Logistic Regression

Epoch IIT Hyderabad

Arin Aggarwal
MA22BTECH11006

1 Introduction

Logistic Regression is a Statistical model used for Classification problems. It predicts if an event belongs to a particular class or not. It uses sigmoid (S shape curve) function to estimate probability for the given class. The output of this regression must be a value between 0 & 1. Here the best line is decided using Maximum Likelihood criteria unlike Least Squares in Linear Regression.

It can have 3 types:

1) **Binomial** : There are only two possible types of dependent variables, such as 0 or 1, Pass or Fail, etc.
2) **Multinomial** :There can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
3) **Ordinal** : There can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

2 Assumptions of Regression

1) **Linearity** : There should be a linear relationship between each explanatory variable and the logit (log of odds) of the response variable. This can be checked using the Box-Tidwell test.
2) **Outliers** : Make sure that there are no significant outliers in the data as they may skew the results.
3) **Multicollinearity** : The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity. Multicollinearity can be checked by 2 main criterions:
   (i **Correlation Matrix** : When computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be close to 0. +1 indicates perfect positive correlation and -1 perfect negative correlation.
   (ii **Variance Inflation Factor(VIF)** : A VIF value of >10 indicates presence of multicollinearity and thus the given variable should be removed.
4) **Normality** : Data needs to be multi-variate normal. This can be checked using Histograms(by checking if the skew is close to 0) .

3 Working

First, the probabilties are converted to log-odds. This makes the y-axis continuous from $+\infty$ to $-\infty$. Then different lines are made on this graph and projections of log-odds (present on the extremes) are taken on the graph. Then the graph of log-odds is coverted back to probabilities and it forms a sigmoid function. Now, for the maximum likelihood, take the sum of $log(p)_s$ for $p > 0.5$ (True value) and $log(1-p)_s$ for $p < 0.5$ (False value). The line that maximises this sum is our best line. P-values should be used to see effectiveness of independent variable (variable will be redundant if it's p-value > 0.05).

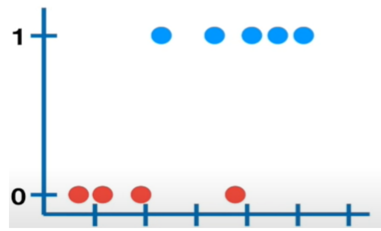

Fig. 4: Given Graph

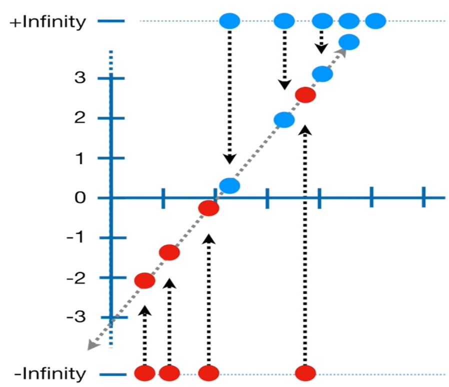

Fig. 4: Graph after log(odds) conversion

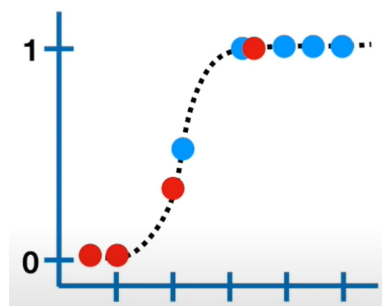

Fig. 4: Final Sigmoid Function

4 Mathematical Formulations

4.1 p and log-odds interconversion

$$odds = \frac{p}{1-p}$$

$$\Rightarrow log(odds) = log(\frac{p}{1-p})$$

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

4.2 R^2 Formula [1]

R^2 requires calculation of Log Likelihood of fitted line (LL_{fit}) and Log Likelihood of null model (LL_{null}).

$$R^2 = 1 - \frac{LL_{fit}}{LL_{null}}$$

[1]This is formula for McFadden's pseudo-R squared. Full maths for it can be checked online. There are other methods to calculate R^2 too.