Logisitic Regression Report

Donal Loitam

July 15, 2022

Contents

1 Introduction

- Logistic Regression is a classification algorithm used for predicting the categorical dependent variable(like Yes/No, True/False, 0/1) using a given set of independent variables.
- Instead of giving the exact value as 0 and 1, it gives the probabilistic values($0 \leq Pr \leq 1$) and $Pr < 0.5 \implies$ (category 1) , else (not category 1)
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

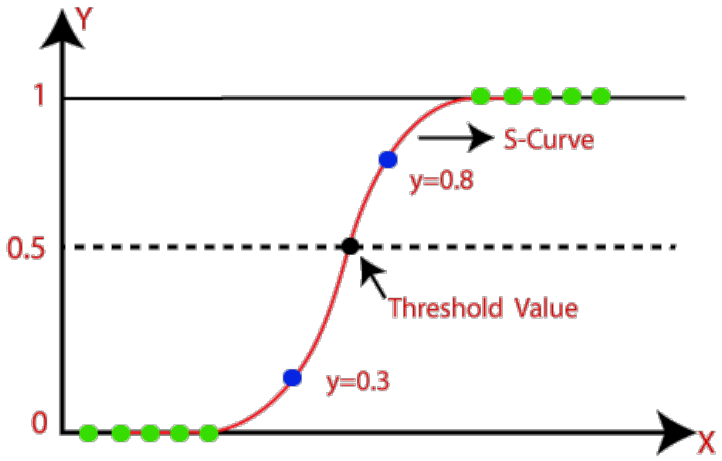

Figure 1: Sigmoid graph

2 Key Points of the Algorithm

As we can see in fig 1, to fit the categorical datas more appropriately we use the sigmoid function $\sigma(z)$ It maps any real value into another value within a range of 0 and 1.

$$\sigma(z) = \frac{1}{1+e^{-z}} \qquad \in [0,1] \tag{1}$$

Define our hypothesis as $h_w(x)$ where w = weights/parameters, x = inputs and $\hat{y}$ be defined as :

$$h_w(x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}} \tag{2}$$

$$\text{Let} \quad \hat{y} = P(y=1|x;w) = h_w(x) \tag{3}$$

$$\therefore P(y=0|x;w) = 1 - h_w(x) \tag{4}$$

$$\hat{y} = \begin{cases} 1 & h(x) \geq 0.5 \\ 0 & h(x) < 0.5 \end{cases} \tag{5}$$

Cost function or the log loss function ($J(w)$) be defined as :

$$L(w) = J(w) = \frac{-1}{m} \sum_{i=1}^{m} [\ ylog(\hat{y}) + (1-y)log(1-\hat{y})] \tag{6}$$

where m = number of datapoints, n = no. of features
Our goal is to choose the parameters so as to minimise the cost function as we train our model, for that we can use Gradient Descent Algorithm We repeat the algorithm until it converges to the global minima i.e update the parameter as follows: (every iteration)

$$w^{(i)} = w^{(i)} - \alpha.(\frac{\partial cost}{\partial w^{(i)}}) \tag{7}$$

$$b = b - \alpha.(\frac{\partial cost}{\partial b}) \qquad \text{where, } b = w^0 \tag{8}$$

where, α = learning rate
After calculations, the bove equations can be simplified to :

$$w_i = w_i - \alpha \sum_{j=1}^{m} (y^{(j)} - h_w(x^{(j)}))x_i^{(j)} \tag{9}$$

$$\tag{10}$$

3 Why does it work ?

It turns out that the log likelihood function $L(w)$ for logistic regression is a concave up graph as in 2 and hence the only minima is a global minima.

What the gradient ($\frac{\partial cost}{\partial w}$) represents is the direction of steepest descent and because gradient is defined that way. While α represents the magnitude of baby step which we will take in the direction of gradient. And since we are taking steps towards the global minima, we will eventually be able to minimse the cost function

Note: α should not be too big because we may never converge to the global minima if we take huge steps Similarly α shouldn't be too small as it will increase computatipnal power and time.

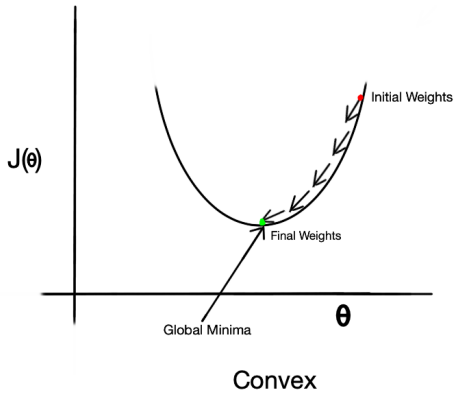

Figure 2: Sigmoid graph

4 Some Questions

1. How do we tackle categorical variables in Logistic Regression ?
Ans. The inputs given to a Logistic Regression model need to be numeric. So we assign each class of the categorical variable a unique numeric value(**dummy variable**), which can then be treated as any other numeric value.
2. What are the assumptions of Logistic Regression ?
Ans.

- There is minimal collinearity among the independent variables i.e. predictors are not correlated.
- The dependent variable must be categorical in nature.

3. Can we solve multiclass classification problems using Logisitic Regression? How?
Ans Yes,we can use a method known as **"one vs all"**. In this method, a number of models are trained, which is equal to the number of classes.
Say for instance, the first model classifies the data as **class 1** or **not class 1**, the second model classifies them into **class 2** or **not class 2** and so on....
This way we can check each data point over all the classes.
4. Why can't we use Mean Square Error(MSE) as cost function for Logistic Regression ?
Ans. In Logistic Regression, we use the sigmoid function to perform a non-linear transformation to obtain the probabilities. If we square this nonlinear transformation, then it will lead to the problem of non-convexity with local minimums and by using gradient descent in such cases, it is not possible to find the global minimum, it may end up to a local minima. Hence in the LOgistic Regression ALgorithm, we used log loss as a cost function as optimizing this function, we can achieve convergence
5. Why can't we use Linear Regression in place of LOgistic Regression for Binary Classification ?
Ans.

- Distribution of error terms : The distribution of data in the case of Linear and Logistic Regression is different. It assumes that error terms are normally distributed. But this assumption does not hold true in the case of binary classification.
- Output of LInear regression: IN Linear regression, the output is continuous(or numeric) which is of not much use for binary classification. On the contrary, Logistic Regression can predict values between 0 and 1 which can then be mapped with probabilty to predict label as 0 or 1