# Advanced Machine Learning – Logistic Regression Optimization Algorithms

Project report

Niziołek Norbert 313395
Zagórski Mateusz 313509

Instructor: mgr Anna Kozak

April 3, 2024

# Contents

# 1 Methodology

## 1.1 Dataset selection

In accordance with the project task, nine datasets were selected – three small ones and six large ones. Aliases to the datasets' names used in the rest of the report are in brackets at the end of every enumeration point.

### 1.1.1 Small datasets

1. **Raisin** – classifying raisins as one of the two species – Kecimen or Besni. 8 variables, 900 samples. (small1)
2. **Banknote Authentication** – detecting if a banknote is genuine or forged based on its image's parameters. 5 variables, 1372 samples. (small2)
3. **Mammographic Mass** – detecting if mammographic mass is malignant or benign based on patient's age and the tumor's attributes. 6 variables, 961 instances. (small3)

### 1.1.2 Large datasets

1. **Ionosphere** – classification of radar data as showing (or not) evidence of some type of structure in the ionosphere. 35 variables, 351 samples. (large1)
2. **Algerian Forest Fires** – detecting a fire based on geographical, time and climate data. 16 variables, 244 samples. The only categorical variable *region* was dropped to satisfy project data requirements. One of the observations was also deleted due to its corrupt fields. (large2)
3. **Breast Cancer Wisconsin (Diagnostic)** – classifying a tumor as malignant or benign based on extensive description of an image. 31 variables, 569 samples. (large3)
4. **Steel Plates Faults** – detecting steel plates faults. 34 variables, 1941 instances. (large4)
5. **QSAR biodegradation** – deciding if a material is biodegradable based on its molecular descriptors. 42 variables, 1055 samples. (large5)
6. **Dry Bean** – classifying beans as one of seven species. 17 variables, 13611 samples. The seven categories were grouped into two to achieve binary classification. (large6)

Large datasets 4 and 5 come from the OpenML repository. All the other datasets come from the UCI repository. The variable numbers include the target variable.

All missing values in the datasets listed above have been imputed using the **KNN** method.

## 1.2 Stopping rule

The stopping rule used for all the algorithms is based on the difference between the parameters estimation in the previous and current iteration. The fitting process is stopped when the following condition is true:

$$\|\beta - \beta'\|_2 < \epsilon,$$

where $\beta$ is the vector of estimated model parameters in the current iteration, $\beta'$ is the same vector for the previous iteration and $\epsilon$ is a tolerance with a default value of $10^{-3}$.

## 1.3 Performance comparison

The performance comparison of the three algorithms is shown in figure 1. The results are averaged over 5 experiments conducted on every dataset on different random train-test splits.

For most of the datasets, the results of all three algorithms are similar. The only notable differences are observed for the *small1*, *large4* and *large6* datasets. In all three of those cases, the IRLS method performed best, followed by SGD, with Adam performing the worst. The general conclusion of all conducted experiments is that the IWLS method is the most stable algorithm and usually provides the best results.

## 1.4 Implementation issues

For some datasets (large2, large3, large4, large5) SGD method turned out to be unstable and some numeric errors occurred in matrix calculations. To prevent fatal errors maximum iterations parameter was limited in these cases as the algorithm wouldn't converge anyway.

Correlated variables detection and deletion have been implemented but it did not result in performance improvement – the opposite effect was observed. The final results presented in this report have been obtained from the original set of variables.

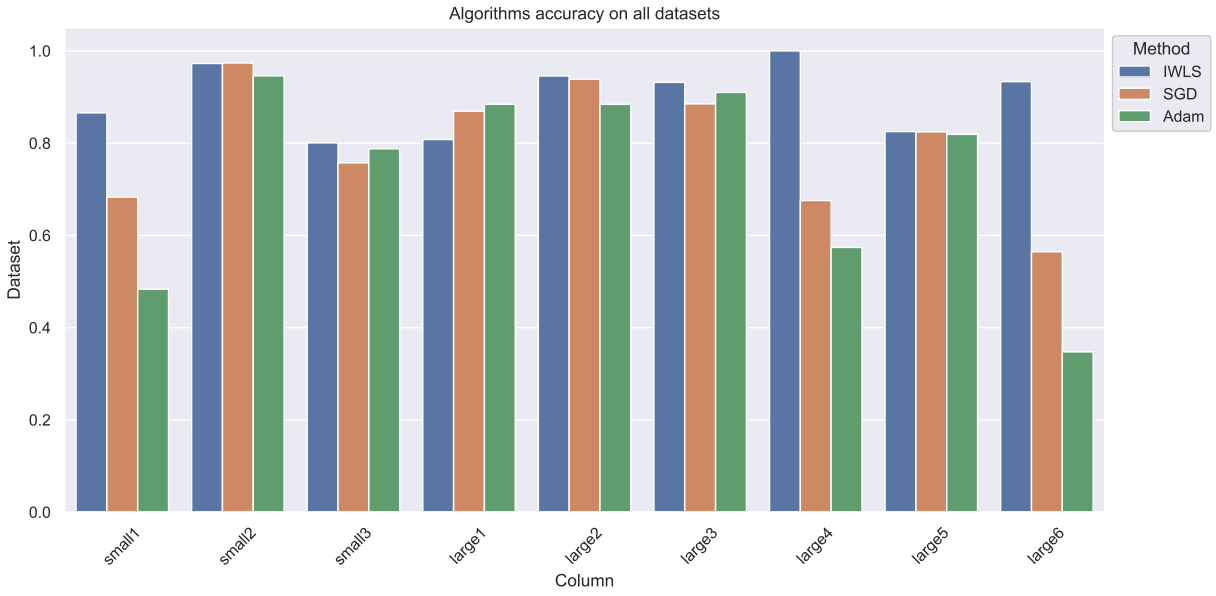


Figure 1: Performance comparison of IRLS, SGD and Adam for all datasets.

## 2 Convergence analysis

The results of the convergence analysis are presented in figure 2. For datasets *small2*, *small3*, *large1* and *large5* all methods are convergent without significant fluctuations during the fitting process. In case of datasets *large2* and *large3* all three algorithms converged, but in case of the SGD algorithms there were some fluctuations during the training process.

In other datasets, one of the methods did not converge. In the case of *small1* and *large6* it was the SGD method and for the dataset *large4* the Adam method did not converge.

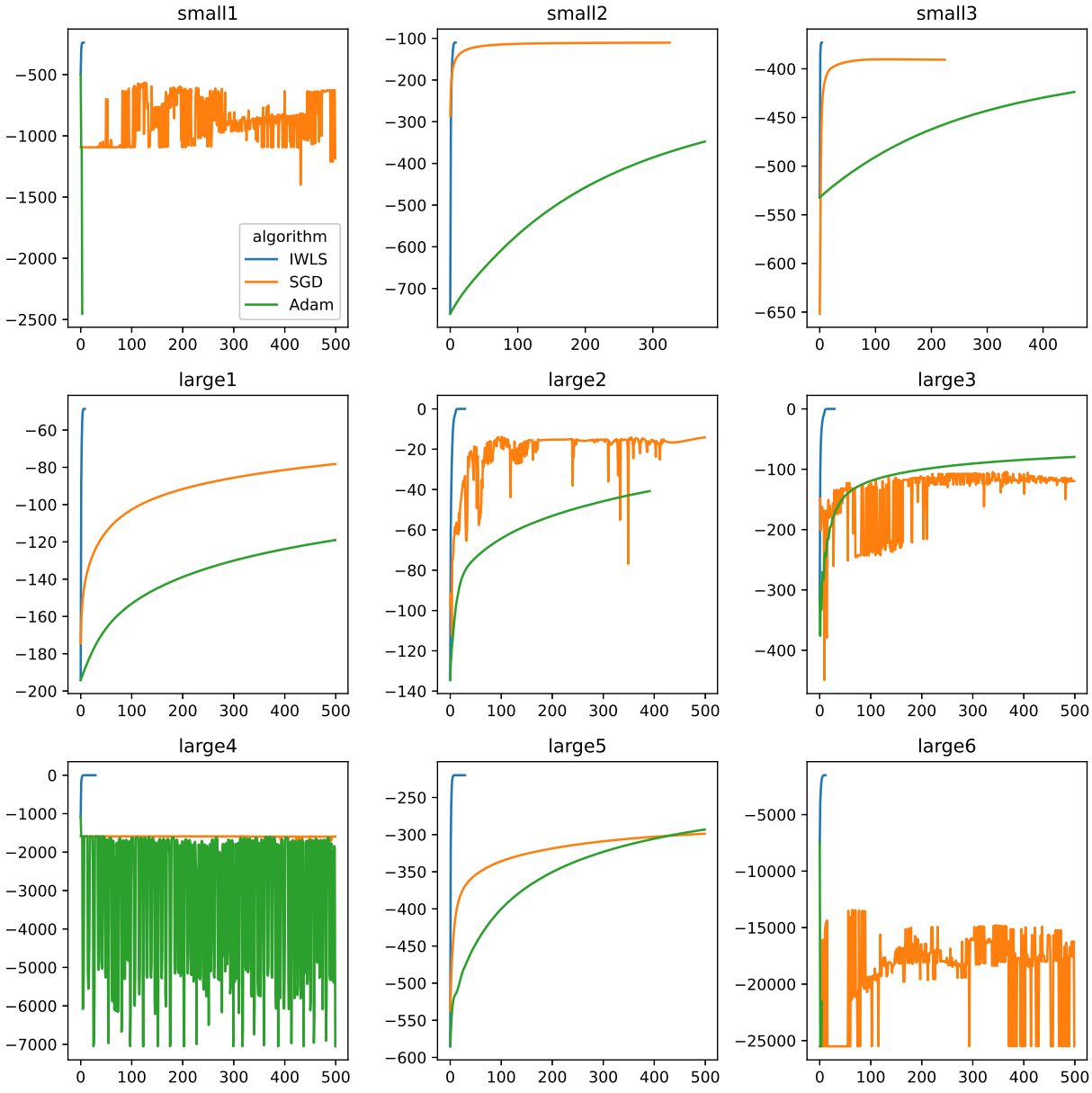


Figure 2: Log likelihood of all three algorithms on all nine datasets. The X axis represents the iteration and the Y axis represents the log likelihood value.

# 3 Comparison of classification performance

The results of the comparison of the implemented algorithms with other popular classification methods are presented in figure 3. The results are an average of three runs for each dataset and each method.

For all of the datasets, the other classification techniques performed similarly to at least one of our algorithms. In case of most of the datasets, the results for all seven tested methods were similar.

In the cases were the three implemented methods differed, all four other methods performed similarly to the IWLS algorithm which also happened to be the best our of the three implemented methods. There are only two exceptions to that rule. The first one is the LDA algorithm for *large4*, where it performed worse than the other methods, but still better than SGD and Adam. The second one is *large5* where five out of six models performed almost exactly the same, decision tree was notably worse and QDA was even worse than that. However, those difference are all within ten percentage points of each other.

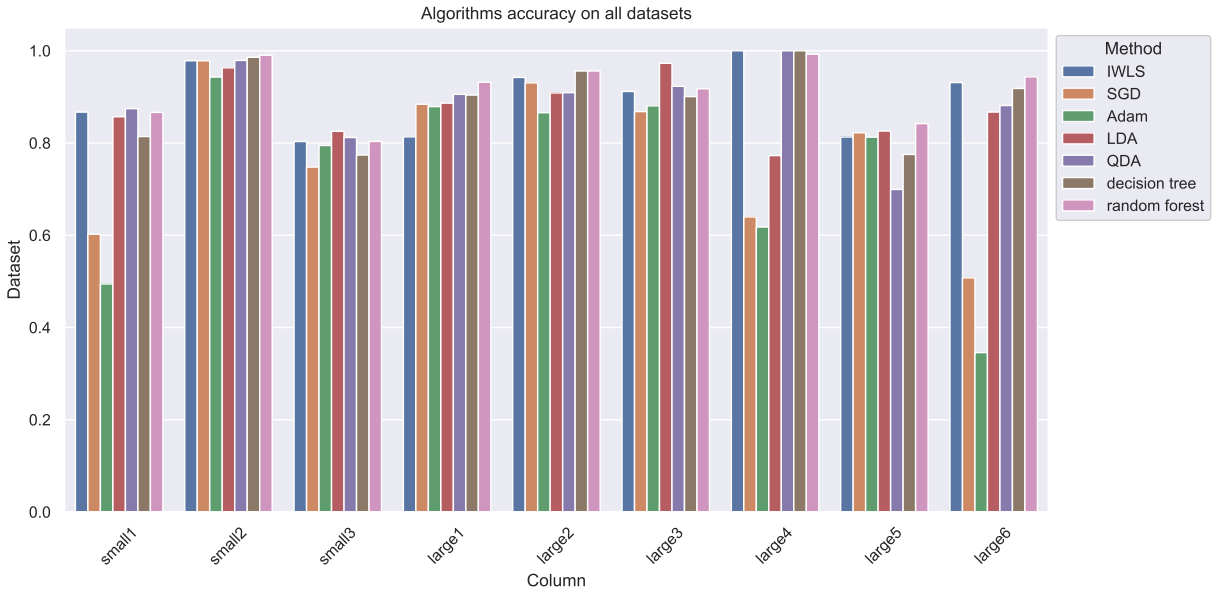


Figure 3: Performance comparison with other classification methods for all datasets.

# 4 Comparison of classification performance of models with and without interactions

The results of testing the effect of interactions on the model's performance are shown in figure 4.

In all cases the model with interactions is slightly better than the model without interactions and the difference is within 3 percentage points. The only outliers are SGD and Adam with *small1* and SGD with *small3* where the difference is larger – 10, 7 and 5 percentage point respectively.

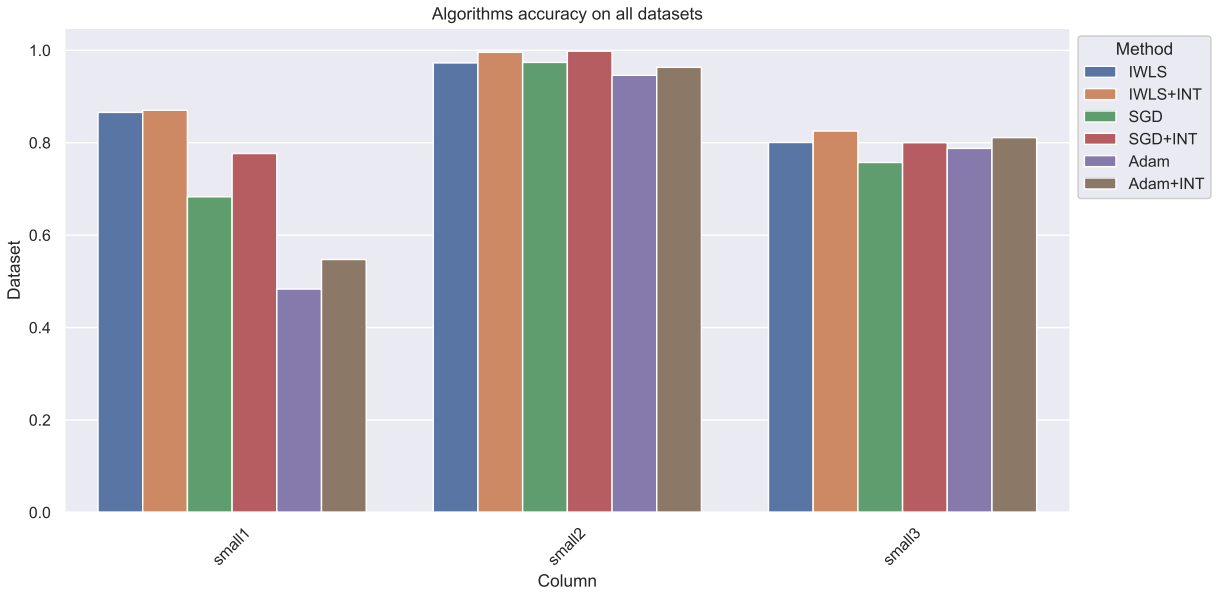


Figure 4: Performance comparison of models with and without interactions for small datasets.