Advanced Machine Learning - Project 1

Mieszko Mirgos, Tomasz Siudalski, Piotr Robak

March 2024

Contents

1 Introduction

This report contains the results of our work for Project 1 of the Advanced Machine Learning course. The goal of the project was to implement and compare in terms of performance different optimization algorithms, namely IWLS, SGD, and ADAM algorithms. Those algorithms were used in logistic regression for parameter estimation. We analyzed the Convergence of the above algorithms, classification performance as well as performance with and without interactions between variables. All of our findings, as well as how we got them are presented below.

2 Methodology

2.1 Datasets

For conducting our experiments we decided to use 9 different data sets, mainly taken from UC Irvine Machine Learning Repository and OpenML Repository. We decided to use six large data sets, each containing more than 10 variables:

- bodyfat - Body fat
- features - Sports articles for objectivity analysis
- fri_c0_1000_25 - Friedman dataset
- phpDQbeeh - SPECTF heart data
- phpGUrE90 - QSAR biodegradation Data Set
- puma32H - Puma 560 dataset

and three smaller ones (where each has at most 10 variables):

- maternal health risk
- php8Mz7BG - phoneme
- rice cammeo osmancik

None of those data sets have missing values which helps in prepossessing and gets rid of the choice of what data augmentation method we should use.

2.2 Preprocessing

To run logistic regression algorithms on the mentioned datasets we had to preprocess them accordingly. Firstly we loaded the datasets from files with various formats ('.arff', '.csv', '.xlsx') and made sure there were no missing values. Fortunately, in all dataframes, there was no absence of data so we did not have to fill in or remove the lacking values. Secondly, we encoded the target variable to 0 and 1. In the case of the maternal health dataset, we combined two categories, 'high risk' and 'medium risk', making 'low risk' the second class. After that, we split the dataframe into train and test datasets. Then we had to remove the collinear variables. High collinearity can lead to a huge variance of the coefficient estimates, causing the predictions to be unstable and unreliable. Our algorithm selected such variables for removal from a dataset in the following way.

- A correlation matrix based on Pearson's correlation coefficient was created.
- Upper triangle of the matrix was extracted (as the correlation is symmetrical).
- For each pair of columns if the correlation between them was higher than 0.7, which is a wildly used threshold, the column with the smaller variance was chosen. We chose such a method as generally variables with higher variance carry more information.

We selected the columns for deletion based on train data and then removed them from both sets. We did not apply any further preprocessing as it drastically changed the outputs of the experiments (especially the convergence analysis).

2.3 Algorithms

The main algorithm that was used for parameter estimation was logistic regression. For optimization algorithms to conduct experiments on we used:

- IWLS - Iterative Reweighted Least Squares
- SGD - Stochastic Gradient Descent
- ADAM - Adaptive Moment Estimation

Which we implemented ourselves (their implementations can be found in files: *iwls.py*, *sgd.py*, *adam.py* respectively). For the stopping rule for logistic regression, we decided to use the following two criteria:

- maximum of 500 epochs
- patience of 20 epochs - if log-likelihood doesn't improve for 20 epochs, then we stop calculations and return to iteration with the best results.

2.4 Experiments

We conducted the following experiments:

- Convergence analysis - done by analyzing the value of the log-likelihood function in each iteration.
- Comparing the performance of logistic regression - between optimizers and compared with LDA, QDA, Decision tree, and Random forest to see how good are those methods in comparison to some of the more popular classification methods.
- Comparing of the model with and without interactions - done by comparing logistic regression models for each optimizer with models for optimizers with interactions.

Each of the experiments was performed on all datasets to gather data that is not dependent on specific data but makes our results more broad.

3 Convergence Analysis

In this section, we present the results of the conducted experiments regarding the convergence analysis for the three optimization algorithms for each dataset. The plots can be seen on the next page. To be able to better see Adam and IWLS, plots without SGD are included in Appendix A.

SGD

In most cases (plots a,b,d,f,g,h), the SGD optimizer behaves very unstably, stopping at a loss much higher than the other two algorithms and oscillating around it, struggling to improve. It may be because of the too frequent updates which create much noise and also the standard learning rate of 0.01 may be too high for the datasets. In the case of the Rice dataset, it does not learn at all. However, for the Phoneme and Friedman datasets, the algorithm managed to achieve a loss similar to other algorithms in a very small number of iterations.

Adam

The Adam optimizer on the contrary to SGD learns very stably, decreasing loss with every iteration and producing a very smooth loss curve. This is thanks to the adaptive estimation of first-order and second-order moments, combined with the batch updates of weights. The algorithm achieves almost identical or slightly higher loss than IWLS.

IWLS

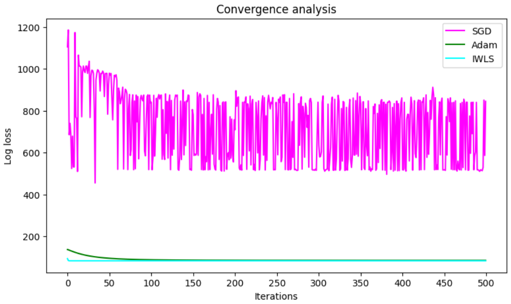
The IWLS optimizer converges by far the fastest of those three. It obtains the smallest loss converging in under 10 iterations, and then stopping or decreasing extremely slowly. The algorithm archives the best results because it optimizes the likelihood function directly, which is well-suited for the probabilistic nature of logistic regression.

Summary

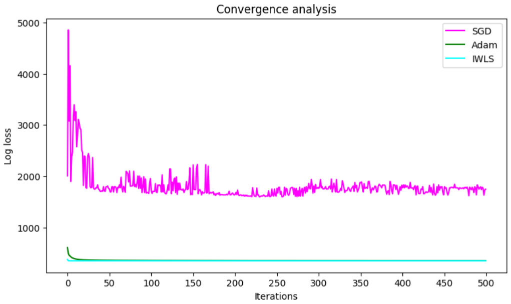
Algorithm	**Avg Num Iter**
SGD	109
Adam	455
IWLS	174

Table 1: Average number of iterations

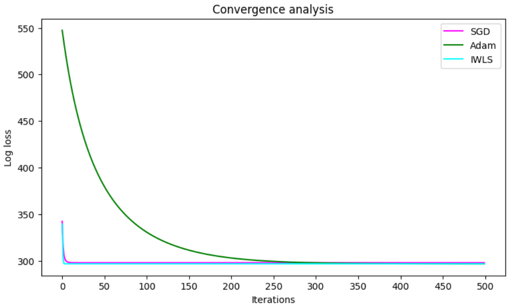
From the conducted analysis we can clearly see that IWLS is the best optimizer out of those 3. SGD algorithm stops in a smaller number of iterations, however, it does not quite find the minimum and its behavior is very unstable. It achieves on average 5.7 times higher loss than IWLS. Truthfully, the IWLS converges faster, it finds the minimum very quickly, in all cases in under 10 iterations and then sometimes decreases extremely slowly. Adam often finds a minimum very close to the one found by IWLS, but it still takes 2.6 times more iterations.



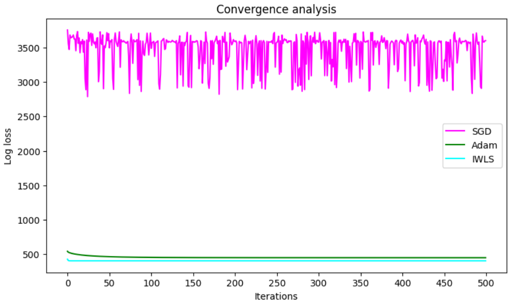
(a) Body Fat



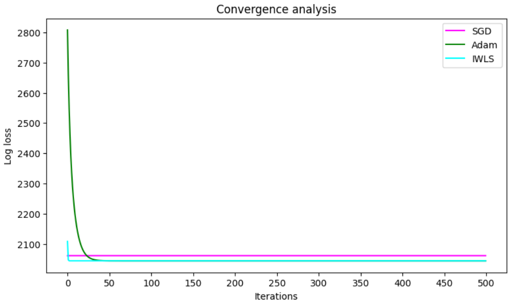
(b) Sports articles



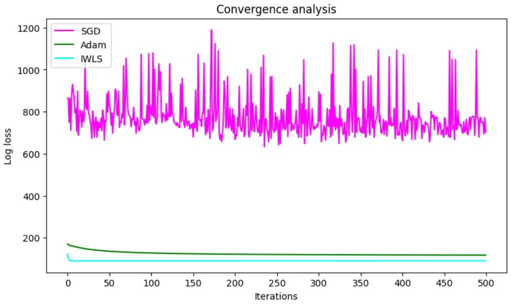
(c) Friedman



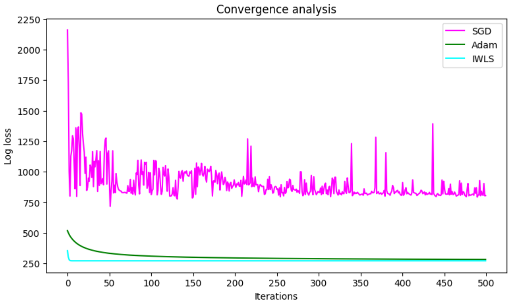
(d) Maternal health risk



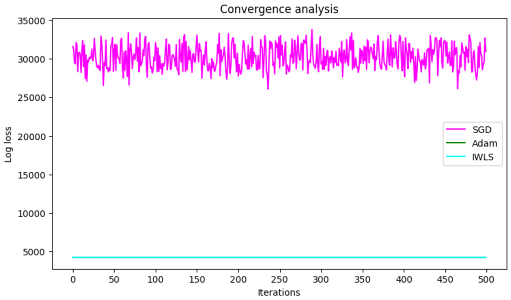
(e) Phoneme



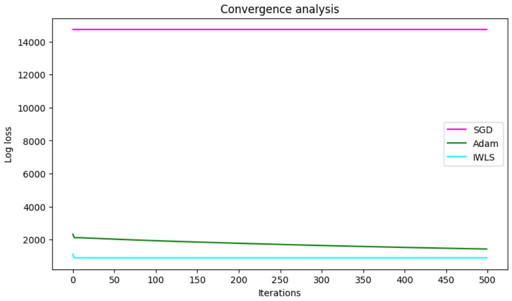
(f) SPECTF



(g) QSAR



(h) Puma



(i) Rice

4 Comparison of classifiers performance

In this section, we will look at how different classifiers fared against different datasets. We will be considering the following classifiers

- lr_iwls - logistic regression with iwls optimizer
- lr_adam - logistic regression with adam optimizer
- lr_sgd - logistic regression with sgd optimizer
- lda - linear discriminant analysis
- qda - quadratic discriminant analysis
- dt - decision trees
- rf - random forest

As presented on 1 tree tree-based classifiers have on average achieved the best results. Logistic regression-based classifiers have achieved noticeably different results depending on the classifier, average LDA and QDA scores fall between Adam and iwls optimizers.

When we reduce the granularity 2 level to individual datasets we see that in many cases the relationship of IWLS > Adam > SGD holds in most cases, but there are outliers. It is noticeable that trees do maintain their superior predictive power in most cases here as well.

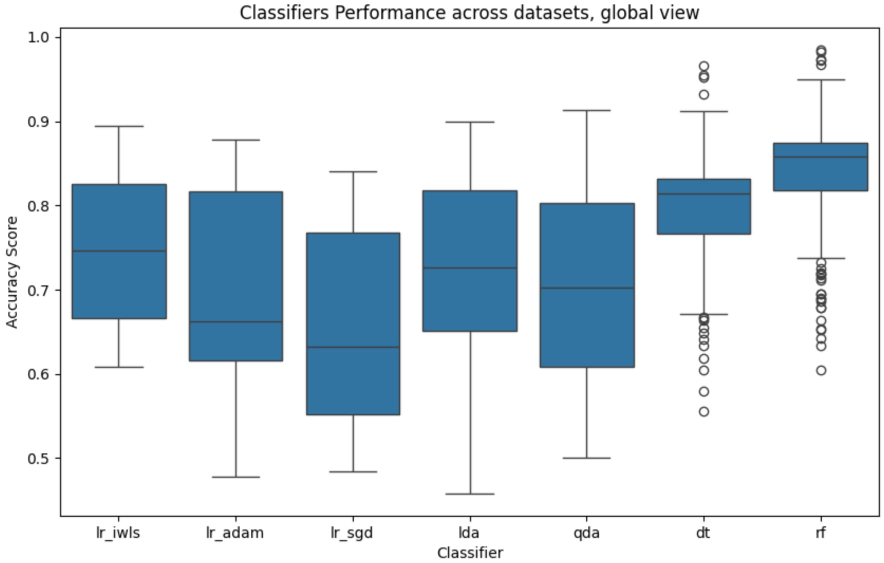


Figure 1: Accuracy scores across all datasets compared

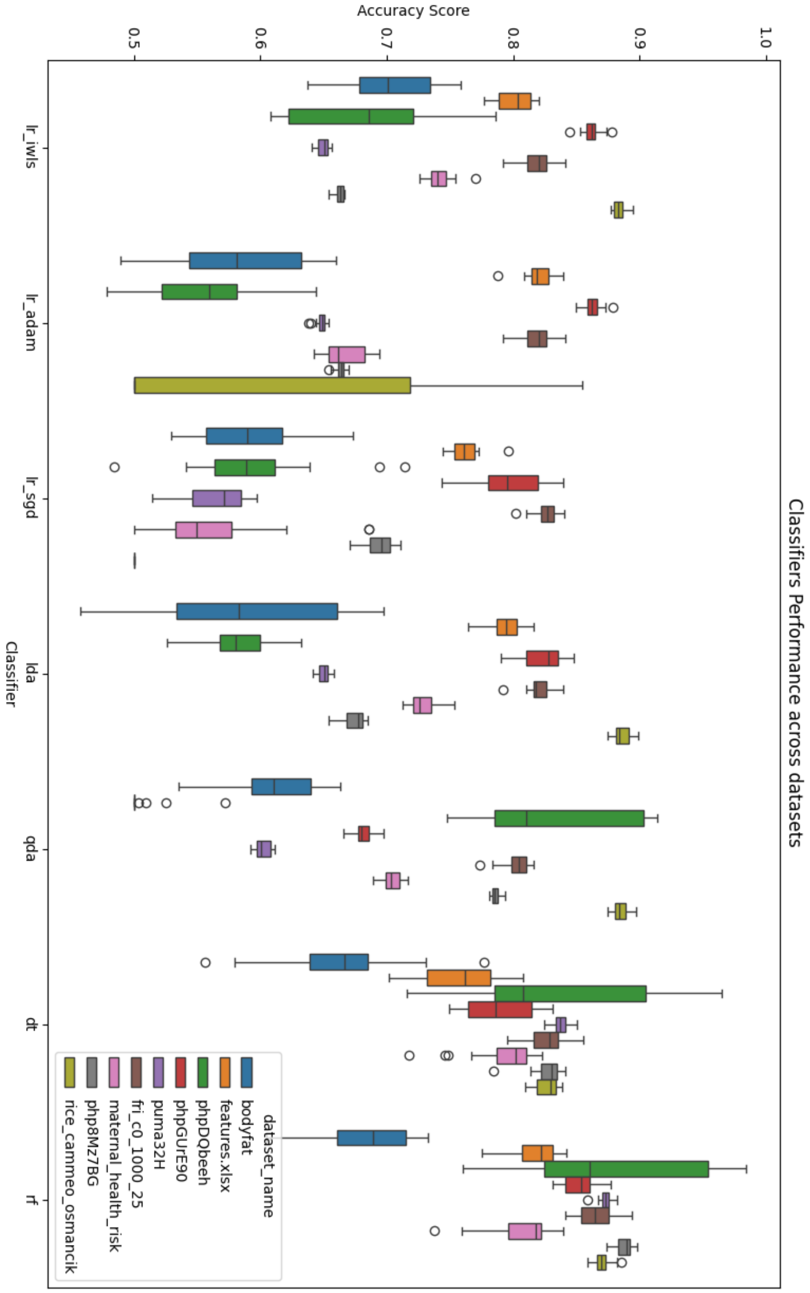


Figure 2: Accuracy scores aggregated across all datasets compared

5 Comparison of classifiers performance for models with and without interactions

In this section, we will evaluate how did introduction of interaction variables influenced the accuracy of the logistic regression classifier on different datasets. The names of classifiers with interactions correspond to the names from 4, classifiers without interactions have the "_no_int" appendix.

As we can see on all graphs in this section in the case of the phoneme dataset (orange) the inclusion of interactions or lack thereof played a noticeably greater role than the choice of optimizer itself on a scale of individual datasets. This behavior was largely consistent across optimizers (with respect to the mean score).

In another case, the maternal health risk dataset, both choices of classifier and interactions seem to have a noticeable effect on accuracy, although this time it was opposite to that of the phoneme dataset.

As confirmed by the plot, the dataset itself was a key factor in whether or not the inclusion of classifiers improved results, as well as in determining how different optimizers fare.

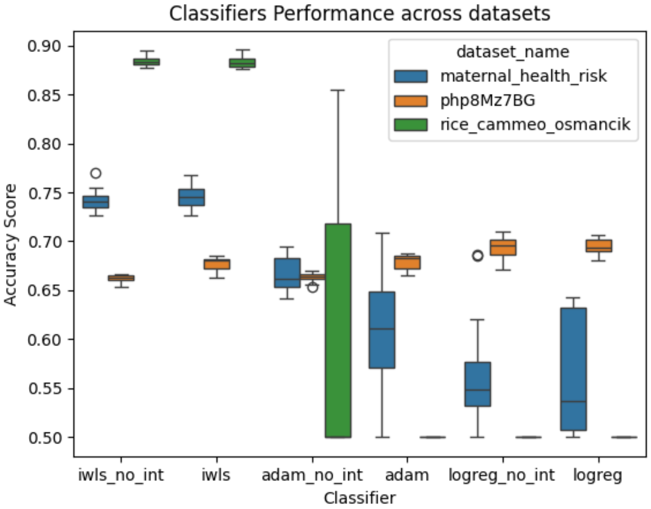
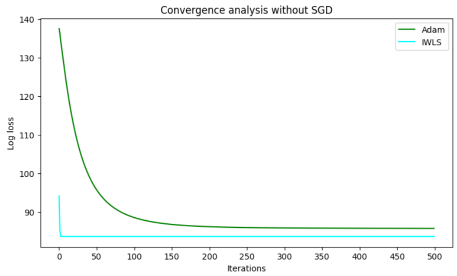


Figure 3: Accuracy scores across all datasets compared

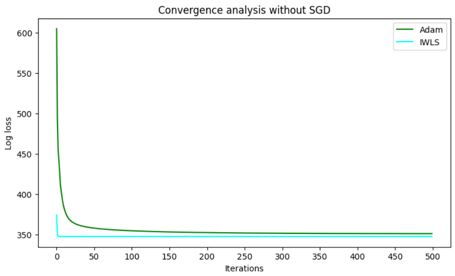
6 Conclusion

All of our tests showed similar results. In all of our tests, the best performance was achieved using the IWLS optimizer. The Second was ADAM, and the last was SGD, which in many cases was way too unstable compared to the previous two. This is quite predictable as SGD is a really simple algorithm and in each iteration, its steps don't become shorter, and as such it has a hard time reaching optimum, only osculating around it. We can see that ADAM which is quite similar but has such characteristics performed much better. When it comes to the performance of logistic regression compared to other algorithms, we concluded that it depends on the optimizer used. When we use SGD - the worst of our optimizers, then it's by far the worst algorithm tested. On the other hand, when we used IWLS (which yielded the best results out of 3 optimizers) our results were better than LDA and QDA, but decision trees and Random forests were still superior and achieved accuracy over 5% point higher.

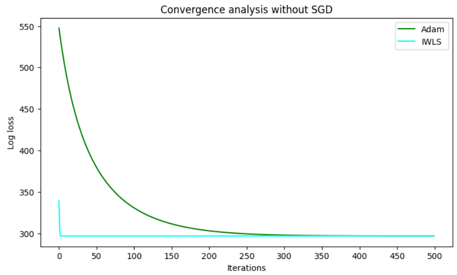
A Plots without SGD



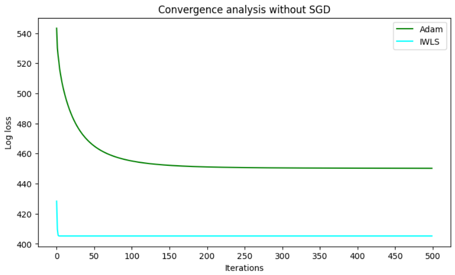
(a) Body Fat



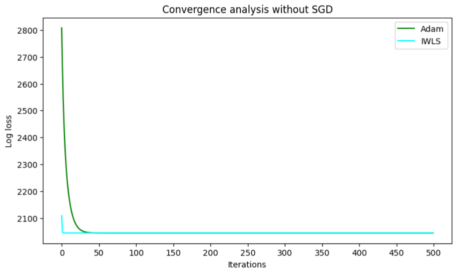
(b) Sports articles



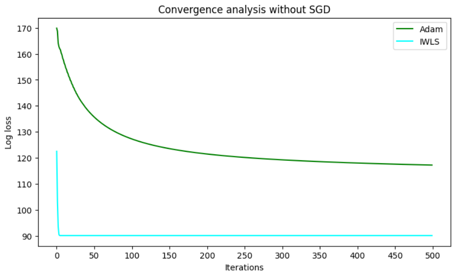
(c) Friedman



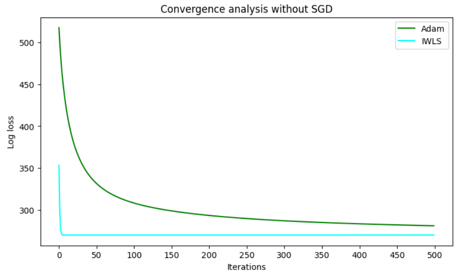
(d) Maternal health risk



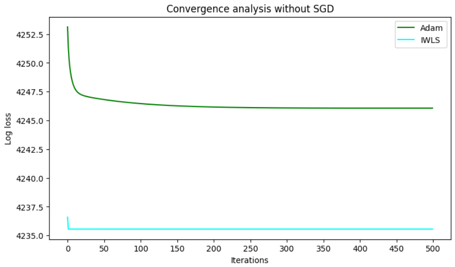
(e) Phoneme



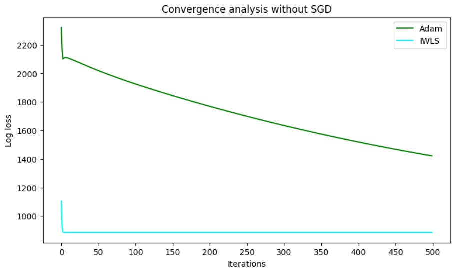
(f) SPECTF



(g) QSAR



(h) Puma



(i) Rice