Assignment 3: Logistic Regression

By: Vani Seth

Project Overview

This document presents the results of a binary classification task using a custom-built logistic regression algorithm. The goal was to classify two species of the Iris flower, Iris-setosa and Iris-versicolor, based on their physical features. The model was trained on 80% of the data and evaluated on the remaining 20%.

Data Visualization

Figure 1 shows the scatter plot with the initial distribution of the two classes based on petal length and petal width. This visualization helps confirm that the classes are linearly separable, making them a good candidate for logistic regression.

Figure 1: Petal Length vs. Petal Width for Iris-setosa and Iris-versicolor

Model Training and Performance

The logistic regression model was trained using gradient descent to minimize the binary cross-entropy loss.

Hyperparameters:

- Learning Rate: 0.1
- Number of Iterations: 2000

Figure 2 shows a plot of the cost function decreasing over the 2000 iterations, which indicates that the model was successfully learning from the training data.

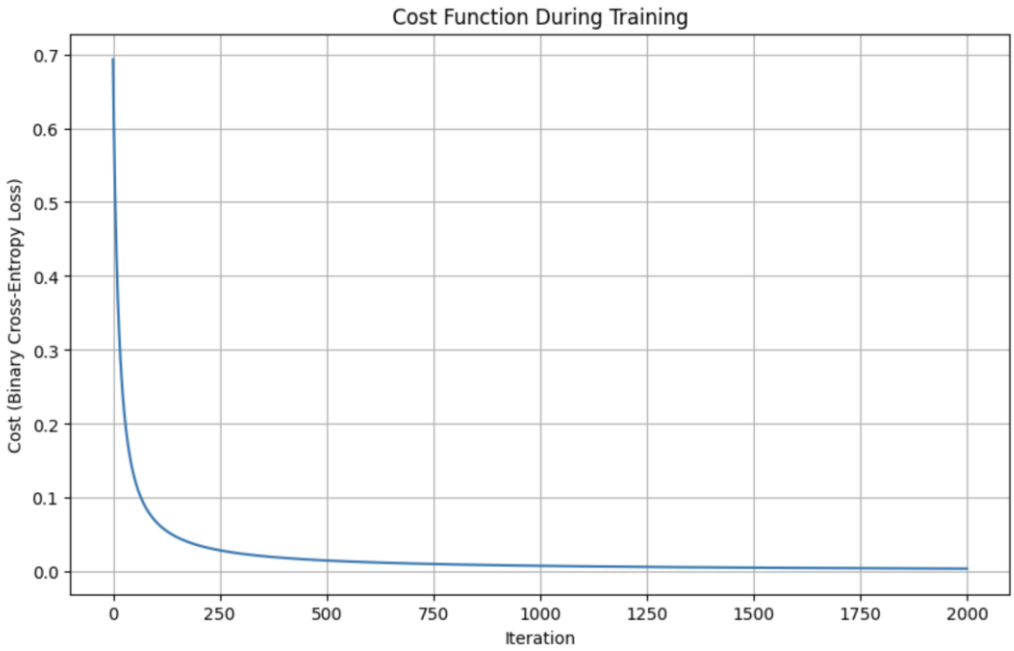

Figure 2: Cost Function During Training

After training, the model learned the following parameters:

- Learned Weights: [-0.63416777 -2.31815188 3.53682415 1.58267453]
- Learned Bias: -0.427972572931216

The weights tell the model how important each of the four flower features is for deciding. Each number in the list corresponds to one of the features (in order: sepal length, sepal width, petal length, and petal width).

- **Positive Weight:** If a feature has a positive weight, a larger value for that feature adds points to the score, making the model lean towards predicting Iris-versicolor (class 1).
- **Negative Weight:** If a feature has a negative weight, a larger value for that feature subtracts points, making the model lean towards Iris-setosa (class 0).
- **Size of the Weight:** The bigger the number (positive or negative), the more influence that feature has.

The bias is a starting score, before the model even looks at the flower's features, it starts with a score of -0.427. This means the model has a very slight initial bias toward guessing Iris-setosa (class 0) if it doesn't have any other information.

Execution Results on Test Data

The trained model was evaluated on the unseen test dataset to measure its performance.

Classification Accuracy: The primary metric for evaluation was classification accuracy, which measures the percentage of correctly predicted labels.

- Classification Accuracy on Test Data: 100%

Predictions vs. Actuals: The following shows a comparison of the model's predictions against the true labels for the test data:

- Predicted Labels: [1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1]
- Actual Labels: [1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 1]

Decision Boundary Visualization

The decision boundary learned by the model is visualized in **Figure 3**. The plot shows the regions where the model predicts each class, overlaid with the actual data points from the test set. This demonstrates how effectively the model separates the two classes.

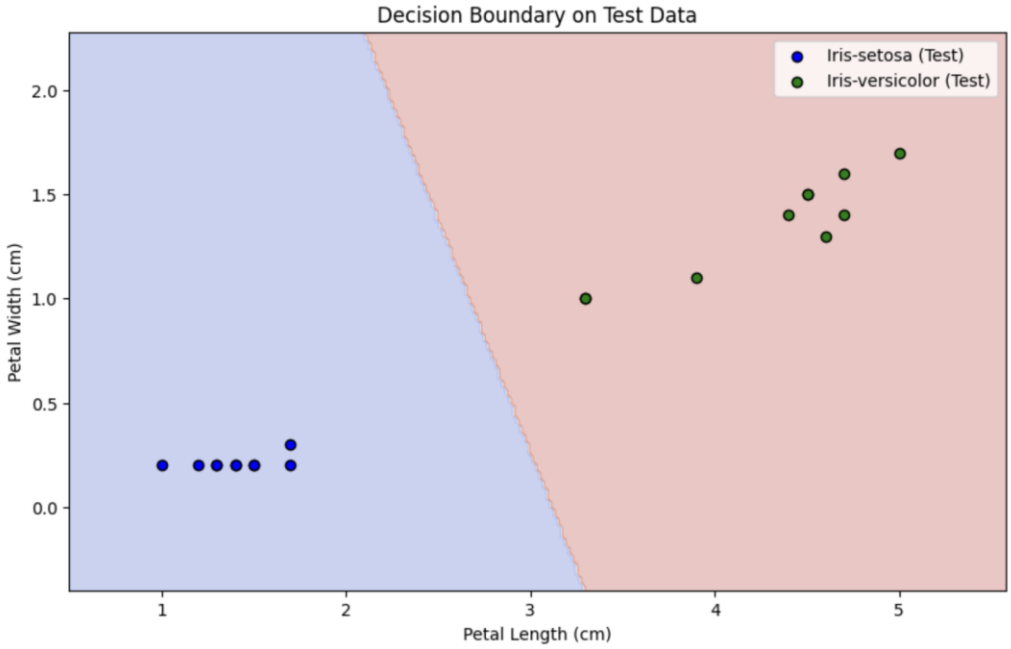

Figure 3: Decision Boundary on Test Data

Conclusion

We implemented a logistic regression model from scratch that is successfully learned to distinguish between Iris-setosa and Iris-versicolor with high accuracy on the test data. The convergence of the cost function and the clear separation shown by the decision boundary confirm the model's effectiveness for this binary classification problem.