Logistic Regression

-MUKESH KUMAR

- Logistic regression is a popular statistical model used for binary classification tasks, where the goal is to predict one of two possible outcomes.

- Despite its name, it's a regression model in name only; in practice, it performs classification.

What is Logistic Regression?

- **Logistic Regression** is a **supervised learning algorithm** used for **binary classification** problems, meaning it predicts whether an input belongs to one of two categories (e.g., spam or not spam, disease or no disease).

- For example, it might predict the probability that a patient has a particular disease (positive class) versus not having it (negative class).

How Logistic Regression Works

- Compute **Linear Combination** → $z=\beta_0+\beta_1X_1+\ldots+\beta_nX_nz$
- Apply **Sigmoid Function** → $\sigma(z)$converts output into a probability (0 to 1).
- Use **Threshold (0.5)** to classify into **Class 0 or Class 1.**
- Optimize coefficients using **Gradient Descent** to minimize **Log Loss.**

The Logistic Function
(Sigmoid Function)

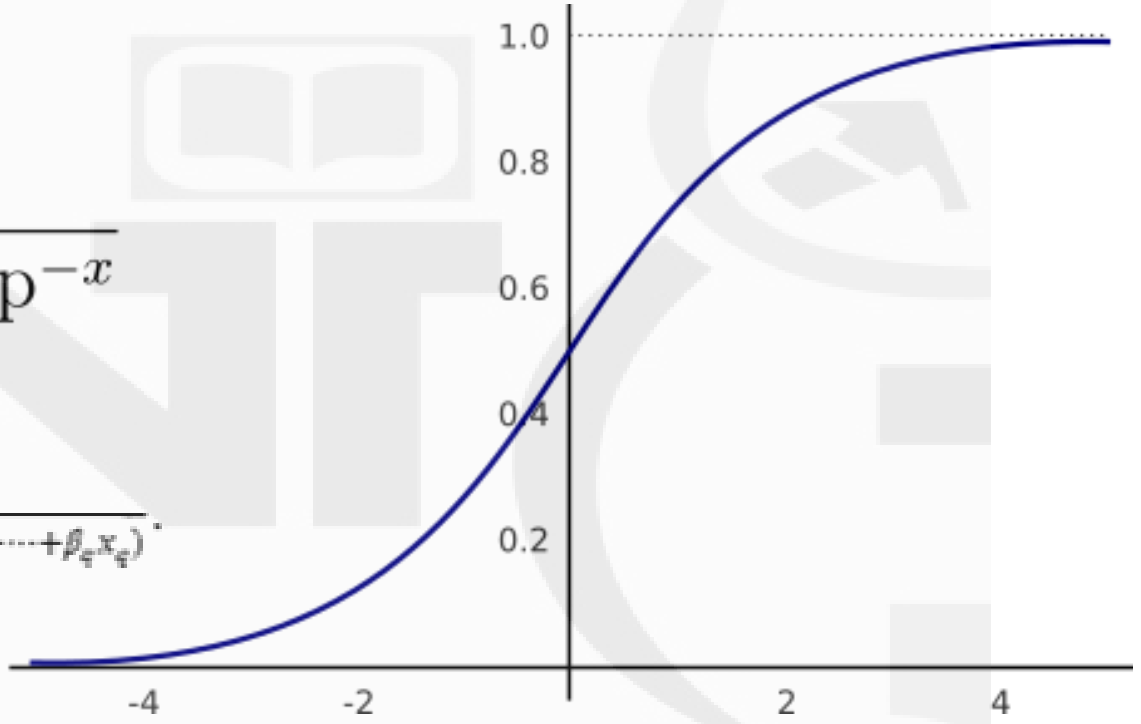
- The core of logistic regression is the logistic function, also known as the sigmoid function.

- The sigmoid function maps any real-valued number into a value between 0 and 1, which is useful for modeling probabilities.

Logistic/Sigmoid Function

$$\sigma(x) = \frac{1}{1 + \exp^{-x}}$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q)}}.$$



- https://www.transum.org/Maths/Activity/Graph/Desmos.asp

How Logistic Regression Works

Step 1: Linear Combination of Inputs:

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

Here, β_0 is the intercept (bias term), $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients, and $x_1, x_2, \ldots, x_n$ are the input features.

Step 2: Apply the Sigmoid Function

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

The output $\hat{y}$ represents the probability that the input belongs to the positive class (e.g., class 1).

Advantages & Disadvantages

✅ Advantages

- Simple, fast, and easy to interpret.
- Works well when features are **linearly separable**.
- Provides **probabilistic interpretation** of predictions.
- Less prone to **overfitting** compared to complex models.

❌ Disadvantages

- Assumes **linear decision boundary**, which may not hold for complex data.
- Not suitable for **multi-class problems** (requires One-vs-All for multi-class).
- Sensitive to **outliers** and **multicollinearity**.

Applications

- **Medical Diagnosis** – Used to predict whether a patient has a particular disease based on symptoms and test results (e.g., predicting diabetes, heart disease, or cancer).
- **Customer Churn Prediction** – Helps businesses determine whether a customer is likely to stop using a service based on behavioral patterns and historical data.
- **Spam Detection** – Classifies emails as spam or not spam based on features like keywords, sender reputation, and frequency of certain words.
- **Credit Scoring & Loan Approval** – Banks use logistic regression to assess whether a borrower is likely to default on a loan based on income, credit history, and other factors.