Logistic Regression for Breast Cancer Wisconsin dataset

Shashank Pathak(15BCE1287)*,
*School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: shashank.pathak2015@vit.ac.in*

***Abstract*—In machine learning, Logistic Regression is a linear model which gives a probablistic approach in classification.The algorithm uses the concept of likelihood.We study the performance of the logistic regression for the classification of the Breast Cancer Wisconsin Dataset.The system developed performs with 36.10% insample accuracy and 32.70% accuracy for the test sample in a two class classification**

1. Introduction

Logistic regression is used to obtain odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest. The main advantage is to avoid confounding effects by analyzing the association of all variables together. In this article, we explain the logistic regression procedure using examples to make it as simple as possible. After definition of the technique, the basic interpretation of the results is highlighted and then some special issues are discussed.

2. Methodology

Logistic regression uses an equation as the representation, very much like linear regression.

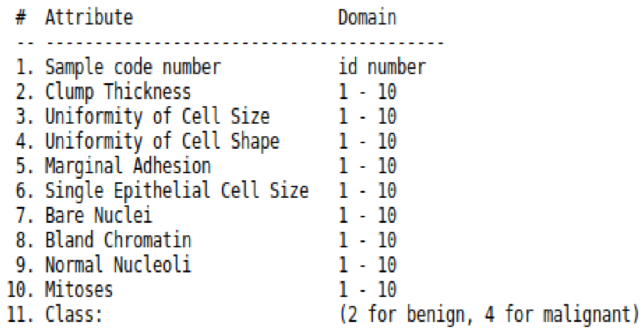
Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary values (0 or 1) rather than a numeric value.

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data.

The actual representation of the model that you would store in memory or in a file are the coefficients in the equation (the beta value or bs).

3. Database - Breast Cancer Wisconsin Dataset(Original)

The Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository is used,to distinguish malignant (cancerous) from benign (non-cancerous) samples.This dataset consist of 699 instances and 11 attributes that would help the classiy the data into the two classes.The attributes are described in the fig.1



```
 #  Attribute                     Domain
 -- -----------------------------------------
 1. Sample code number            id number
 2. Clump Thickness               1 - 10
 3. Uniformity of Cell Size       1 - 10
 4. Uniformity of Cell Shape      1 - 10
 5. Marginal Adhesion             1 - 10
 6. Single Epithelial Cell Size   1 - 10
 7. Bare Nuclei                   1 - 10
 8. Bland Chromatin               1 - 10
 9. Normal Nucleoli               1 - 10
10. Mitoses                       1 - 10
11. Class:                        (2 for benign, 4 for malignant)
```

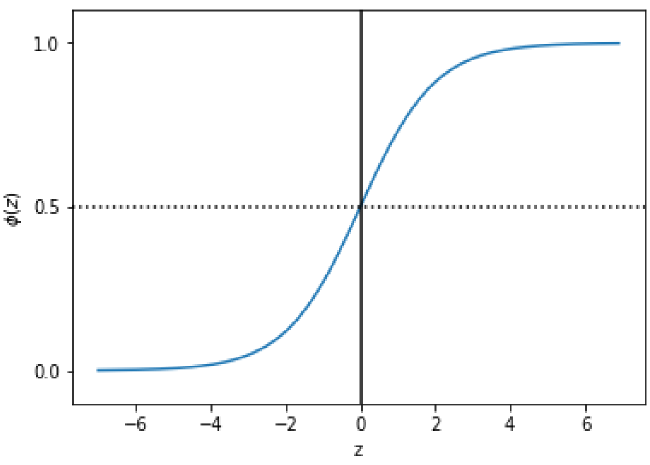
Figure 1. Breast Cancer Wisconsin Dataset Attributes

4. Algorithm

The coefficients (Beta values b) of the logistic regression algorithm must be estimated from your training data. This is done using maximum-likelihood estimation.

Maximum-likelihood estimation is a common learning algorithm used by a variety of machine learning algorithms, although it does make assumptions about the distribution of your data (more on this when we talk about preparing your data).

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{\ln\left(1+e^{-y_n \mathbf{w}^{\intercal}\mathbf{x}_n}\right)}_{\mathrm{e}\left(h(\mathbf{x}_n),y_n\right)}$$



The best coefficients would result in a model that would predict a value very close to 1 (e.g. malignant) for the default class and a value very close to 0 (e.g. benign) for the other class. The intuition for maximum-likelihood for logistic regression is that a search procedure seeks values for the coefficients (Beta values) that minimize the error in the probabilities predicted by the model to those in the data (e.g. probability of 1 if the data is the primary class).

5. Experiments

First we used Logistic Regression from scikit learn library to classify the dataset on a reduced set of attributes namely:

'ClumpThickness', 'UCellSize', 'UCellShape','MargAdhes', 'EpiCellSize', 'BareNuc', 'Chromatin', 'Nucleoli' and 'Mitoses'. After this I did some preprocessing i.e. scaling the data and taking the standardized values.Then I plotted the decision surface using matplotlib and got a visual aspect of logistic regression.

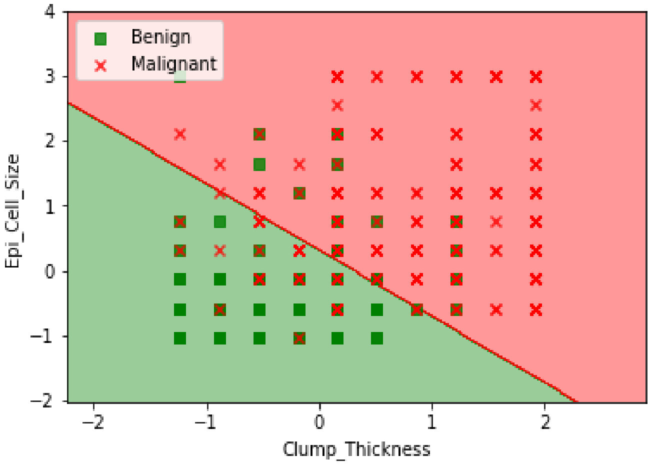


Figure 2.

6. Conclusion

After the classification we can see that the classification was not very effective, there were datapoints which belong to wrong category.This indeed suggest that the Logistic Regression model is not fit to classifify the data into two class Benign and Malignant fot the breast cancer dataset The system developed performs with 36.10% insample accuracy and 32.70% accuracy for the test sample in a two class classification