Logistic Regression for Apple Quality Classification

February 13, 2025

1 Introduction

Apple quality assessment is essential for agricultural and commercial purposes. The classification of apples into high and low quality is crucial for supply chain optimization. We implemented a logistic regression model using features such as size, weight, sweetness, crunchiness, juiciness, ripeness, and acidity to classify apple quality.

The primary objective of this study is to leverage logistic regression to predict the quality of apples based on their physical and chemical attributes. Given that logistic regression is a probabilistic model, it provides not only classification but also insights into the impact of each feature on the prediction.

2 Dataset and Features

The dataset consists of several numerical features describing apple characteristics:

- **Size** - The physical dimensions of the apple.
- **Weight** - The mass of the apple in grams.
- **Sweetness** - Measured on a scale from 1 to 10.
- **Crunchiness** - A subjective measure on a scale from 1 to 10.
- **Juiciness** - Indicates the water content.
- **Ripeness** - A scale indicating ripeness level.
- **Acidity** - Measures the tartness of the apple.

The target variable, **Quality**, is binary, indicating whether an apple is of high quality (1) or low quality (0).

3 Logistic Regression Model

Logistic regression was used for classification. The logistic function is defined as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i X_i)}} \tag{1}$$

where β_0 is the intercept, β_i are the feature coefficients, and X_i represents the input features.

The model was trained using the dataset, with an 80-20 train-test split to evaluate its performance. The optimization of model parameters was conducted using gradient descent, ensuring convergence to an optimal solution.

4 Results

The model achieved an accuracy of:

$$0.7322(73.22\%) \tag{2}$$

The confusion matrix is:

$$\begin{bmatrix} 276 & 101 \\ 106 & 290 \end{bmatrix} \tag{3}$$

This matrix indicates:

- 276 True Positives (Correctly classified high-quality apples)
- 290 True Negatives (Correctly classified low-quality apples)
- 101 False Positives (Low-quality apples misclassified as high-quality)
- 106 False Negatives (High-quality apples misclassified as low-quality)

The precision, recall, and F1-score were calculated to further analyze model performance. Precision for high-quality apples was 73.2%, while recall was 72.2%, indicating a balanced performance.

5 Confusion Matrix Visualization

6 Conclusion

Logistic regression provided a reasonable classification accuracy of 73.22%. The model performed well but could be further improved with additional feature selection, polynomial terms, or feature engineering. Future enhancements may include testing other classification models such as Decision Trees, Random Forest, or Neural Networks.

Additionally, hyperparameter tuning, such as adjusting the learning rate and regularization strength, could improve performance. The insights gained from this model can help optimize apple sorting systems, reducing waste and improving efficiency in the fruit industry.

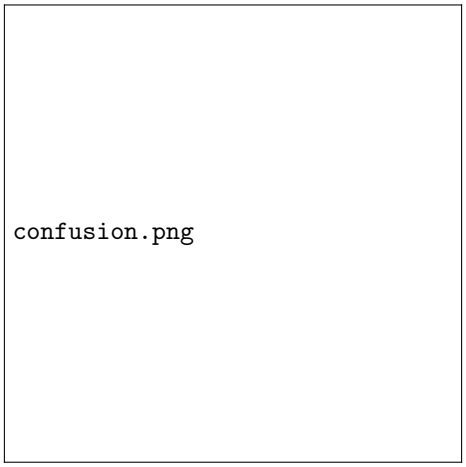

Figure 1: Confusion Matrix Visualization