Project 1 - Advanced Machine Learning

Piotr Bielecki, Tomasz Krupiński

March 2024

Abstract

This report summarizes the work done for a project regarding the implementation and performance comparison of different optimization algorithms for logistic regression, utilizing multiple different, real-life datasets. The algorithms discussed in this report are IWLS (Iteratively (Re)Weighted Least Squares), SGD (Stochastic Gradient Descent) and ADAM (Adaptive Moment Estimation). In addition, convergence analysis of log-likelihood is performed. Finally, performance is tested with and without interactions between data.

1 Methodology

(Task 1) 9 Datasets were selected for this project, 3 of which contain fewer than 10 features. All values visible in table 1 were obtained after initial preprocessing of the datasets, which included removing highly correlated features (absolute value of the correlation greater than 0.8). Figure 1 represents correlation matrices of *rice* dataset before and after having removed highly correlated columns.

dataset name	number of columns	number of instances
adult	6	32561
raisin	3	900
rice	4	3810
default	16	30000
cancer	14	569
spam	55	4601
eeg	10	14980
churn	12	5000
ionosphere	33	351

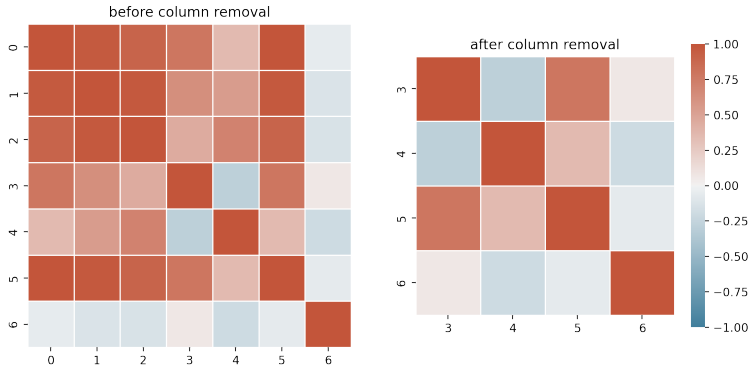
Table 1: datasets used in the project

(Task 3.1) Our proposal for the stopping rule is as such: we keep track of mean absolute error between true labels and predicted ones. Training process stops if the difference between current iteration's error and previous one is smaller than a certain tolerance parameter (e.g. 10^{-4}) for 10 iterations in a row. (essentially, if the progress is halts, training stops)

(Task 3.2) During training, every dataset is split into train/test data, where training data is 80% of the instances. Figure 2 presents average balanced accuracy measured on test data, for each dataset and each of the three algorithms trained on the same 5 train/test splits. For Stochastic Gradient Descent batch size of 64 was used, while the other algorithms were given the whole training dataset in a single batch.

IWLS performed best on 6 out of those 9 datasets, including all 3 of the smaller datasets. SGD was victorious only once, by a landslide to IWLS. There was one dataset, for which neither of our algorithms performed any better than random.

Figure 1: rice dataset correlation matrices before and after removal of highly correlated columns



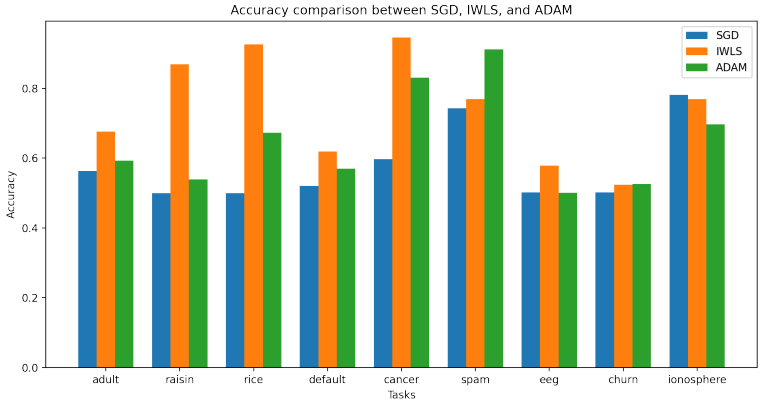


Figure 2: balanced accuracy for different datasets, for different optimization algorithms

2 Convergence analysis

Log-likelihood convergence in logistic regression is a pivotal aspect indicating the model's optimization progress.For this reason convergence analysis of log-likelihood function has been performed on training data for each optimization algorithm and for each dataset. Figures below depict log-likelihood function throughout iterations, coloured by learning rate parameter. Three figures are depicted (instead of all 9), as they are representative of the three classes of results we identified.

Figure 3 represents best case scenario, where each algorithm seems to converge without difficulty.

Figure **??** shows, that for some datasets its more difficult to achieve convergence for SGD or ADAM optimizers. log-likelihood for SGD seems to be able to converge only when learning rate parameter reaches 10^{-6} (purple line), whilst ADAM usually prefers larger learning rates.

Figure **??** shows that some datasets prove difficult for SGD to converge with respect to log-likelihood. Nevertheless, one can notice that the troughs of the function appear higher and higher as the training progresses.

Log-likelihood plots for SGD seem messy, as if it was unable to learn in some cases. At the same time, IWLS plot looks exactly the same for each of the 9 datasets tested. This result should not be surprising, as IWLS is an optimization technique tailored for logistic regression, aimed at estimation of the maximum likelihood parameters of logistic regression - that's exactly why the plots are so unambiguous. as for ADAM optimizer, it utilizes momentum mechanism, which could be the reason as to why in some cases the plot is "spiky" - it seems as if ADAM overshoots the target, then overcorrects its mistake back and forth, eventually converging. What is also interesting, is the fact that both IWLS and ADAM seem to do better with larger learning rates, while SGD performs acceptable only after learning rate drops, sometimes by several orders of magnitude, compared to the other two.

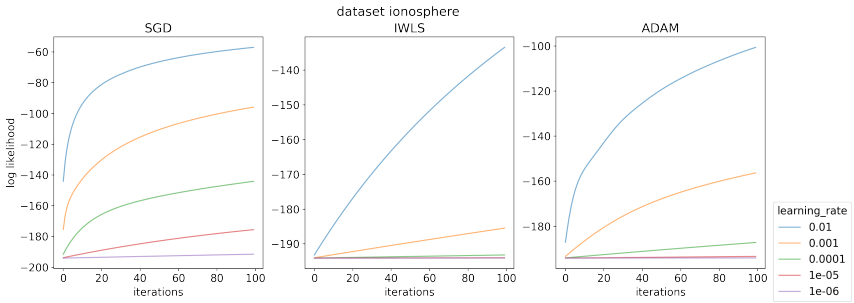


Figure 3: Best case scenario, all three algorithms seem to converge

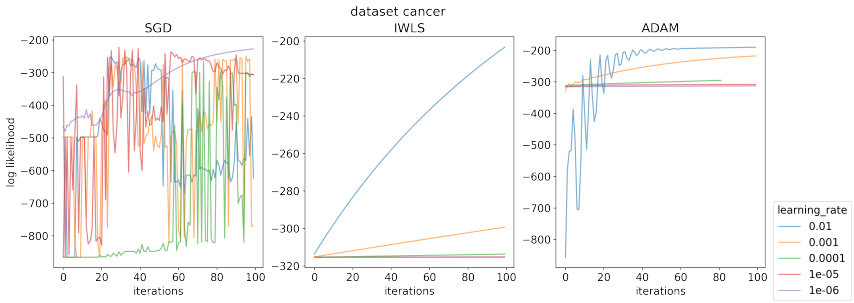


Figure 4

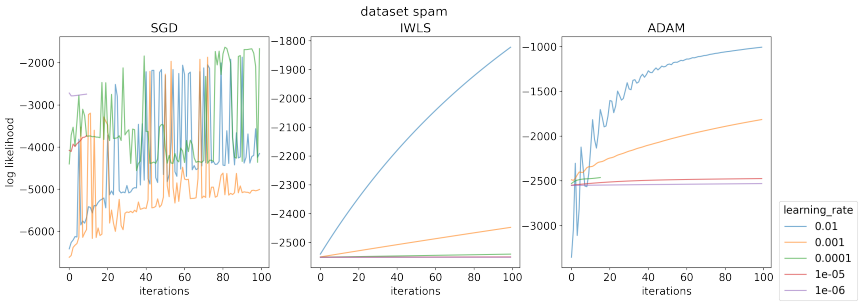


Figure 5

3 Comparison of classification performance

The goal of this task, was to compare the performance of Logistic Regression with SGD, IWLS and ADAM optimization algorithms with other popular models: LDA, QDA, Decision Tree and Random Forest.

In order to compare them in a fair manner, each dataset was split into train and test datasets five times. The results were averaged over those five splits for each dataset. The metric, displayed in figure 6, was balanced accuracy.

Figure 6 represents the comparison between those models. The data shows, that overall, the already implemented models were better than our logistic regression with different optimization algorithms. However, IWLS algorithm matched the results of the sklearn's implementations in 6 out of 9 datasets. SGD performed well only on one dataset - spam. On other datasets it achieved nothing better, than the accuracy of a random model. ADAM algorithm was in-between, as it achieved better results than SGD but more often than not was it worse than IWLS.

The IWLS algorithm was the best out of the three implemented by us for logistic regression. Still, it did not surpass the ones already implemented by sklearn.

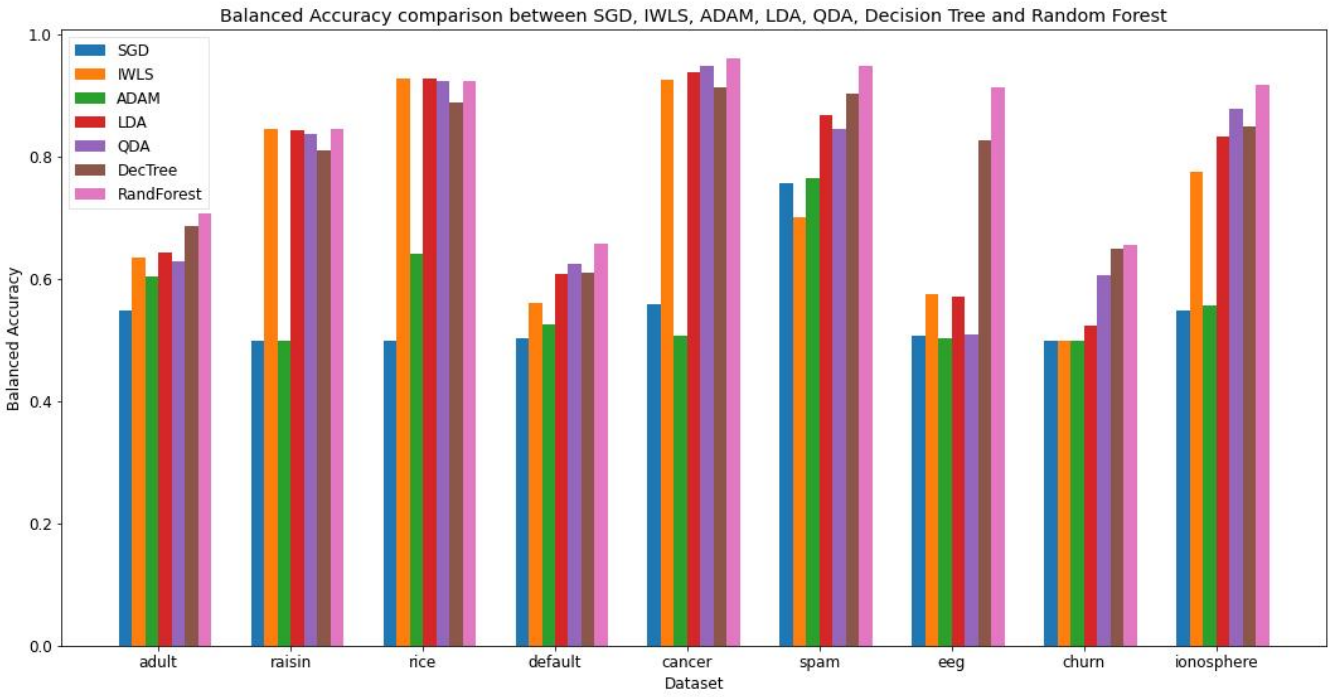


Figure 6: Comparison of the Logistic Regression models and other popular models

4 Comparison of classification performance of models with and without interactions

In this final chapter we have compare each model of logistic regression with and without the interactions. The models with interactions simply include, in addition to the original input variables, products between the features:

$$X_1 \times X_2, X_1 \times X_3, ..., X_2 \times X_3,.$$

In the end, if the initial data has n features then by performing the interactions it has $n + (n * (n - 1))/2$ features. Because of that, we compare the results only on the datasets with fewer than 10 features.

The same as in previous chapters:

Each dataset was split into train and test datasets five times, for a fairer comparison. The splits were the same for every model. The results were averaged over those five splits. The metric used was balanced accuracy.

There are 6 bars per dataset in figure 7, Each bar shows balanced accuracy of the particular mode for each dataset. On adult dataset, adding interactions improved every model by a little. On raisin dataset however, the improvement was much higher for SGD. The IWLS and ADAM improvements stayed approximately the same. The surprising part is the last dataset - rice. This time, interactions did not help in model performance, in fact IWLS and ADAM algorithms achieved worse metrics. For the SGD, balanced accuracy stayed the same.

Therefore, we cannot completely come to a consensus that the interactions in the models increase the effectiveness of the models.

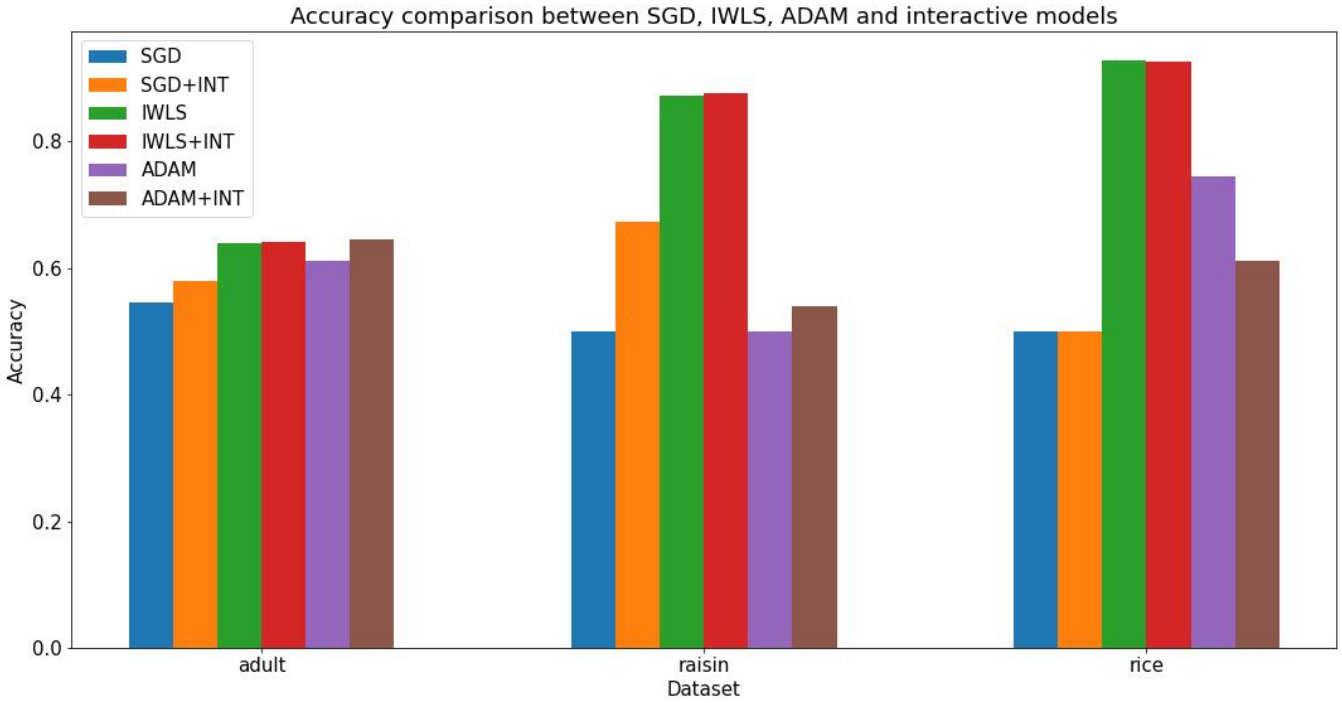


Figure 7: Comparison of the Logistic Regression models and interaction models