Logistic Regression: Theory and Case Study

August 30, 2024

Introduction to Logistic Regression

- Logistic regression is a statistical method for modeling binary outcomes (0 or 1, true or false, success or failure).
- It is widely used in classification problems where the goal is to predict the probability of a certain class or event.
- Common applications include medical diagnosis, spam detection, credit scoring, and marketing.

Logistic Regression Model

Model Equation:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \quad (1)$$

- p: Probability of the positive class (e.g., success)
- x_i: Independent variables (features)
- β_0: Intercept term
- β_i: Coefficients (weights)

Sigmoid Function

- The logistic regression model outputs probabilities using the sigmoid function:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)}} \tag{2}$$

- The sigmoid function maps any real-valued number into a value between 0 and 1.

Sigmoid function

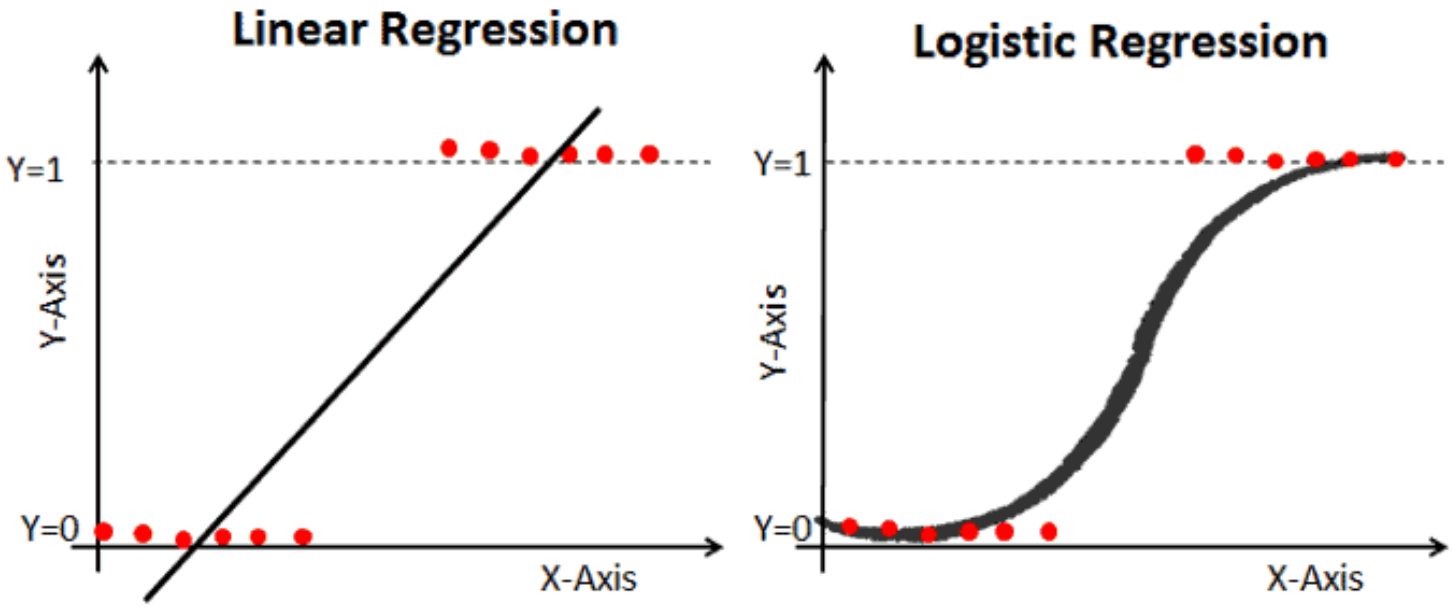

Figure: Sigmoid function curve

Cost Function and Maximum Likelihood Estimation (MLE)

Cost Function for Logistic Regression:

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \log(h_\beta(x_i)) + (1 - y_i)\log(1 - h_\beta(x_i))] \tag{3}$$

where:

- $h_\beta(x_i) = \frac{1}{1+e^{-\beta^T x_i}}$
- y_i: True label (0 or 1)
- x_i: Feature vector for the i-th example
- m: Number of training examples

Case Study: Predicting Customer Purchase

Dataset: Customer Demographics and Purchase History

- Features: Age, income, gender, browsing behavior, etc.
- Target: Whether a customer will purchase a product (1: Purchase, 0: No Purchase)

Objective: Predict the likelihood of a customer making a purchase based on their demographic and behavioral data.

Model Training and Evaluation

Training the Model:

- Split data into training and test sets
- Fit the logistic regression model using training data

Evaluation Metrics:

- Accuracy: Proportion of correctly predicted outcomes
- Confusion Matrix: Table summarizing true vs. predicted classifications
- Precision, Recall, F1 Score: Metrics to evaluate model performance on imbalanced datasets

Confusion Matrix and Classification Report

Confusion Matrix:

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

Classification Metrics:

- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- F1 Score $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Conclusion

- Logistic regression is a powerful and interpretable model for binary classification problems.
- Important to understand assumptions and limitations, especially with non-linear relationships.
- Case study demonstrated logistic regression in predicting customer purchase behavior.
- Future work could include using more advanced models or techniques like regularization to improve performance.