Logistic Regression

Shuvrajeet Das

May 7, 2021

Abstract

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

1 Introduction

A statistical model typically used to model a binary dependent variable with the help of logistic function. Another name for the logistic function is a sigmoid function and is given by:

$$F(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \tag{1}$$

This function assists the logistic regression model to squeeze the values from (-k,k) to (0,1).Here w is reffered as the weights. Logistic regression is majorly used for binary classification tasks; however, it can be used for multiclass classification.

The reason behind this is that just like Linear Regression, logistic regression starts from a linear equation. However, this equation consists of log-odds which is further passed through a sigmoid function which squeezes the output of the linear equation to a probability between 0 and 1. And, we can decide a decision boundary and use this probability to conduct classification task.

1.1 Maths behind Logistic Regression

We could start by assuming p(x) be the linear function. However, the problem is that p is the probability that should vary from 0 to 1 whereas p(x) is an unbounded linear equation. To address this problem, let us assume, log p(x) be a linear function of x and further, to bound it between a range of (0,1), we will use logit transformation. Therefore, we will consider $logp(x)/(1-p(x))$. Next, we will make this function to be linear:

$$log\frac{p(x)}{1-p(x)} = wX + b \tag{2}$$

After solving for $p(x)$:

$$p(x) = \frac{e^{W+b}}{1-e^{W+b}} \tag{3}$$

To make the logistic regression a linear classifier, we could choose a certain threshold, e.g. 0.5.In the later part the decision of the value of the threshold will be discussed.

Since Logistic regression predicts probabilities, we can fit it using likelihood. Therefore, for each training data point x, the predicted class is y. Probability of y is either p if y=1 or 1-p if y=0. Now, the likelihood can be written as:

$$L(b,W) = \prod_{i=1}^{n} p(x_i)^{y_i}(1-p(x_i))^{1-y_i} \tag{4}$$

The multiplication can be transformed into a sum by taking the log:

$$\begin{aligned} l(b,W) &= logL(b,w) \\ &= \sum_{i=0}^{n} y_i logp(x_i) + (1-y_i)log(1-p(x_i)) \\ &= \sum_{i=0}^{n} log(1-p(x_i)) + \sum_{i=0}^{n} y_i log\frac{p(x_i)}{1-p(x_i)} \end{aligned} \tag{5}$$

Maximum Likelihood Estimation: Gradient for weights

$$\begin{aligned}
\frac{\partial l}{\partial w} &= \frac{\partial}{\partial w} y_i log(p(x_i)) + \frac{\partial}{\partial w}(1-y_i)log(1-p(x_i)) \\
&= \left[\frac{y_i}{p(x_i)} + \frac{1-y_i}{1-p(x_i)}\right]\frac{\partial}{\partial w}p(x_i) \quad \because p(x_i) = \sigma(wx+b) \\
&= \left[\frac{y_i}{p(x_i)} + \frac{1-y_i}{1-p(x_i)}\right]\sigma(wx+b)(1-\sigma(wx+b))x \\
&=\rightarrow -(y-\sigma(wx))x \qquad\qquad \because \sigma(x) = \frac{1}{1-e^{-x}}
\end{aligned} \tag{6}$$

Similarly for bias,

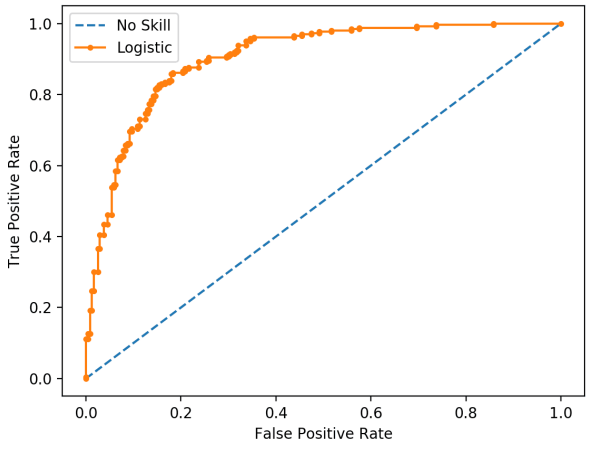
$$\begin{aligned}
\frac{\partial l}{\partial b} &= \frac{\partial}{\partial b} y_i log(p(x_i)) + \frac{\partial}{\partial b}(1-y_i)log(1-p(x_i)) \\
&= \left[\frac{y_i}{p(x_i)} + \frac{1-y_i}{1-p(x_i)}\right]\frac{\partial}{\partial b}p(x_i) \quad \because p(x_i) = \sigma(wx+b) \\
&= \left[\frac{y_i}{p(x_i)} + \frac{1-y_i}{1-p(x_i)}\right]\sigma(wx+b)(1-\sigma(wx+b)) \\
&=\rightarrow -(y-\sigma(wx)) \qquad\qquad \because \sigma(x) = \frac{1}{1-e^{-x}}
\end{aligned} \tag{7}$$

Back to the topic of choosing the threshhold:

A ROC curve is a diagnostic plot that evaluates a set of probability predictions made by a model on a test dataset.

A set of different thresholds are used to interpret the true positive rate and the false positive rate of the predictions on the positive (minority) class, and the scores are plotted in a line of increasing thresholds to create a curve.

Figure 1: ROC curve.



The false-positive rate is plotted on the x-axis and the true positive rate is plotted on the y-axis and the plot is referred to as the Receiver Operating Characteristic curve, or ROC curve.

A diagonal line on the plot from the bottom-left to top-right indicates the "curve" for a no-skill classifier (predicts the majority class in all cases), and a point in the top left of the plot indicates a model with perfect skill.

The curve is useful to understand the trade-off in the true-positive rate and false-positive rate for different thresholds. The area under the ROC Curve, so-called ROC AUC, provides a single number to summarize the performance of a model in terms of its ROC Curve with a value between 0.5 (no-skill) and 1.0 (perfect skill).

Now we define:
Sensitivity = TruePositive / (TruePositive + FalseNegative)
Specificity = TrueNegative / (FalsePositive + TrueNegative)

Where:
Sensitivity = True Positive Rate
Specificity = 1 – False Positive Rate

The Geometric Mean or G-Mean is a metric for imbalanced classification that, if optimized, will seek a balance between the sensitivity and the specificity.

G-Mean = $\sqrt{Sensitivity * Specificity}$

Given that we have already calculated the Sensitivity (TPR) and the complement to the Specificity when we calculated the ROC Curve, we can calculate the G-Mean for each threshold directly. Once calculated, we can locate the index for the largest G-mean score and use that index to determine which threshold value to use.

$\therefore$ we get the threshold value of our model.

2 Conclusion

Thus our model can predict better.