Logistic Regression

Ta Quang Minh

April 27, 2024

1 Introduction

Inspite of its name, Logistic regression is a statistical method used for binary classification tasks, rather than a regression method. Unlike linear regression, which predicts continuous values, logistic regression predicts the probability that an instance belongs to a particular class. It is widely used in various fields such as healthcare, finance, marketing, and social sciences.

2 Mathematical forumalation

Logistic regression models the probability that a binary outcome variable Y belongs to a particular class based on one or more independent variables X. It uses the logistic function (sigmoid function) to map the linear combination of the independent variables to the probability of the outcome:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)}}$$

Where:

- $P(Y = 1|X)$ is the probability that Y equals 1 given the values of X.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients (parameters) of the model.
- $X_1, X_2, \ldots, X_n$ are the independent variables.
- e is the base of the natural logarithm.

In logistic regression, the model is trained by optimizing a loss function known as the binary cross-entropy (also called log loss) or logistic loss function. This loss function quantifies the difference between the predicted probabilities and the true binary labels.

Let's denote:

- y_i as the true binary label (0 or 1) for the i-th observation.
- p_i as the predicted probability that the i-th observation belongs to class 1.

The binary cross-entropy loss for logistic regression is defined as:

$$\text{Binary Cross-Entropy Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right]$$

Where:

- N is the total number of observations.
- log denotes the natural logarithm.

The goal during training is to minimize this loss function, typically using optimization algorithms such as gradient descent or its variants. The binary cross-entropy loss is a convex function, ensuring that gradient-based optimization methods converge to a global minimum.

The loss can also be interpreted as the inverse of the log-likelihood of the predicted distribution given by the logistic function (in which case we want to maximize the likelihood).

3 Decision Boundary of Logistic Regression

In logistic regression with two classes (binary classification), the decision boundary is determined by the coefficients (parameters) of the model. It is either a line in two-dimensional space (for two features) or a hyperplane in higher-dimensional space (for more than two features). Hence logistic regression belongs to the class of linear classifier, and it can not work well with non-linear separable data.

In the case that there are two features X_1 and X_2, the decision boundary is given by the equation:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

4 Example

The dataset used in this experiment was generated using the make_blobs function from scikit-learn. It contains 200 samples with two features and two classes.

The logistic regression model was trained on 80% of the data and tested on the remaining 20%.

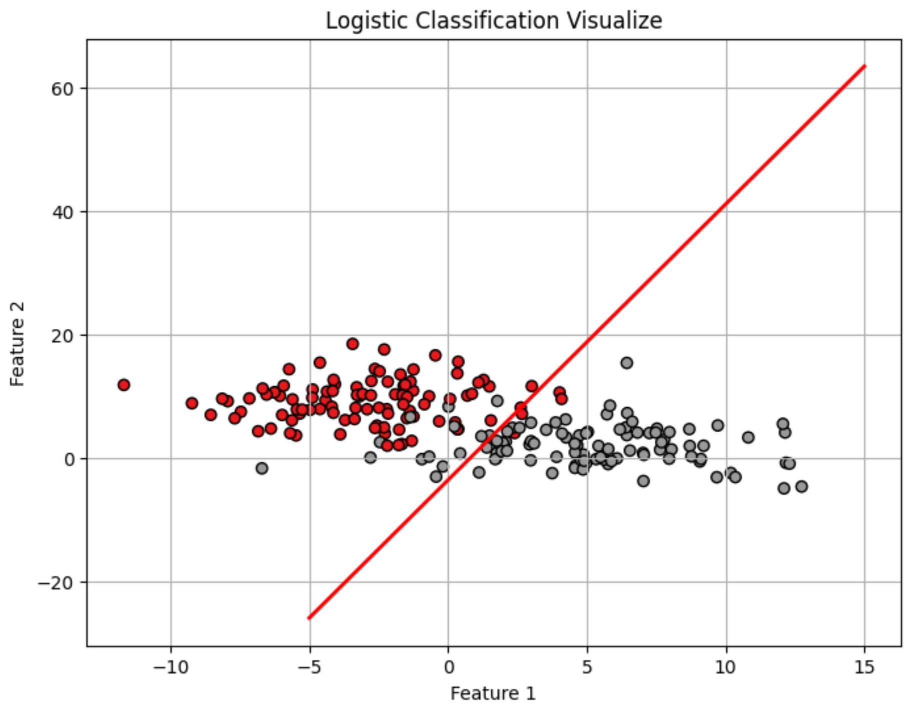

Figure 1: Logistic Regression

We compared the predicted value of the test set with the true value to get the following classification report.

Class	Precision	Recall	F1-Score	Support
0	0.96	1.00	0.98	23
1	1.00	0.94	0.97	17
Accuracy				0.97
Macro Avg	0.98	0.97	0.97	40
Weighted Avg	0.98	0.97	0.97	40

Table 1: Classification Report

5 Applications of Logistic Regression

Logistic regression finds applications in various domains, including:

- **Medical Diagnosis:** Predicting the likelihood of a patient having a particular disease based on symptoms and test results.

- **Credit Scoring:** Assessing the creditworthiness of individuals based on their financial attributes.
- **Customer Churn Prediction:** Predicting whether a customer will churn (leave) a service based on their behavior and demographics.
- **Sentiment Analysis:** Classifying text data (e.g., reviews, tweets) as positive or negative sentiment.

6 Advantages of Logistic Regression

Logistic regression offers several advantages, including:

- **Interpretability:** The coefficients in logistic regression provide insights into the relationship between the independent variables and the probability of the outcome.
- **Efficiency:** Logistic regression is computationally efficient and can handle large datasets with relatively low computational resources.
- **Probability Output:** Logistic regression provides probabilistic predictions, allowing for flexible decision-making thresholds.

7 Limitations of Logistic Regression

Despite its advantages, logistic regression has some limitations:

- **Assumption of Linearity:** Logistic regression assumes a linear relationship between the independent variables and the log odds of the outcome, which may not always hold true.
- **Binary Outcome:** Logistic regression is limited to binary classification tasks and cannot be directly applied to multi-class classification problems without modifications.
- **Sensitivity to Outliers:** Logistic regression can be sensitive to outliers, which may affect model performance.

8 Extensions of Logistic Regression

Several extensions of logistic regression exist to address its limitations and cater to more complex scenarios, including:

- **Multinomial Logistic Regression:** Generalizes logistic regression to handle multi-class classification problems. It uses the softmax function instead of the logistic sigmoid function used in binary logistic regression

- **Regularized Logistic Regression:** Introduces regularization terms (e.g., L1, L2 regularization) to prevent overfitting and improve generalization.
- **Ordinal Logistic Regression:** Extends logistic regression to handle ordinal outcome variables with ordered categories.

9 Conclusion

Logistic regression is a versatile and widely used statistical method for binary classification tasks. By modeling the probability of binary outcomes, logistic regression provides interpretable predictions and can be applied to various real-world problems across different domains. In this report, we have presented an overview of the model, as well as its variants, applications, advantages, and limitations.