Logistic Regression
Epoch IIT Hyderabad

Chakka Surya Saketh
AI22BTECH11005

Introduction

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. It is widely employed in various fields, including medicine, marketing, finance, and social sciences, for its ability to handle classification problems.
The main reason to prefer this over linear regression is because the latter one is highly sensitive to outliers. Logistic regression is almost similar to linear regression but it squeezes the output in between 0 and 1

Logistic Function

$$P = \sigma(b + WX) = \frac{1}{1 + e^{-(b+WX)}}$$

This is called the sigmoid function. It squeezes a straight line into a curve.

Cost Function

If we use the same cost function as in the case of linear regression it will give a non convex graph with many local minima.
Hence we use a cost function called the log loss function.

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} -(y_i \log y_{pred} + (1 - y_i)\log(1 - y_{pred}))$$

where

$$y_{pred} = \sigma(\theta^T X_i)$$

Gradient Descent

Gradient descent changes the value of our weights in such a way that it always converges to minimum point. At first gradient descent takes a random value of our parameters from our function. The gradient descent algorithm finds the slope of the loss function at that particular point and then in the next iteration, it moves in the opposite direction to reach the minima.

$$\frac{dP}{dX} = \frac{1}{1 + e^{-(b+WX)}}(1 - \frac{1}{1 + e^{-(b+WX)}})$$

And consider the learning rate be η

$$\theta_{new} = \theta_{old} - \eta\frac{\partial J(\theta)}{\partial \theta}$$

After all the calculations

$$\theta_{new} = \theta_{old} - \eta(\sigma(\theta^T X_i) - y)X_i$$

Assumptions

1) Linearity of Logits: The relationship between the predictor variables and the log-odds of the binary outcome should be approximately linear.

2) Independence: Statistical analyses rely on the assumption that observations are independent and not influenced by one another to avoid biased results i.e there should not be any visible patterns in the error terms.The absence of this phenomenon is known as Autocorrelation.

3) No Multicollinearity: To avoid redundancy and multicollinearity issues, independent variables should not be strongly correlated with each other, preserving the stability of multiple regression models.

Evaluation metrics

1) Confusion matrix: It is used to find amount of values predicted correctly and wrongly. Here false positives(FP) are type 1 errors and false negatives(FN) are type 2 errors.
2) Classification report: This includes 3 parameters namely
   Precision
   Recall
   F1 score
3) Accuracy score: This is the usual metric which predicts the overall accuracy of the model.

$$Accuracy = \frac{FP + FN}{TP + FP + TN + FN}$$

4) ROC Curve: “Receiver Operating Characteristic Curve” is the score which lies between 0 to 1. 0 stands for Bad and 1 stands for Good. 0.5 is better. If ROC score is 0.78 then it means 78% of predicted values are correct and rest 22% are predicted wrongly.