Advanced Machine Learning Project 1

Łukasz Tomaszewski
Patryk Rakus
Michał Tomczyk

March 2024

Contents

1 Introduction

The task of this project was to implement and compare performance of three different optimization algorithms for logistic regression: Iterative Reweighted Least Squares (IWLS), Stochastic Gradient Descent (SGD) and Adaptive Moment estimation (ADAM). We have implemented the methods in question from scratch, using Python and tested them on different dataset corresponding to binary classification problem. We have proposed a stopping rule for the algorithms, analysed their convergence depending on the number of iterations, compared their performance with each other and with 4 alternative popular classification methods and compared their performance on models trained with and without interactions between variables.

2 Methodology

2.1 Data

To properly check the performance of implemented algorithms, we have selected 9 binary classification datasets. Six of them had more than 10 variables (and were considered large datasets) and three had at most 10 (and were considered small). The following datasets were selected:

- **League of Legends Diamond Games (First 15 Minutes)** (large dataset) - a dataset consisting of team statistics on the first 15 minutes of LoL matches, with the task of predicting their outcomes.
- **Pollen** (small dataset) - binarized version of a synthetic dataset describing pollen.
- **Heart Attack Analysis & Prediction Dataset** (large dataset) - a dataset for heart attack classification based on medical information.
- **Spambase** (large dataset) - the classification task for this dataset is to determine whether a given email is spam or not.
- **Ionosphere** (large dataset) - the targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere.
- **Phoneme** (small dataset) - The aim of this dataset is to distinguish between nasal (class 0) and oral sounds (class 1).
- **Phishing websites** (large dataset) - the dataset consisting of information about different websites, with the task of predicting whether a site is used for phishing. The variables describe the properties of the URL of a website (e.g. whether the address contains the '' sign), its HTML and Javascript content (e.g. if the right ckick is disabled) and information about the domain (e.g. its age).
- **Banknote Authentication** (small dataset) - a dataset with the task of predicting whether a banknote is fake, based on features extracted from a wavelet transform of the image of a banknote
- **Climate Model Simulation** (large dataset) - a dataset focused on predicting climate model simulation outcome (whether the model has crashed or not), based on various parameters of said model.

The dataset and preprocessing necessary for using them are summarised in a table 1 (when dropping correlated columns, Pearson correlation was calculated and columns with more than 0.8 correlation were dropped).

Table 1: Datasets used

Name	**Noteworthy processing**	**Shape**
League of Legends Diamond Games (First 15 Minutes) [1]	Dropped id column and columns with variance equal to 0. Dropped duplicate rows. Dropped correlated columns.	48632 x 12
Pollen [3]	Dropped correlated columns, mapped target variable from (N, P) to (0,1).	3848 x 5
Heart Attack Analysis & Prediction Dataset [7]	Dropped correlated columns.	303 x 14
Spambase [2]	Dropped correlated columns	4601 x 56
Ionosphere [8]	Dropped correlated columns, mapped target variable from (b, g) to (0, 1)	351 x 32
Phoneme	Mapped target variable from (1,2) to (0, 1)	5404 x 6
Phishing Websites [6]	Dropped correlated columns, decoded target variable from binary to numeric and mapped from (-1, 1) to (0, 1)	11055 X 28
Banknote Authentication [4]	Decoded target variable from binary to numeric and mapped from (1, 2) to (0, 1)	1372 X 5
Climate Model Simulation [5]	Decoded target variable from binary to numeric	540 X 21

2.2 Early stopping

For an early stopping rule, we check whether the improvement in log-likelihood is larger than 0.0001, on the training set, for each in the last 10 epochs (number of epochs can be changed with a hyper-parameter). If this condition is not satisfied, the model loads the best coefs encountered during training and ends training.

2.3 Training and evaluation

Each model was trained and evaluated (with balanced accuracy metric) using cross-validation with 5 splits. Training and test data are standardized for each split using scikit-learn's StandardScaler. The only exception is convergence analysis, where each model was ran only once, on the entire dataset.

3 Convergence analysis

As mentioned in the above section 2.3, for the convergence analysis each model was trained once on the whole dataset. We have analysed the log-likelihood on the training data over the number of epochs (and, in the same time, the number of epochs for which the learning stopped). The results are visualised in the figure 1. It can be observed that in general IWLS needs the least number of iterations to perform early stopping, while still achieving satisfactory log-likelihood. ADAM converges the most slowly, its log-likelihood over iterations function is the most smooth and the stopping rule is not called for 5 datasets. For SGD, the convergence varies the most - for some dataset it needs less than 50 epochs to call the stopping rule, while for other it converged after few hundred.

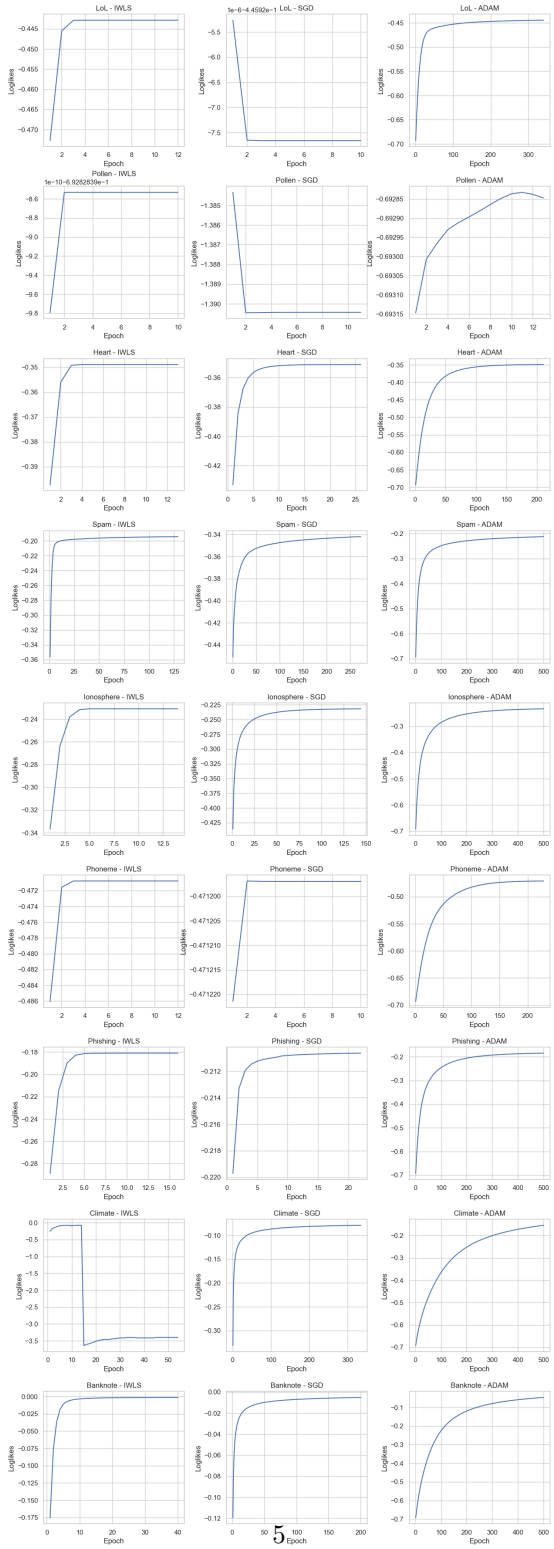

Figure 1: Convergence of log-likelihood over the number of iteration for each dataset and algorithm

4 Comparison of classification performance

The performance of each optimization algorithm varies for each dataset. Thus, it is hard to generalize which method achieves the best results. As well, the implemented methods sometimes perform better than other popular methods and sometimes worse. It is worth to note the Spam dataset, for which SGD achieves much lower results than other implemented optimizers. For LoL, Heart, Phoneme and Phishing datasets IWLS, SGD and ADAM perform nearly identical. Phishing dataset is also noteworthy, as QDA achieves significantly lower balanced accuracy than other methods. For Climate dataset IWLS is very unstable (its performance varies greatly between each split), while for Pollen dataset SGD is almost completely stable. It should not be surprising that for most datasets Random Forest Classifier performs better than Decision Tree. Especially for LoL dataset Decision tree achieves visibly lower score than any other algorithm. Between LDA and QDA, similarly to ADAM, IWLS and SGD different datasets different methods perform better.

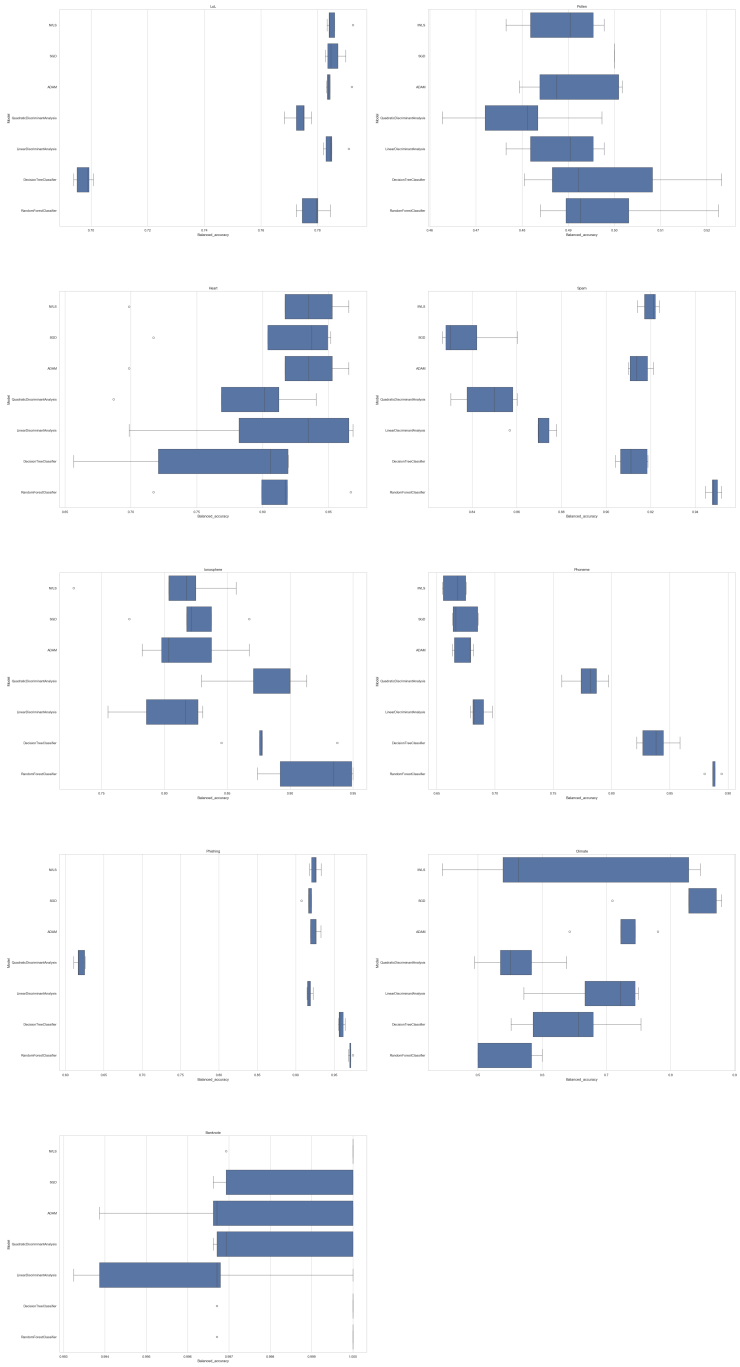

Figure 2: Performance of different classification algorithms on different datasets

5 Comparison of classification performance of models with and without interactions

In the case of small datasets we compared two versions of the logistic regression. One version used original input variables and the other used in addition products of two variables. For Pollen dataset adding interactions worsened the average results in most cases. The only exception is SGD model, for which balanced accuracy stayed the same and was equal 0.5 for all models trained on different train test splits. In the case of IWLS one model performed better when interactions were introduced and it was the best score achieved for this dataset. On Phoneme dataset all methods performed better when interactions were added. For all splits IWLS and ADAM achieved higher balanced accuracy when the models were fitted on dataset with interactions and those results were better than the best result obtained when original dataset was used. In the case of Banknote dataset the results of all models were more stable when interactions were applied. For IWLS and SGD single outliers were eliminated. The biggest improvement was seen for ADAM method, where the results on original dataset had the biggest standard deviation. After applying interactions all models obtained the perfect score (balanced accuracy equal to 1) on every train test split.

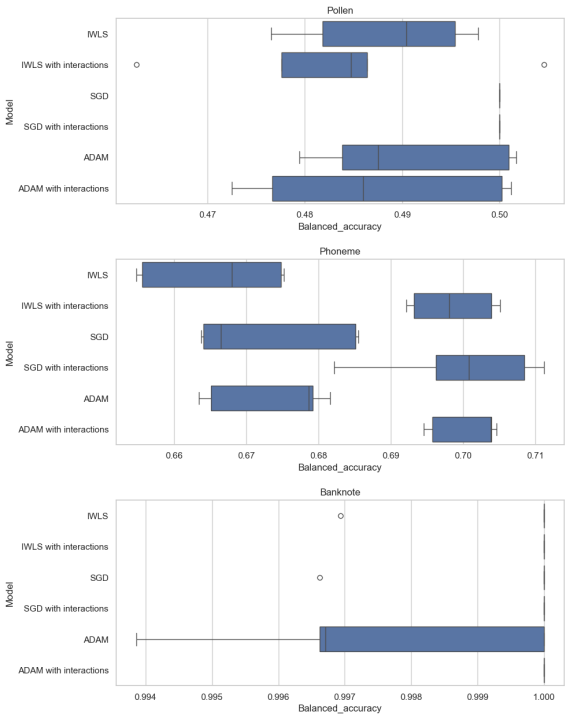

Figure 3: Comparison of performance of models with and without interactions for each small dataset and algorithm

References

[1] Fattori B. League of Legends Diamond Games (First 15 Minutes). Kaggle. https://www.kaggle.com/datasets/benfattori/league-of-legends-diamond-games-first-15-minutes/data.

[2] Reeber Erik Forman George Hopkins, Mark and Jaap Suermondt. Spambase. UCI Machine Learning Repository, 1999. DOI: https://doi.org/10.24432/C53G6X.

[3] Vanschoren J. pollen. OpenML. ID: 871.

[4] Volker Lohweg. Banknote Authentication. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C55P57.

[5] D. D. Lucas, R. Klein, J. Tannahill, D. Ivanova, S. Brandon, D. Domyancic, and Y. Zhang. Failure analysis of parameter-induced simulation crashes in climate models. *Geoscientific Model Development*, 6(4):1157–1171, 2013.

[6] Rami Mohammad and Lee McCluskey. Phishing Websites. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C51W2X.

[7] Rahman R. Heart Attack Analysis Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

[8] Wing S. Hutton L. Sigillito, V. and K. Baker. Ionosphere. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5W01B.