Logistic Regression in Machine Learning

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.

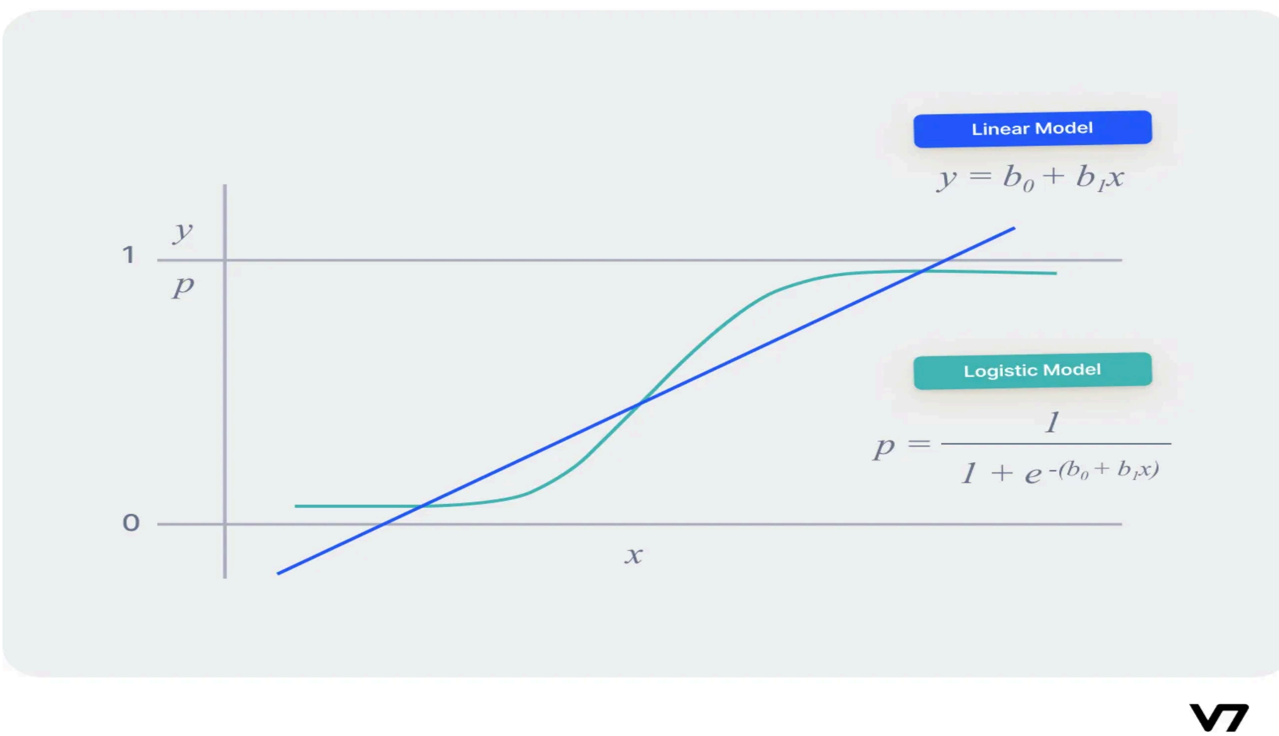
For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used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

1. Data

2. Compute Predictions

A. Compute the Linear Function

The model calculates a **linear combination** of the input features (X), weights (w), and bias (b):

$$Z = w_1X_1 + w_2X_2 + ... + w_nX_n + b$$

B. Apply the Sigmoid Function

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

This function outputs values between **0 and 1**, representing the probability that the sample belongs to a particular class.

3. Compute the Cost Function

Logistic Regression uses the **Binary Cross-Entropy Loss (Log Loss)** to measure how well the model is performing:

$$J(w, b) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where:

- y_i is the actual label (0 or 1).
- $\hat{y}_i$ is the predicted probability.
- m is the number of samples.

4. Optimize Model Parameters (Gradient Descent)

To minimize the cost function, **Gradient Descent** is used to update the weights:

$$w_j = w_j - \alpha \frac{\partial J}{\partial w_j}$$

where:

- α is the learning rate.
- $\frac{\partial J}{\partial w_j}$ is the gradient of the cost function with respect to w_j.

5. Make Predictions

After training, the model predicts class labels by applying a threshold (usually 0.5):

$$y_{\text{pred}} = \begin{cases} 1, & \text{if } \sigma(Z) \geq 0.5 \\ 0, & \text{if } \sigma(Z) < 0.5 \end{cases}$$

How to Evaluate Logistic Regression Model?

- **Accuracy**: Accuracy provides the proportion of correctly classified instances.
- **Precision**: Precision focuses on the accuracy of positive predictions.

$$\frac{TP}{TP + FP} = \text{Precision}$$

- **Recall** (Sensitivity or True Positive Rate): Recall measures the
- proportion of correctly predicted positive instances among all actual positive instances.

$$\frac{TP}{TP + FN} = \text{Recall}$$

-
- **F1 Score**: F1 score is the harmonic mean of precision and recall.

- $$\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 = F1$$
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: The ROC curve plots the true positive rate against the false positive rate at various thresholds. AUC-ROC measures the area under this curve, providing an aggregate measure of a model's performance across different classification thresholds.

Code Example

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Example dataset
X = [[2.5], [3.0], [3.5], [4.0], [4.5]]  # Study hours
y = [0, 0, 1, 1, 1]  # Pass (1) or Fail (0)

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train model
model = LogisticRegression()
model.fit(X_train, y_train)

# Predict
y_pred = model.predict(X_test)
print("Accuracy:", accuracy_score(y_test, y_pred))
```

Limitations of Logistic Regression

- Linear Decision Boundary: It assumes a linear relationship between features and the log-odds of the target.
- Overfitting: It can overfit if there are too many features.
- Not Suitable for Complex Relationships: It may not perform well on non-linear data.

for learn more

[link]
[link]