Logistic Regression Classifier Implementation on the Wine Dataset

Tommy Liu and Pavel Filippov

November 2025

Introduction

The project aims to implement and evaluate the Logistic Regression (LR) Classifier in Java and apply it to the *wine* dataset with assistance from OVA and AVA Classifiers which the project team has implemented earlier in the semester. The main deliverable will be an implementation of the Logistic Regression Classifier supporting binary and multiclass classifications (using OVA/AVA). In addition to algorithm implementation, there will be an experiment section focusing on the performance assessment of the LR Classifier on the wine dataset, as well as a comparison between the Naive Bayes and the Logistic Regression Classifiers. In addition, some of the results from the experiment will be presented using Apache Superset, which is an open-source data visualization platform that supports interactive dashboards showing classification performance and feature importance.

Timeline and Tasks

Nov 11 - 16: LR Classifier Implementation

The project team is going to implement the LR Classifier in Java. As the LR Classifier is primarily a binary classifier, an implementation that supports binary classification will be sufficient. To apply the classifier to the wine dataset with multiple labels, in the next phase, the project team is going to integrate the classifier with multiclass classification (OVA/AVA) which was implemented earlier in the semester.

Nov 17 - 21: LR Classifier Validation and OVA/AVA Code Correction

The project team is going to create a small-scale (a few examples) sample data and manually perform the LR classification. Using this manual classification result as a reference, the team is going to validate the implementation of the LR Classifier and resolve issues and discrepancies. Meanwhile, since eventually the LR Classifier will be part of the classifier factory and would be used to perform multiclass classification using the OVA/AVA approach, the team is also going to revise their implementation of OVA/AVA from Assignment 6 based on the Gradescope feedback. By Nov 21, the team should have the functional LR Classifier integrated into the classifier factory and usable AVA and OVA classifiers.

Nov 21 - 25: Experimentation

To evaluate the LR Classifier, the primary metrics will be the accuracy of the testing dataset from 10-fold cross-validation and the runtime for training and testing. These metrics will also be used for the comparison between the Naive Bayes Classifier and the LR Classifier. Before the comparison, the team should try different combinations of hyperparameters, loss functions, and regularization methods to find the best combination for each classifier. This tuning process is similar to the process in the gradient descent assignment. To enhance presentation quality, the team will utilize Apache Superset as an interactive visualization layer over the model outputs.

Nov 30 - Dec 1: Project Write-Up and Presentation Practice

The project team is going to organize the code and provide careful documentation to the code, and upload the project to GitHub. Meanwhile, the team should summarize their work and findings and write a 2-page executive summary for the project. The team should also prepare the slides for the presentation and practice the oral delivery before the presentation.

Resources

Below are some resources the project team considers using.

- **Textbook, Chapter 9.6:** This chapter of the book describes the mechanism of Logistic Regression. This will serve as the basis for the classifier implementation.
- **Lecture Notes:** Lecture 14 from the course website also includes an overview of the LR Classifier. It also introduces various candidates for loss functions and regularizers. This will also help the project team with classifier implementation.
- **ML Assignment 6:** The assignment contains the implementation of OVA and AVA classifiers. The project team will integrate the LR Classifier into OVA and AVA to support multiclass classification.
- **Wine Dataset:** The *wine* dataset from Assignment 6 will be used for this project.
- **YouTube Channel – Artem Golubnichy:** The project team will use it to install Apache Superset and learn its fundamentals. This channel contains several basic examples for running simple visualizations.
- **Apache Superset:** The project team will use Apache Superset, a data visualization platform, to study and present some results from training and algorithm comparison. The specific results that will be visualized in Apache Superset are yet to be determined and will depend on the amount of time available for the team to study the platform.