Logistic Regression: Theory, Loss Functions, and Optimization

Naman Goel

October 9, 2025

Abstract

This report provides a comprehensive overview of logistic regression, a fundamental algorithm used for binary classification tasks in machine learning. Core components covered include the mathematical formulation, the sigmoid activation function, the cross-entropy loss function, and parameter estimation via gradient descent. Illustrative figures are included to enhance understanding.

1 Introduction

Logistic regression is a widely used classification algorithm that models the probability of a binary outcome as a function of input features using the logistic function [1]. Unlike linear regression, which predicts continuous values, logistic regression outputs probabilities bounded between 0 and 1, suitable for classification tasks such as spam detection, medical diagnosis, and many others.

2 Mathematical Model

Given an input vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, logistic regression models the probability of a positive class ($y = 1$) as:

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

where

$$z = \mathbf{w}^\top \mathbf{x} + b$$

with $\mathbf{w}$ as the weights, b as the bias term, and $\sigma(\cdot)$ the sigmoid function [2].

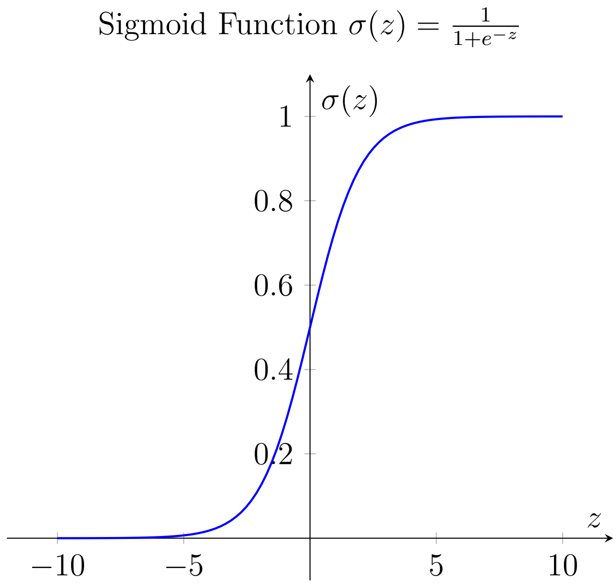

Figure 1: The sigmoid function maps any real-valued number into the interval (0,1), enabling probability interpretation.

3 Cross-Entropy Loss Function

To quantify the difference between predicted probabilities and true class labels, logistic regression uses the cross-entropy (log loss) function:

$$L = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \right]$$

where m is the number of samples, $y^{(i)} \in \{0, 1\}$ is the true label, and $\hat{y}^{(i)} = \sigma(z^{(i)})$ is the predicted probability [?].

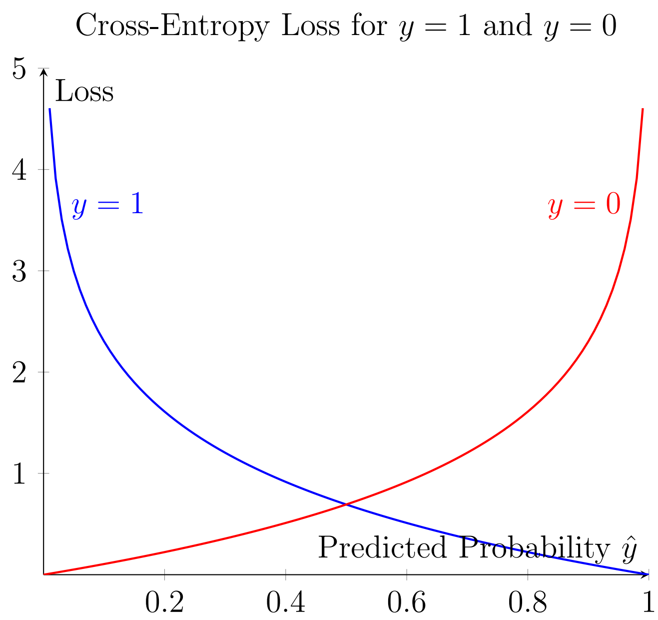

Figure 2: Cross-entropy loss curves illustrating penalty as predicted probability diverges from true label for positive (blue) and negative (red) classes.

4 Gradient Descent Optimization

To find optimal parameters $\mathbf{w}$ and b minimizing the cross-entropy loss, gradient descent is widely used. The update rules for parameters at iteration t are:

$$w_j^{t+1} = w_j^t - \alpha \frac{\partial L}{\partial w_j}, \quad b^{t+1} = b^t - \alpha \frac{\partial L}{\partial b}$$

where α is the learning rate, and gradients are computed as

$$\frac{\partial L}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) x_j^{(i)}, \quad \frac{\partial L}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)$$

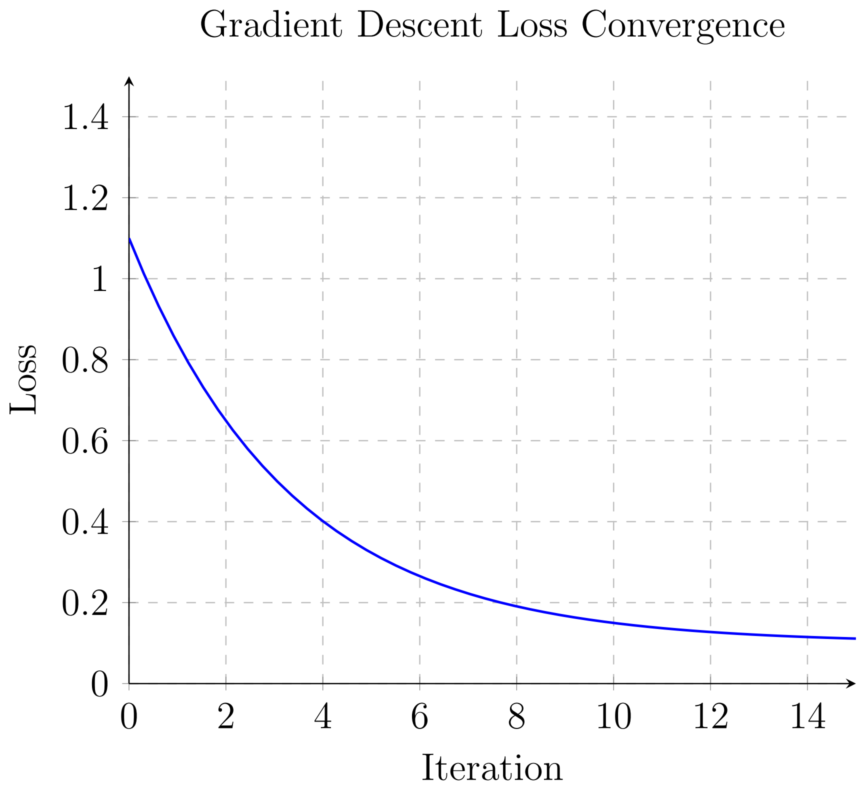

Figure 3: An example curve of loss decreasing over iterations of gradient descent optimization.

5 Conclusion

Logistic regression forms a cornerstone of binary classification tasks in machine learning with its probabilistic model and interpretable parameters. The sigmoid function enables probability output, the cross-entropy loss measures prediction error effectively, and gradient descent provides a practical optimization technique. Understanding these components is essential for applying logistic regression confidently.

References

[1] Logistic Regression. *Wikipedia*. `https://en.wikipedia.org/wiki/Logistic_regression`

[2] Andrew Ng, Machine Learning, Coursera Lecture Notes.