Logistic Regression in Machine Learning

Logistic regression is a **supervised machine learning algorithm** used for **classification tasks** where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression is a statistical algorithm which analyze the relationship between two data factors. The article explores the fundamentals of logistic regression, it's types and implementations.

What is Logistic Regression?

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems.

Key Points:

- Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

Logistic Function – Sigmoid Function

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Types of Logistic Regression

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.

2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as “cat”, “dogs”, or “sheep”
3. **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

Assumptions of Logistic Regression

We will explore the assumptions of logistic regression as understanding these assumptions is important to ensure that we are using appropriate application of the model. The assumption include:

1. Independent observations: Each observation is independent of the other. meaning there is no correlation between any input variables.
2. Binary dependent variables: It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories SoftMax functions are used.
3. Linearity relationship between independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable should be linear.
4. No outliers: There should be no outliers in the dataset.
5. Large sample size: The sample size is sufficiently large

Terminologies involved in Logistic Regression

Here are some common terms involved in logistic regression:

- **Independent variables:** The input characteristics or predictor factors applied to the dependent variable’s predictions.
- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds:** It is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds:** The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.
- **Coefficient:** The logistic regression model’s estimated parameters, show how the independent and dependent variables relate to one another.

- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- Maximum likelihood estimation: The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

How does Logistic Regression work?

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.

Cost function in Logistic Regression in Machine Learning

Logistic Regression is one of the simplest classification algorithms we learn while exploring machine learning algorithms. In this article, we will explore cross-entropy, a cost function used for logistic regression.

What is Logistic Regression?

Logistic Regression is a statistical method used for binary classification. Despite its name, it is employed for predicting the probability of an instance belonging to a particular class. It models the relationship between input features and the log odds of the event occurring, where the event is typically represented by the binary outcome (0 or 1).

The output of logistic regression is transformed using the logistic function (sigmoid), which maps any real-valued number to a value between 0 and 1. This transformed value can be interpreted as the probability of the instance belonging to the positive class.

Why do we need Logistic Regression?

Even when we have a linear regression **algorithm, why do we need another logistic regression algorithm?**

To answer this question first we need to understand the **problem behind the linear regression for the classification task.** Let's consider an example of predicting diabetes based on sugar levels. In this example, we have a dataset with two variables: sugar levels (independent variable) and the likelihood of having diabetes (dependent variable, 1 for diabetic and 0 for non-diabetic).

As observed, the linear regression line () fails to accurately capture the relationship between sugar levels and the likelihood of diabetes. It assumes a linear relationship, which is not suitable for this binary classification problem. **Extending this line will result in values that are less than 0 and greater than 1, which are not very useful in our classification problem.**

This is where logistic regression comes into play, using the sigmoid function to model the probability of an instance belonging to a particular class.

What is Cost Function?

A cost function is a mathematical function that calculates the difference between the target actual values (ground truth) and the values predicted by the model. A function that assesses a machine learning model's performance also referred to as a **loss function or objective function**. Usually, the objective of a machine learning algorithm is to reduce the error or output of cost function.

Why Mean Squared Error cannot be used as cost function for Logistic Regression

Let's consider the Mean Squared Error (MSE) as a cost function, but it is not suitable for logistic regression due to its nonlinearity introduced by the sigmoid function.

Log Loss for Logistic regression

Log loss is a classification evaluation metric that is used to compare different models which we build during the process of model development. It is considered one of the efficient metrics for evaluation purposes while dealing with the soft probabilities predicted by the model.

Conclusion

The choice of cost function, log loss or cross-entropy, is significant for logistic regression. It quantifies the disparity between predicted probabilities and actual outcomes, providing a measure of how well the model aligns with the ground truth.

Frequently Asked Questions (FAQs)

1. What is the difference between the cost function and the loss function for logistic regression?

The terms are often used interchangeably, but the cost function typically refers to the average loss over the entire dataset, while the loss function calculates the error for a single data point.

2.Why logistic regression cost function is non convex?

The sigmoid function introduces non-linearity, resulting in a non-convex cost function. It has multiple local minima, making optimization challenging, as traditional gradient descent may converge to suboptimal solutions.

3.What is a cost function in simple terms?

A cost function measures the disparity between predicted values and actual values in a machine learning model. It quantifies how well the model aligns with the ground truth, guiding optimization.

4.Is the cost function for logistic regression always negative?

No, the cost function for logistic regression is not always negative. It includes terms like -log(h(x)) and -log(1 – h(x)), and the overall value depends on the predicted probabilities and actual labels, yielding positive or negative values.

5.Why only sigmoid function is used in logistic regression?

The sigmoid function maps real-valued predictions to probabilities between 0 and 1, facilitating the interpretation of results as probabilities. Its range and smoothness make it suitable for binary classification, ensuring stable optimization.