CS-233 Project Milestone 1 Report

Ali Bendaoud, Jonathan Pilemand, Léonard Lemaire

April 2025

Contents

1 K-Means Clustering

1.1 Method

1.2 Experiment/Results

1.3 Conclusion

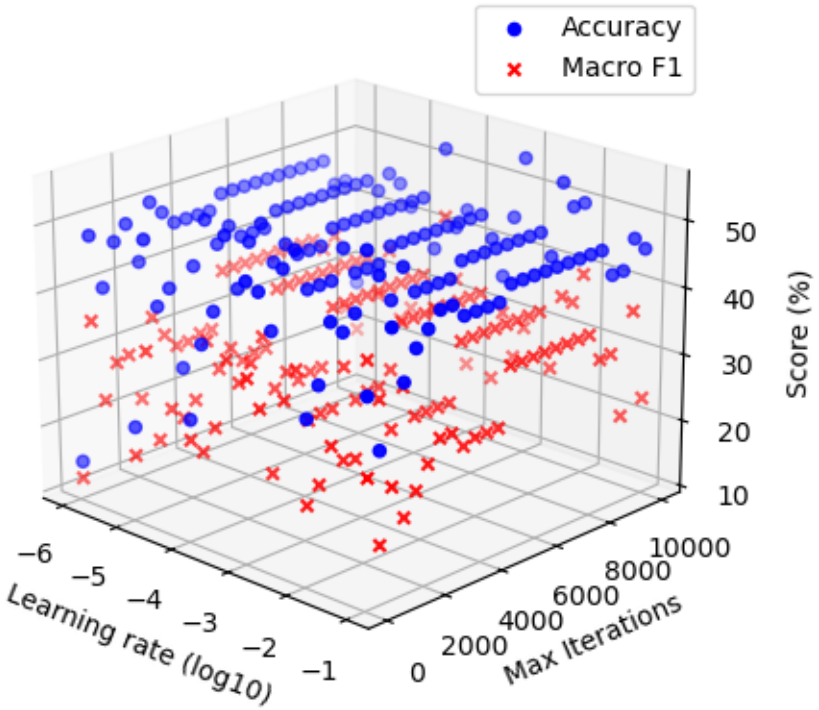
2 Multi-class Logistic Regression

2.1 Method

For this method, the idea is to construct a weight vector, through a given maximum number of iterations and a given learning rate, which will then be applied to the data to be classified. Before using this method, we need to preprocess the data. First, we normalize all data samples to avoid a scale imbalance between features. Then, since logistic regression is not distance-based like KNN and K-Means, we add a bias term equal to 1 to the data to make the classification more flexible and precise.

2.2 Experiment/Results

The challenge for this model is to find the best hyperparameters for classification, that is, to maximize accuracy and macro F1 scores for certain values of learning rate and maximum iterations.
Our idea was to use different values of learning rates and maximum iterations and iterate over them and then rank them by their subsequent values of accuracy and macro F1 score. Also, we prioritized the macro F1 score over the accuracy because it describes individual class accuracy instead of overall accuracy, which for this dataset seemed more appropriate since this is a heart disease database, meaning there would be more healthy people than sick, leading to an imbalanced dataset.
We used a wide range for each parameter, from 1e-6 to 10 in an exponential manner (i.e. 1e-6, 1e-5, ..., 10) for the learning rate and from 100 to 10000 in a linear manner (i.e. 100, 500, 1000, ...) for the number of maximum iterations. The first problem we encountered was that the learning rate was too high when the exponents increased and we got some errors. So we had to reduce the learning rate range to [1e-6, 1e-1]. After this search, we got the best results for a learning rate of order 1e-4 and maximum iterations of order 9200.



2.3 Conclusion

After our search for the best hyperparameters, we can check the final performance of our model, which is the following:

Logistic Regression	Training	Test	Validation
Accuracy	64.979%	61.667%	47.917%
F1 Score	0.44110	0.32047	0.36121

Even though accuracy decreases on the validation set, the F1 score remains pretty consistent, which means the model is not vulnerable to the imbalance of the classes.

3 KNN

3.1 Method

3.2 Experiment/Results

3.3 Conclusion

4 Conclusion