

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}$$

classification model

odd

Introduction to Logistic Regression

fundamental model

0-1 classification

1. Introduction to Logistic Regression

Logistic Regression is a fundamental algorithm used for binary classification problems. Despite the name, it is a classification model that predicts the probability that a given input belongs to a particular category.

2. Mathematical Foundation

The model estimates the probability $P(Y = 1 | X)$ using the logistic (sigmoid) function:

$$P(Y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}}$$

The predicted output is interpreted as a probability, and a threshold (typically 0.5) is used to classify into class 0 or 1.

3. Applications in Data Science

Logistic regression is widely used in various domains, including:

- ✓ • **Healthcare:** Predicting disease presence (e.g., diabetes, cancer).
- ✓ • **Marketing:** Customer churn prediction.
- ✓ • **Finance:** Credit risk modeling, fraud detection.
- ✓ • **Human Resources:** Predicting employee attrition.
- ✓ • **Email filtering:** Spam vs. non-spam classification.

4. Pros of Logistic Regression

- **Simple and interpretable:** Coefficients represent feature importance.
- **Fast to train:** Suitable for large datasets.

- Works well with linearly separable data.
- Probabilistic output: Useful for decision-making under uncertainty.

5. Cons of Logistic Regression

- Assumes linearity: Can underperform when the true relationship is non-linear.
- Not ideal for high-dimensional sparse data.
- Sensitive to outliers and multicollinearity.
- Can be limited in capturing complex patterns compared to tree-based models or neural networks.

6. When to Use Logistic Regression

- When your target variable is binary or categorical.
- When you need a fast, interpretable baseline model.
- When the relationship between features and log-odds is roughly linear.

7. Summary

Logistic regression is an essential tool for any data scientist. It's simple, interpretable, and often performs surprisingly well on real-world binary classification problems. Understanding its assumptions, strengths, and limitations is crucial for building better models.