**Logistic Regression in Machine Learning**

Logistic Regression is a supervised machine learning algorithm used for classification problems. Unlike linear regression which predicts continuous values it predicts the probability that an input belongs to a specific class. It is used for binary classification where the output can be one of two possible categories such as Yes/No, True/False or 0/1. It uses sigmoid function to convert inputs into a probability value between 0 and 1. In this article, we will see the basics of logistic regression and its core concepts.

**Types of Logistic Regression**

Logistic regression can be classified into three main types based on the nature of the dependent variable:

1. **Binomial Logistic Regression**: This type is used when the dependent variable has only two possible categories. Examples include Yes/No, Pass/Fail or 0/1. It is the most common form of logistic regression and is used for binary classification problems.
2. **Multinomial Logistic Regression**: This is used when the dependent variable has three or more possible categories that are not ordered. For example, classifying animals into categories like "cat," "dog" or "sheep." It extends the binary logistic regression to handle multiple classes.

**Assumptions of Logistic Regression**

Understanding the assumptions behind logistic regression is important to ensure the model is applied correctly, main assumptions are:

1. **Independent observations**: Each data point is assumed to be independent of the others means there should be no correlation or dependence between the input samples.
2. **Binary dependent variables**: It takes the assumption that the dependent variable must be binary, means 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 model assumes a linear relationship between the independent variables and the log odds of the dependent variable which means the predictors affect the log odds in a linear way.
4. **No outliers**: The dataset should not contain extreme outliers as they can distort the estimation of the logistic regression coefficients.
5. **Large sample size**: It requires a sufficiently large sample size to produce reliable and stable results.

**Understanding Sigmoid Function**

1. The sigmoid function is a important part of logistic regression which is used to convert the raw output of the model into a probability value between 0 and 1.

2. This function takes any real number and maps it into the range 0 to 1 forming an "S" shaped curve called the sigmoid curve or logistic curve. Because probabilities must lie between 0 and 1, the sigmoid function is perfect for this purpose.

3. In logistic regression, we use a threshold value usually 0.5 to decide the class label.

* If the sigmoid output is same or above the threshold, the input is classified as Class 1.
* If it is below the threshold, the input is classified as Class 0.

This approach helps to transform continuous input values into meaningful class predictions.

**How does Logistic Regression work?**

Logistic regression model transforms the [linear regression](https://www.geeksforgeeks.org/ml-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.

we use the [sigmoid function](https://www.geeksforgeeks.org/derivative-of-the-sigmoid-function/) where the input will be z and we find the probability between 0 and 1. i.e. predicted y.

*σ*(*z*)=1/1+*e*−*z*



As shown above the sigmoid function converts the continuous variable data into the probability i.e between 0 and 1.

**Terminologies involved in Logistic Regression**

Here are some common terms involved in logistic regression:

1. **Independent Variables:**These are the input features or predictor variables used to make predictions about the dependent variable.
2. **Dependent Variable**: This is the target variable that we aim to predict. In logistic regression, the dependent variable is categorical.
3. **Logistic Function**: This function transforms the independent variables into a probability between 0 and 1 which represents the likelihood that the dependent variable is either 0 or 1.
4. **Odds**: This is the ratio of the probability of an event happening to the probability of it not happening. It differs from probability because probability is the ratio of occurrences to total possibilities.
5. **Log-Odds (Logit)**: The natural logarithm of the odds. In logistic regression, the log-odds are modeled as a linear combination of the independent variables and the intercept.
6. **Coefficient**: These are the parameters estimated by the logistic regression model which shows how strongly the independent variables affect the dependent variable.
7. **Intercept**: The constant term in the logistic regression model which represents the log-odds when all independent variables are equal to zero.

**How to Evaluate Logistic Regression Model?**

Evaluating the logistic regression model helps assess its performance and ensure it generalizes well to new, unseen data. The following metrics are commonly used:

1. **Accuracy:** Accuracy provides the proportion of correctly classified instances.  
   Accuracy=TruePositives+TrueNegatives/Total
2. **Precision:** Precision focuses on the accuracy of positive predictions.  
   Precision=True Positives/TruePositives+FalsePositives
3. **Recall (Sensitivity or True Positive Rate):** Recall measures the proportion of correctly predicted positive instances among all actual positive instances.  
   Recall=TruePositives/TruePositives+FalseNegatives

**F1 Score:**F1 score is the harmonic mean of precision and recall.  
F1Score=2∗Precision∗Recall/Precision+Recall

**Differences Between Linear and Logistic Regression**

Logistic regression and linear regression differ in their application and output. Here's a comparison:

| **Linear Regression** | **Logistic Regression** |
| --- | --- |
| Linear regression is used to predict the continuous dependent variable using a given set of independent variables. | Logistic regression is used to predict the categorical dependent variable using a given set of independent variables. |
| It is used for solving regression problem. | It is used for solving classification problems. |
| In this we predict the value of continuous variables | In this we predict values of categorical variables |
| In this we find best fit line. | In this we find S-Curve. |
| Least square estimation method is used for estimation of accuracy. | Maximum likelihood estimation method is used for Estimation of accuracy. |
| The output must be continuous value, such as price, age etc. | Output must be categorical value such as 0 or 1, Yes or no etc. |
| It required linear relationship between dependent and independent variables. | It not required linear relationship. |

**Advantages of Logistic Regression**

1. **Simple and Easy to Implement**
   * It’s intuitive and easy to implement compared to more complex models.
2. **Computationally Efficient**
   * Logistic regression is not resource-intensive and works well with smaller datasets.
3. **Works Well with Linearly Separable Classes**
   * Very effective when the relationship between the independent variables and the log-odds of the dependent variable is linear.
4. **Probability Estimates**
   * Provides class probabilities (between 0 and 1), which is useful in risk assessment and ranking.
5. **No Need for Feature Scaling**
   * Not heavily affected by feature scaling (though regularization might require it).
6. **Interpretable Model**
   * Coefficients are easy to understand, making it a good choice when model explainability is important.

**Disadvantages of Logistic Regression**

1. **Assumes Linearity in the Logit**
   * Assumes a linear relationship between independent variables and the log odds, which may not always hold.
2. **Not Suitable for Complex Relationships**
   * Can't capture non-linear relationships unless transformed features or polynomial terms are added.
3. **Sensitive to Irrelevant Features**
   * Including too many irrelevant or correlated features can reduce model performance.
4. **Limited to Binary or Multinomial Outputs**
   * Basic logistic regression is for binary classification; multinomial logistic regression is more complex.
5. **Prone to Overfitting**
   * Especially if the number of features is much larger than the number of observations (high-dimensional data).
6. **Cannot Handle Missing Values**
   * Logistic regression doesn’t natively handle missing data.

**Applications of Logistic Regression**

1. **Medical Diagnosis**
   * Predicting whether a patient has a disease (e.g., cancer, diabetes) based on symptoms or test results.
2. **Credit Scoring**
   * Assessing whether a person is likely to default on a loan (good/bad credit risk).
3. **Marketing and Customer Retention**
   * Predicting whether a customer will buy a product or churn.
4. **Political Forecasting**
   * Predicting election outcomes (win/loss) based on campaign data.
5. **Fraud Detection**
   * Determining whether a transaction is fraudulent or not.
6. **Social Science Research**
   * Analysing factors that influence binary outcomes like employment status, voting behavior, etc.