Advanced Machine Learning Project 1

Dutt Salveen Singh, Patryk Prusak

Warsaw University of Technology

April 3, 2024

Contents

1 Methodology

We have implemented logistic regression with 3 variations of optimization (Adam, IWLS, SGD). For each of the 10 datasets, each of the 7 models has been trained and tested on 6 different splits and 5 different seeds to obtain some level of statistical significance. In addition, the 4 additional datasets have been introduced in the form of a small dataset with interactions by adding columns resulting from the multiplication of features. The additional datasets have been trained and tested only on the implemented logistic regression variations. In total 2460 experiments have been performed. Each dataset has been prepared for classification by dropping highly correlated features (mostly those with correlation > 0.9) and mapping the binary classes to either 1 or 0. As for the datasets, we decided to choose the following:

Small Dataset	**Number of Attributes**
Blood Transfusion Service Center[1]	5
Iris[2]	5
Diabetes[3]	9
Pollen[4]	6

Table 1: Small Datasets Choice

Large Dataset	**Number of Attributes**
KC2[5]	22
PC3[6]	38
Madelon[7]	501
Biodegradable chemicals[8]	42
Japanese Vowels[9]	15
Vehicles[10]	19

Table 2: Big Datasets Choice

Each optimizer process happens over a number of iterations, we have implemented an early stopping rule indicating that the optimizer achieved satisfactory results or in other words converged. The stopping rule is: if the absolute difference between consecutive loss functions is smaller than 0.00001 stop. For the loss function, cross-entropy loss is used for each algorithm

$$l = -(y\log(p) + (1-y)\log(1-p))$$

. As a performance measure we use balanced accuracy[11]

$$Balanced\ Accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2}$$

Figure 1: Balanced Accuracy

2 Convergence Analysis

Looking at the Figures 2 and 3. we can easily see that the fastest convergence on most datasets occurs with the use of the IWLS algorithm. Adam and SGD often have trouble converging within 500 iterations (the number of iterations is capped at 500), for example in the 'vehicle' dataset. Some interesting conclusions from the loss functions are that Adam has exceptional trouble converging on the 'biodegradable' dataset within 500 iterations whereas IWLS and SGD converge within 1 to 5 iterations. For the 'diabetes with interactions' dataset, SGD does not converge quickly and the loss function momentarily grows back up at the end of the 500 iterations mark. In the case of the 'pollen' dataset, SGD converged in 1 iteration for both with and without interactions, however not in the way we would expect since the loss has grown. This might indicate that perhaps for SGD a different stopping rule might be more optimal, with a smaller (more strict) sopping rule coefficient. Although all models nearly always manage to reduce the loss over iterations, SGD's loss function graph at times presents itself with lots of direction changes regarding error growth. SGD also is quite sensitive to higher learning rates, for Adam and IWLS we have chosen a learning rate of 0.001, however for SGD a learning rate of 0.00000001 was necessary to obtain satisfactory results and avoid overflows. IWLS seems to converge quite quickly in many cases with good results, however, its implementation presents itself with the need for matrix inversion, which fails in the case of singular matrices. Adam is more reliable in that sense.

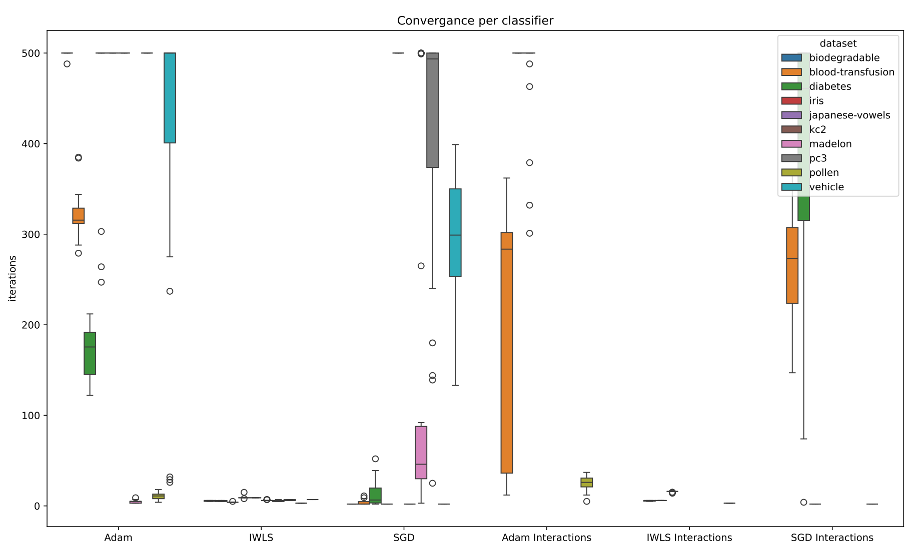

Figure 2: Convergence per classifier

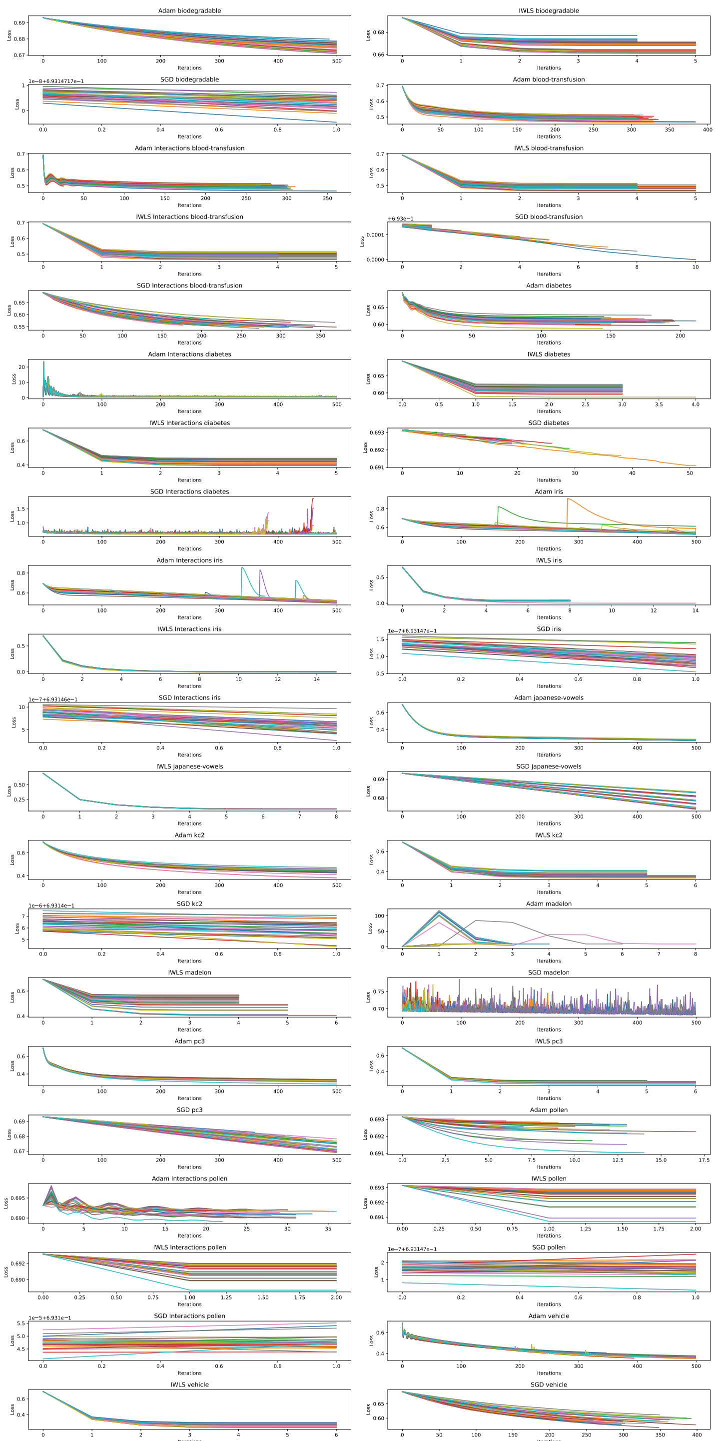

Figure 3: Convergence per datasets

3 Comparison of Classification Performance

Apart from logistic regression with Adam, IWLS, and SGD we also tested LDA, QDA, Decision Tree, and Random Forest. In Figure 4 we compare the accuracies of all classifiers per dataset. In general, accuracies are quite similar for all models with Decision Tree, Random Forest, and QDA achieving higher results for larger datasets such as 'vehicle' or 'japanese-vowels'. In Figure 5 one can see that the SGD variation performs the worst, with results falling mostly close to 0.5. As previously QDA, Decision Tree, and Random Forest achieve higher results on average than the rest of the models, however, in general, the differences are in the 0.1 range.

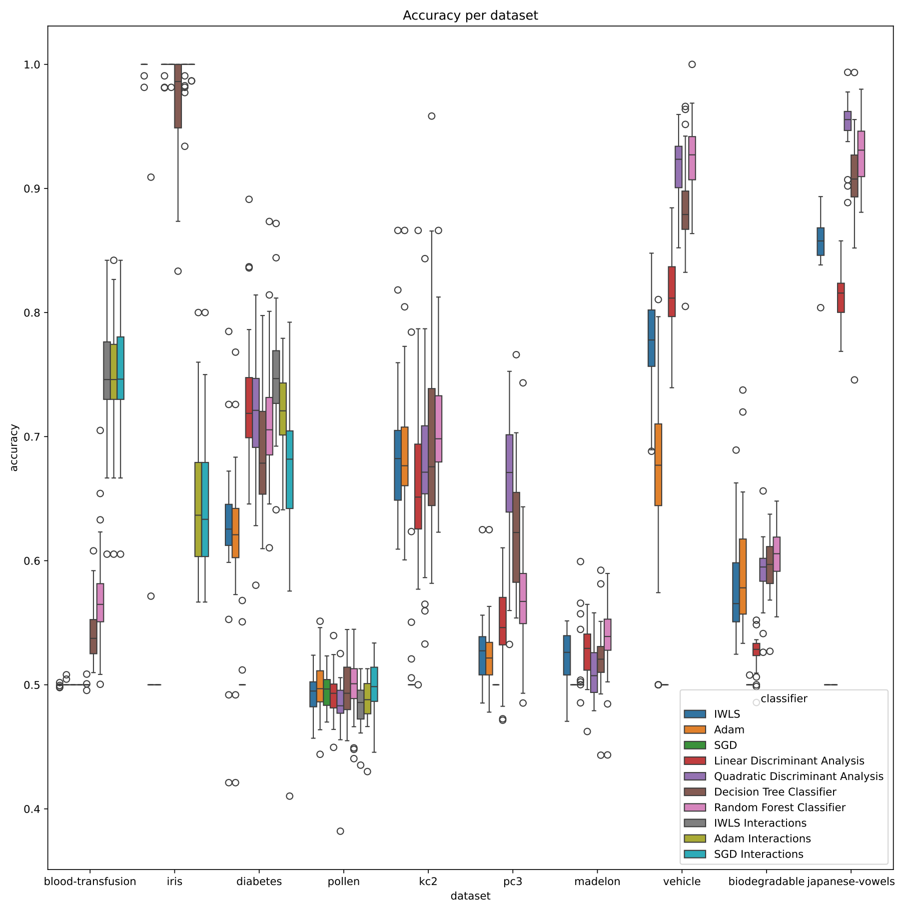

Figure 4: Average Balanced Accuracy comparison per dataset

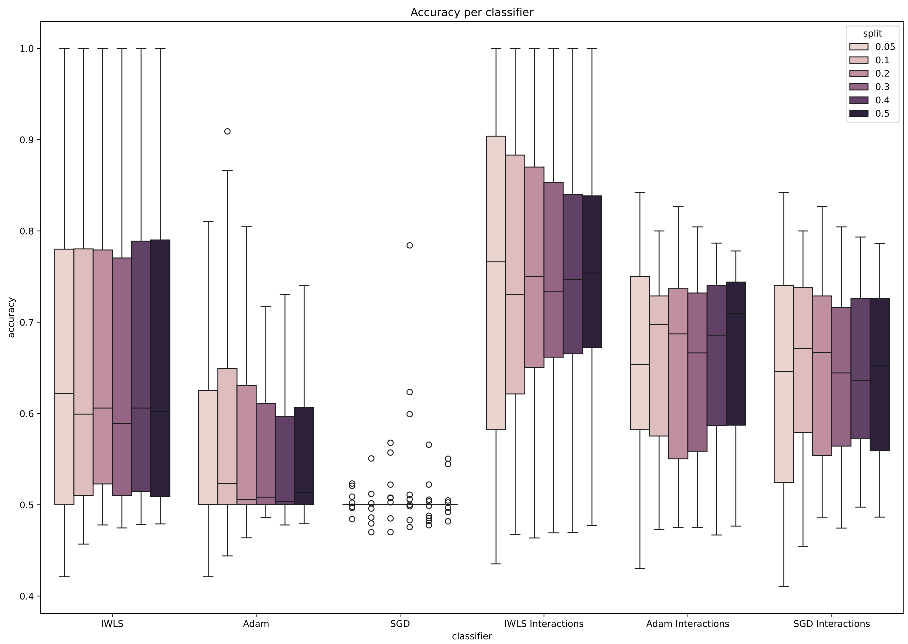

Figure 5: Average Balanced Accuracy comparison by split

4 Comparison of classification performance of models with and without interactions

We have included additional variation for the small datasets with interactions between variables. Each of the implemented optimizers has then been tested on both of the datasets. In Figure 6 boxplots of accuracies for all logistic classifier variations are visible, whereas in Figure 7 the results have been grouped by datasets. In general, each optimizer achieves higher accuracy for datasets with interactions at the cost of higher variance, IWLS obtains the highest variance, with mean balanced accuracy oscillating between 0.5 and 0.6. Adam and SGD offer similar results, with Adam having better accuracy. Comparing the results per dataset, one can see that the algorithms have troubles with the pollen dataset (with and without interactions), achieving a mean of 0.5 for balanced accuracy. IWLS is surprisingly good at classifying the Iris dataset, achieving a balanced accuracy of 1. Similarly to the per split, here also variations with interactions perform better on average, achieving similar results for each optimizer (with IWLS in the lead in Iris). In conclusion, models with interactions score higher on the balanced accuracy metric but introduce more variance.

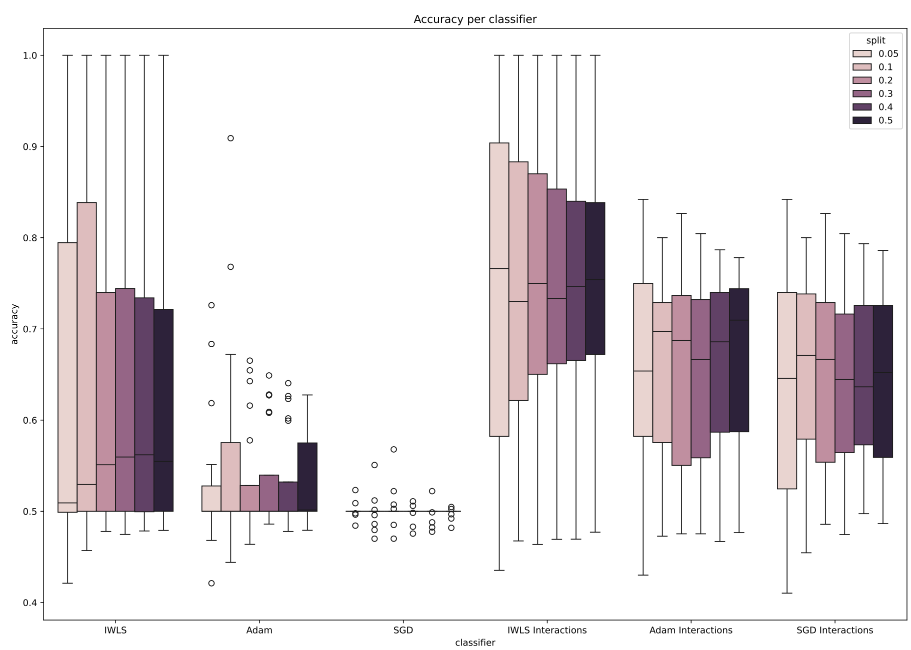

Figure 6: Accuracy comparison for logistic regression by split

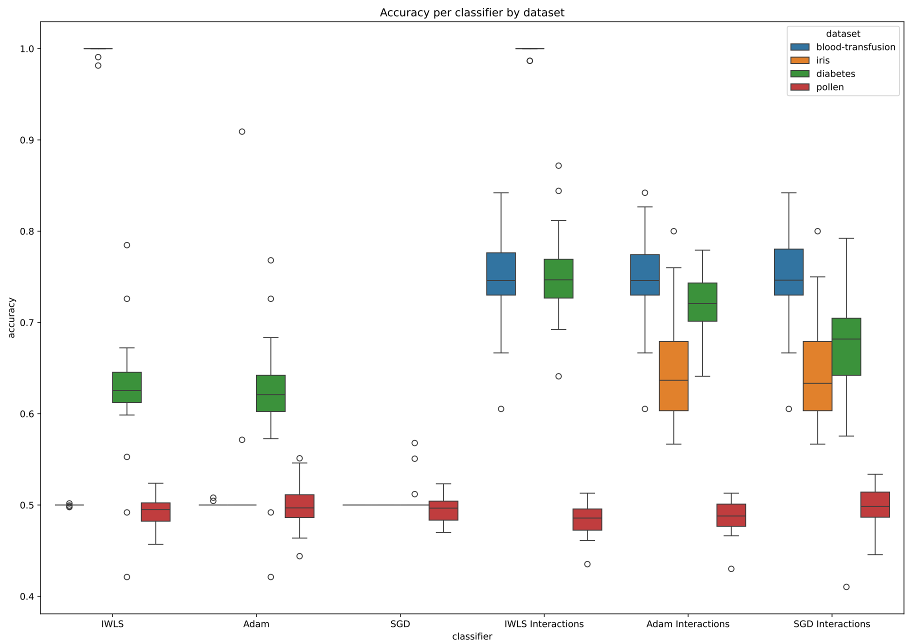

Figure 7: Accuracy comparison for logistic regression by dataset

List of Tables

List of Figures

References

[1] *Blood Transfusion Service Center*. URL: https://archive.ics.uci.edu/dataset/176/blood+transfusion+service+center. (accessed: 26.03.2024).

[2] *Iris*. URL: https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html. (accessed: 03.04.2024).

[3] *Diabetes*. URL: https://archive.ics.uci.edu/dataset/34/diabetes. (accessed: 03.04.2024).

[4] *Pollen*. URL: https://openml.org/search?type=data&status=active&qualities.NumberOfClasses=%3D_2&id=871&sort=runs. (accessed: 03.04.2024).

[5] *KC2*. URL: https://www.openml.org/search?type=data&sort=runs&id=1063&status=active. (accessed: 03.04.2024).

[6] *PC3*. URL: https://www.openml.org/search?type=data&sort=runs&id=1050. (accessed: 03.04.2024).

[7] *Madelon*. URL: https://archive.ics.uci.edu/dataset/171/madelon. (accessed: 03.04.2024).

[8] *qsar-biodeg*. URL: https://www.openml.org/search?type=data&status=active&id=1494. (accessed: 03.04.2024).

[9] *Japanese Vowels*. URL: https://archive.ics.uci.edu/dataset/128/japanese+vowels. (accessed: 03.04.2024).

[10] *Vehicle*. URL: https://openml.org/search?type=data&status=active&qualities.NumberOfClasses=%3D_2&id=994&sort=runs. (accessed: 03.04.2024).

[11] *Balanced Accuracy*. URL: https://neptune.ai/blog/balanced-accuracy. (accessed: 03.04.2024).