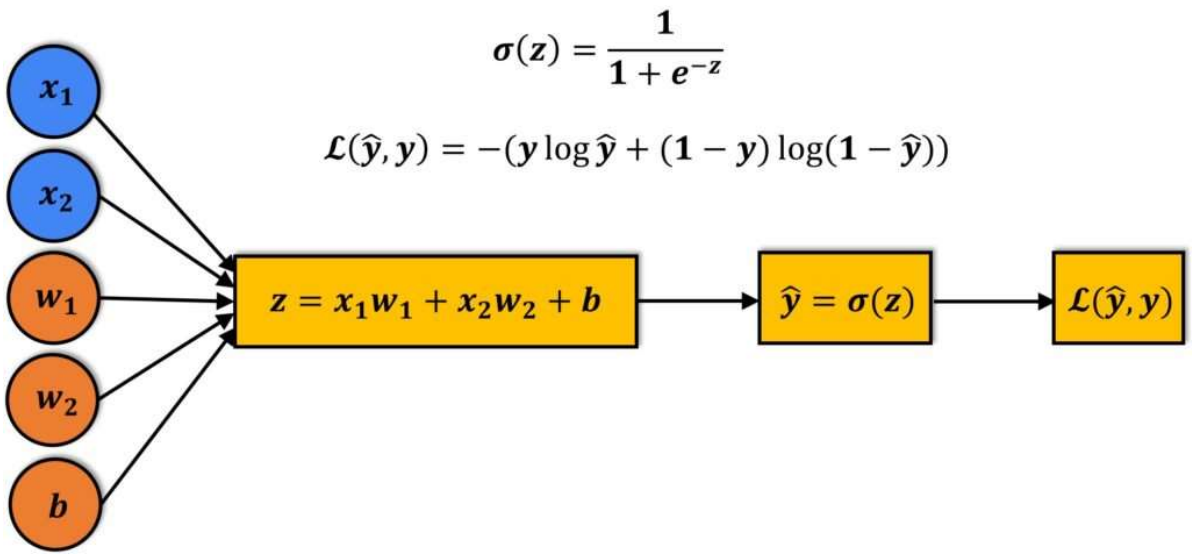
Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023



The image that depicts the working of the Logistic regression model



A linear equation (z) is given to a sigmoidal activation function (σ) to predict the output (ŷ).

To evaluate the performance of the model, we calculate the loss. The most commonly used loss function is the mean squared error.

But in logistic regression, as the output is a probability value between 0 or 1, mean squared error wouldn't be the right choice. So, instead, we use the cross-entropy loss function.

The cross-entropy loss function is used to measure the performance of a classification model whose output is a probability value.

Fr. Conceicao Rodrigues College of Engineering

Bandstand Bandra (West) Mumbai 400053

Department of Computer Engineering

AY - 2022-2023

III IMPLEMENT THE FOLLOWING PROBLEM STATEMENTS

Logistic Regression

1. An experiment is done to test the effect of a toxic substance on insects. At each of six dose levels, 250 insects are exposed to the substance and the number of insects that die is counted, the data is tabulated as below. Find the logistic equation.

Dose	SampSize	Deaths
1	250	28
2	250	53
3	250	93
4	250	126
5	250	172
6	250	197

2. Students in STAT 200 at Penn State were asked if they have ever driven after drinking. They also were asked, "How many days per month do you drink at least two beers?" Find the logistic equation.

Variable	Value	Count
DrivDrnk	Yes	122 (Event)
	No	127
	Total	249

3. There is a dataset given which contains the information of various users obtained from the social networking sites. There is a car making company that has recently launched a new SUV car. So the company wanted to check how many users from the dataset, wants to purchase the car. Build a Machine Learning model using the Logistic regression algorithm.

IV CODE WITH OUTPUT