LOGISTIC REGRESSION

Logistic regression is a supervised learning algorithm that makes use of logistic functions to predict the probability of a binary outcome

- Logistic regression is defined as a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.

What Is Logistic Regression?

- Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation.
- The model delivers a binary outcome limited to two possible outcomes: yes/no, 0/1, or true/false.
- Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes.
- It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.
- For example, 0 – represents a negative class; 1 – represents a positive class.
- Logistic regression is commonly used in binary classification problems where the outcome variable reveals either of the two categories (0 and 1).
- Some examples of such classifications and instances where the binary response is expected or implied are:
- 1. Determine the probability of heart attacks: With the help of a logistic model, medical practitioners can determine the relationship between variables such as the weight, exercise, etc
- ., of an individual and use it to predict whether the person will suffer from a heart attack or any other medical complication.
- 2. Possibility of enrolling into a university: Application aggregators can determine the probability of a student getting accepted to a particular university or a degree course in a college by studying the relationship between the estimator variables, such as GRE, GMAT, or TOEFL scores.
- 3. Identifying spam emails: Email inboxes are filtered to determine if the email communication is promotional/spam by understanding the predictor variables and applying a logistic regression algorithm to check its authenticity.

Key advantages of logistic regression



1. Easier to implement machine learning methods: A machine learning model can be effectively set up with the help of training and testing. The training identifies patterns in the input data (image) and associates them with some form of output (label). Training a logistic model with a regression algorithm does not demand higher computational power. As such, logistic regression is easier to implement, interpret, and train than other ML methods.
2. Suitable for linearly separable datasets: A linearly separable dataset refers to a graph where a straight line separates the two data classes. In logistic regression, the y variable takes only two values. Hence, one can effectively classify data into two separate classes if linearly separable data is used.
3. Provides valuable insights: Logistic regression measures how relevant or appropriate an independent/predictor variable is (coefficient size) and also reveals the direction of their relationship or association (positive or negative).

Types of Logistic Regression with Examples

1. ***Binary logistic regression***

- Binary logistic regression predicts the relationship between the independent and binary dependent variables. Some examples of the output of this regression type may be, success/failure, 0/1, or true/false.
- Examples:
- Deciding on whether or not to offer a loan to a bank customer: Outcome = yes or no.
- Evaluating the risk of cancer: Outcome = high or low.
- Predicting a team's win in a football match: Outcome = yes or no.

2. ***Multinomial logistic regression***

- A categorical dependent variable has two or more discrete outcomes in a multinomial regression type. This implies that this regression type has more than two possible outcomes.
- Examples:
- Let's say you want to predict the most popular transportation type for 2040. Here, transport type equates to the dependent variable, and the possible outcomes can be electric cars, electric trains, electric buses, and electric bikes.
- Predicting whether a student will join a college, vocational/trade school, or corporate industry.
- Estimating the type of food consumed by pets, the outcome may be wet food, dry food, or junk food.

3. ***Ordinal logistic regression***

- Ordinal logistic regression applies when the dependent variable is in an ordered state (i.e., ordinal). The dependent variable (y) specifies an order with two or more categories or levels.
-
- Examples: Dependent variables represent,
- Formal shirt size: Outcomes = XS/S/M/L/XL
- Survey answers: Outcomes = Agree/Disagree/Unsure
- Scores on a math test: Outcomes = Poor/Average/Good

LOGISTIC REGRESSION FOR BINARY CLASSIFICATION

Logistic regression is one of the most widely used algorithms for binary classification, where the output is restricted to two categories (e.g., Yes/No, 0/1, True/False). It predicts probabilities using the sigmoid function and then applies a threshold (commonly 0.5) to assign class labels.

How it works

- Logistic regression is a supervised machine learning algorithm used for binary classification, where the target variable has two possible outcomes (for example: yes/no, spam/not spam, 0/1).
- Unlike linear regression, which predicts a continuous value, logistic regression predicts the probability that an input belongs to the positive class.
- The predicted probability is always between 0 and 1 because it uses the sigmoid (logistic) function:
- Thresholding: If probability ≥ 0.5 → Class 1; else → Class 0.

$$P(y=1|x) = \frac{1}{1+e^{-z}}$$

where

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- β_0: intercept (bias)
- $\beta_1, \beta_2, \ldots$: feature coefficients
- $x_1, x_2, \ldots$: input features

Workflow of Logistic Regression for Binary Classification

Step 1: Collect the Dataset
- Gather a dataset containing:
- Input features (X): Independent variables (e.g., age, income, study hours).
- Target variable (Y): Binary outcome (0 or 1).
- Example:
-

Study Hours	Pass
2	0
4	0
6	1
8	1

Step 2: Preprocess the Data

- Prepare the data before training:
- Handle missing values.
- Encode categorical variables.
- Scale/normalize features (optional but often beneficial).
- Split the dataset into training and testing sets (e.g., 80% training, 20% testing).

Step 3: Train the Logistic Regression Model

- The model learns the relationship between input features and the binary target.
- First, it computes a linear combination:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where:

- β_0 = intercept
- β_i = coefficients
- x_i = input features

Step 4: Apply the Sigmoid Function

- The linear output is converted into a probability between 0 and 1 using the sigmoid function:

- For example:
- $z=2.5 \rightarrow$ Probability = 0.924
- $z=-1.5 \rightarrow$ Probability = 0.182

$$P(y=1|x) = \frac{1}{1+e^{-z}}$$

where

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- β_0: intercept (bias)
- $\beta_1, \beta_2, \ldots$: feature coefficients
- $x_1, x_2, \ldots$: input features

Step 5: Apply a Decision Threshold

- The predicted probability is compared with a threshold (usually 0.5) to assign the final class.
- Decision rule:
- If Probability ≥ 0.5 → Class = 1
- If Probability < 0.5 → Class = 0
- Example:
-

Probability	Predicted Class
0.91	1
0.73	1
0.42	0
0.15	0

Step 6: Optimize the Model

- The model adjusts its coefficients to minimize the binary cross-entropy (log loss).
- Optimization is typically performed using:
-
- Gradient Descent
- Stochastic Gradient Descent (SGD)
- Advanced optimizers such as L-BFGS

Step 7: Evaluate the Model

- Test the trained model on unseen data using evaluation metrics such as:
  - Accuracy
  - Precision
  - Recall
  - F1-score
  - ROC-AUC
- A confusion matrix summarizes the predictions:

Actual \ Predicted	0	1
0	True Negative (TN)	False Positive (FP)
1	False Negative (FN)	True Positive (TP)

Step 8: Make Predictions

- Use the trained model to classify new data.
- Example:
- Study hours = 7
- Predicted probability = 0.88
- Since 0.88 > 0.5, the predicted class is Pass (1).