Wine Classification Using Multi-Class Logistic Regression

Introduction

Wine classification is a significant application in machine learning, enabling the categorization of wines based on their characteristics. This project focuses on classifying wines into three distinct classes: Class 1 (non-alcoholic), Class 2 (non-alcoholic), and Class 3 (alcoholic). By leveraging Multi-Class Logistic Regression, a supervised learning algorithm, the model aims to predict the class of a wine sample based on its features, such as chemical composition or sensory attributes. This project demonstrates the application of AI and ML techniques in addressing real-world classification problems.

Problem Statement

The challenge is to accurately classify wines into three categories—two non-alcoholic (Classes 1 and 2) and one alcoholic (Class 3)—based on a dataset of wine features. The dataset may include attributes like acidity, sugar content, pH, or alcohol percentage. Misclassification can lead to incorrect labelling, impacting quality control or consumer trust. The goal is to develop a robust Multi-Class Logistic Regression model that minimizes classification errors and achieves high accuracy across all classes.

Algorithm

Multi-Class Logistic Regression: This algorithm extends binary logistic regression to handle multiple classes using the "one-vs-rest" (OvR) or "softmax" approach. In this project, the softmax regression is used to compute probabilities for each class (1, 2, and 3) and assign the wine sample to the class with the highest probability. ~~The model optimizes the cost function using gradient descent to find the best-fitting parameters for classification.~~

Objective

* To develop a Multi-Class Logistic Regression model for classifying wines into three categories: Class 1 (non-alcoholic), Class 2 (non-alcoholic), and Class 3 (alcoholic).
* To achieve high classification accuracy and minimize misclassification errors.
* To evaluate the model’s performance using metrics like accuracy, precision, recall, and F1-score.

Motivation

Wine classification has practical applications in the beverage industry, including quality assurance, product labelling, and consumer safety. Non-alcoholic and alcoholic wines have distinct markets, and accurate classification ensures compliance with regulations and consumer expectations. this project provides hands-on experience with a real-world dataset, multi-class classification techniques, and model evaluation, enhancing my understanding of supervised learning algorithms.

Methodology

* Data Collection: Obtain a dataset containing wine samples with features (e.g., acidity, sugar content, alcohol percentage) and corresponding class labels (1, 2, or 3).
* Data Preprocessing: Clean the dataset by handling missing values, normalizing/ standardizing features, and splitting it into training (80%) and testing (20%) sets.
* Model Training: Implement Multi-Class Logistic Regression using the softmax function to predict class probabilities. ~~Train the model on the training dataset using gradient descent.~~
* Model Evaluation: Assess the model’s performance on the test set using metrics like accuracy, precision, recall, and F1-score. Generate a confusion matrix to analyze class-wise performance.
* Visualization: Plot decision boundaries, confusion matrix, and feature importance to interpret the model’s behavior.
* ~~Optimization: Fine-tune hyperparameters (e.g., learning rate, regularization strength) to improve model performance.~~

Flow Chart

Block Diagram

Glimpse

This project implements a Multi-Class Logistic Regression model to classify wines into three classes based on their features. The model processes a dataset, trains on normalized features, and predicts whether a wine belongs to Class 1 (non-alcoholic), Class 2 (non-alcoholic), or Class 3 (alcoholic). The results are visualized through confusion matrices and decision boundaries, providing insights into the model’s performance and feature importance.

Flow of the Program

* Import Libraries: Use Python libraries like NumPy, Pandas, Scikit-learn, and Matplotlib for data handling, model implementation, and visualization.
* Load Dataset: Read the wine dataset (e.g., CSV file) into a Pandas DataFrame.
* Preprocess Data: Handle missing values, encode labels, and normalize features.
* Split Data: Divide the dataset into training and testing sets.
* Train Model: Fit the Multi-Class Logistic Regression model using Scikit-learn’s LogisticRegression(multi\_class='multinomial').
* Predict and Evaluate: Generate predictions on the test set and compute evaluation metrics.
* Visualize Results: Plot the confusion matrix and decision boundaries using Matplotlib/Seaborn.
* ~~Optimize: Adjust hyperparameters to enhance model performance.~~
* ~~Generate Report: Summarize results in a classification report.~~

Results

The model is expected to achieve an accuracy of over 85% on the test set, with balanced precision and recall across all three classes. The confusion matrix will highlight the model’s ability to distinguish between non-alcoholic (Classes 1 and 2) and alcoholic (Class 3) wines. Decision boundary plots will visualize how the model separates the classes based on key features like alcohol percentage and acidity. Hyperparameter tuning is expected to further improve performance, reducing misclassifications.

References

* Scikit-learn Documentation: Logistic Regression. Available at: <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>
* UCI Machine Learning Repository: Wine Dataset. Available at: <https://archive.ics.uci.edu/ml/datasets/wine>
* Multi-Class Logistic Regression: A Friendly Guide to Classifying the Many: <https://medium.com/@jshaik2452/multi-class-logistic-regression-a-friendly-guide-to-classifying-the-many-4a590c2e6c26>