Logistic Regression

Vishwa Raghava Reddy Mamindla

1 Introduction

Logistic Regression is a machine learning algorithm used for binary classification tasks. It predicts the probability of an input belonging to a particular class, making it suitable for problems such as spam detection and medical diagnosis.

2 Mathematical Intuition

The sigmoid function, also referred to as the logistic function, serves as the basis of logistic regression. Any input is converted by the sigmoid function to a probability represented by a value between 0 and 1. It's defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Here, z is a linear combination of the features (x_i) and their corresponding weights (w_i):

$$z = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

The logistic function converts (z) into a probability estimate, or the likelihood that the given input belongs to the positive class.

3 Assumptions

Logistic Regression assumes:

- **Linearity:** The log-odds of the projected probability and the features are considered to be linearly related.
- **Errors:** Errors between actual class labels and predicted probabilities are assumed to be independent.

- **Multicollinearity:** To prevent unstable coefficient estimations, features shouldn't be strongly correlated.
- **Sample Size:** A considerably large dataset is necessary for proper coefficient estimates and meaningful results.

4 Maximum Likelihood Estimation

In Logistic Regression, the ideal set of weight coefficients (w_i) is the one that maximizes the likelihood of observing the given data. Maximum Likelihood Estimation (MLE) is used to achieve this. The probability function for (N) data points is written as follows:

$$L(w) = \prod_{i=1}^{N} \left(P(y_i = 1|x_i; w)\right)^{y_i} \left(1 - P(y_i = 1|x_i; w)\right)^{1-y_i}$$

Here, y_i represents the actual class label of the i-th sample, and $P(y_i = 1|x_i; w)$ is the predicted probability of the i-th sample that it belongs to class 1.

Training involves finding the (w) values that maximize the likelihood. Gradient descent can be used for optimization. After training, the decision boundary is determined by the sigmoid function. The point where $P(y = 1|x; w) = 0.5$ corresponds to the decision boundary and is obtained by solving $z = 0$.

5 Applications

Logistic Regression finds applications across domains:

- **Spam Detection:** Classifying emails as spam or non-spam.
- **Medical Diagnosis:** Predicting the presence of a disease based on medical test results.
- **Credit Risk:** Assessing the likelihood of loan default.
- **Customer Segmentation:** Categorizing customers into distinct segments.

6 Summary

Logistic Regression is a fundamental method for binary classification that predicts class probabilities using the sigmoid function. Its mathematical foundations, along with the log-likelihood maximisation, offer a strong context for understanding its behaviour. Logistic Regression can be used for a variety of real-world applications by adopting the assumptions and understanding the training process.