Warsaw University of Technology

Implementation and comparison of IWLS, SGD, ADAM

Advance Machine Learning

authored by

Filip Kucia (335724)

Szymon Trochimiak (307459)

Michał Taczała (303775)

supervisor

Katarzyna Woźnica

Warsaw 2024

Contents

1 Methodology

Remark: all of the calculations and plots were send/created in Weights and Biases (Wandb).

1.1 Small Datasets with a few variables

Heart Disease UCI Repository: A dataset for indicating fasting blood sugar ¿ 120 mg/dl with 14 health indicators but only 6 of them (age, trestbps, chol, thalach, oldpeak) are numerical, so they were included and considered small dataset. [6].

Apple Quality Kaggle: A dataset for assessing apple quality (good or bad) using 7 variables plus a target for quality [3].

Fertility UCI Repository: Dataset described with variables for predicting human fertility [5].

1.2 Big Datasets with many variables

Breast Cancer Wisconsin UCI Repository: Diagnostic dataset with 30 features for classifying tumors as malignant or benign [4].

Airline Passenger Satisfaction Kaggle: Contains 25 variables related to passenger experience for satisfaction or dissatisfaction. [1].

Water Quality Kaggle: Measures water quality (good or bad) through 25 chemical variables [9].

Predict Students Dropout UCI Repository: Predicts student dropout and academic success with 36 features [8].

Algerian Forest Fire UCI Repository: Involves meteorological data for predicting forest fires in Algeria [2].

Mine vs. Rock OpenML: Binary classification dataset with 60 sonar signal features for identifying mines and rocks [7].

1.3 Class imbalance within datasets

Worth to mention that 2 of 3 of our small datasets have high class imbalance. In Big datasets only Water quality has high class imbalance.

Dataset	**% of observations in class 1**	**% of observations in class 0**	**No. of Observations**
Small Datasets			
Heart Disease	14.85%	**85.15%**	303
Apple Quality	**50.1%**	49.9%	4000
Fertility	**88.0%**	12.0%	100
Big Datasets			
Breast Cancer	37.26%	**62.74%**	569
Airline Passenger Satisfaction	43.45%	**56.55%**	129880
Water Quality	11.41%	**88.59%**	7996
Predict Student Dropout	**67.88%**	32.12%	4424
Algerian Forest Fire	**56.38%**	43.62%	243
Mine vs. Rock	46.63%	**53.37%**	208

Table 1: Dataset Class Distribution and Size

1.4 Data Preprocessing Steps

For the **Water Quality Dataset**, non-numeric values in the 'is_safe' column were mapped to binary values. For the **Algerian Forest Fire Dataset**, labels in the 'Classes' column are stripped of whitespaces and standardized in case; missing values are also handled. In the **Student Dropout and Academic Success Dataset**, 'Graduate' and 'Enrolled' labels are grouped into a single class, and categorical labels in the 'Target' column are mapped to numerical values.

Aside from converting categorical variables into numeric values, we removed all columns that had correlation values higher than 0.7 with other columns. If there were any NaN's or string that could not be converted into floats and other non-interpretable values we removed them.

1.5 A Stopping rule

For all of the algorithms, we proposed the following stopping rule: Our function that stops the training: `should_stop_convergence`, is defined as:

$$\text{should\_stop\_convergence}(\Delta J, \epsilon) = \begin{cases} \text{True} & \text{if } |\Delta J| < \epsilon \\ \text{False} & \text{otherwise} \end{cases} \tag{1}$$

where:

- ΔJ is the change in the loss function between iterations.
- ϵ is the convergence threshold, with a default value of 1×10^{-6}.

In case the algorithm would not decrease the loss function, we applied another limit of 500 iterations maximum, even if the stopping rule is not fulfilled.

1.6 Balanced Accuracy

For all algorithms, we used a **Balanced Accuracy** is which calculated as the average of *sensitivity* and *specificity*. It is particularly useful in imbalanced datasets, where the prevalence of one class greatly exceeds the other. The formula is:

$$\text{Balanced Accuracy} = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{2}$$

where:

- *Sensitivity* (also known as the true positive rate or recall) is the proportion of true positives correctly identified by the model:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

- *Specificity* (also known as the true negative rate) is the proportion of true negatives correctly identified by the model:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

2 Convergence Analysis

For the convergence analysis, we measured the value of a log-likelihood function for every algorithm. The results from the Breast Cancer Dataset are as follows:

The value of the LogLikelihood decreases with the number of epochs for every optimizer. However, we can see that the values are very different. Another thing worth mentioning is the fact, that the log-likelihood value is not proportional to the obtained accuracy, as for IWLS we got a much higher likelihood value than for SGD which has achieved a worse performance.

3 Comparison of Classification Performance

We can see, that the performances of methods are very dependent on the dataset. Moreover, some methods perform better than others for different datasets. It's also worth mentioning, that some methods for some datasets have about 50% accuracy, which means that they don't perform better than randomly choosing target values.

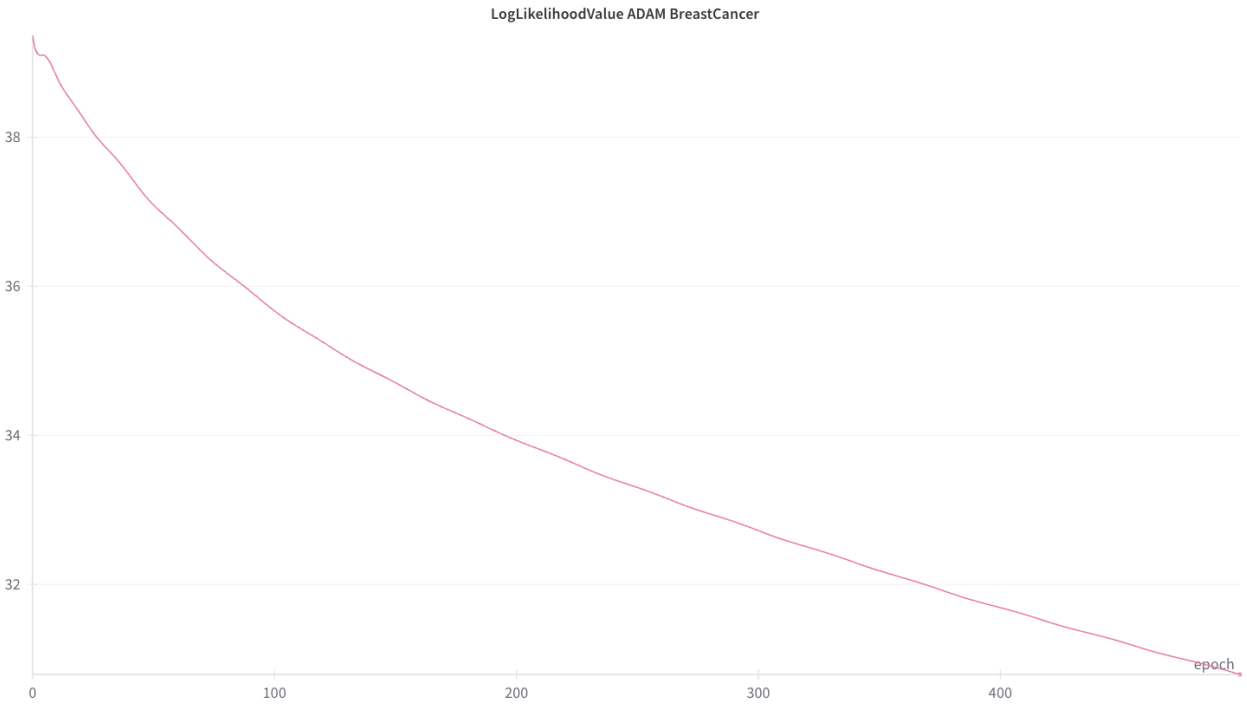

Figure 1: ADAM LogLikelihood - Breast Cancer Results

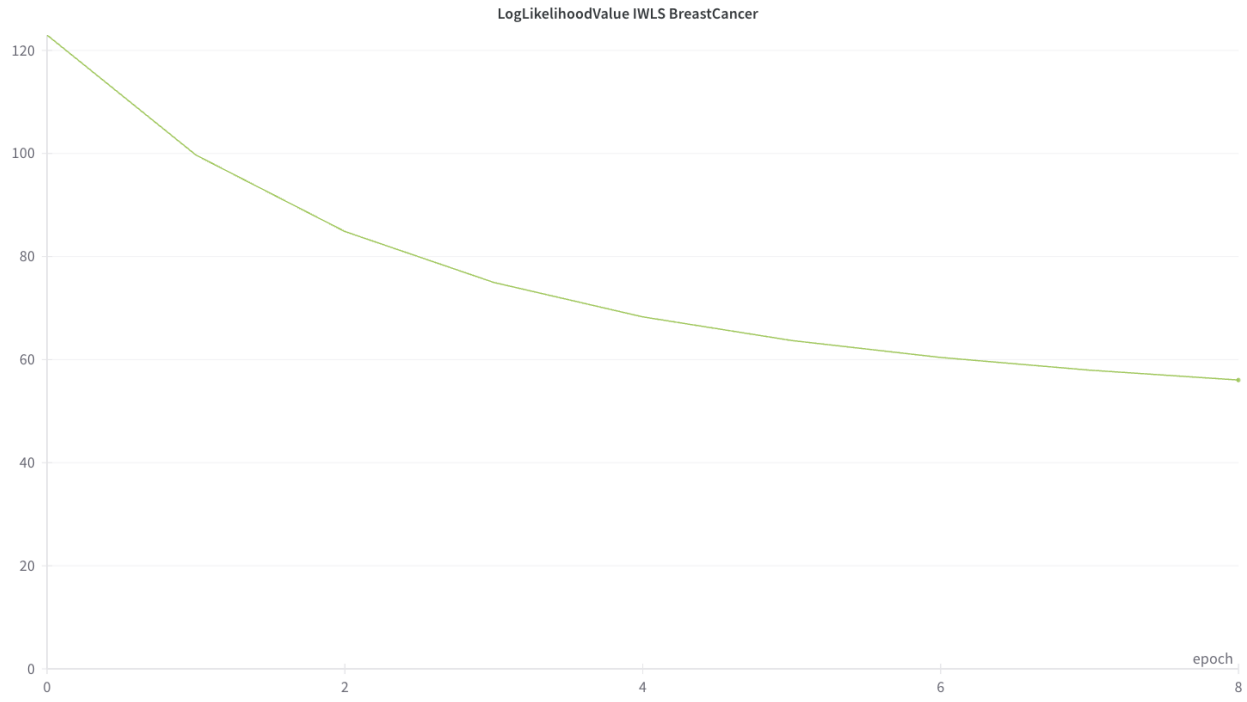

Figure 2: IWLS LogLikelihood - Breast Cancer Results

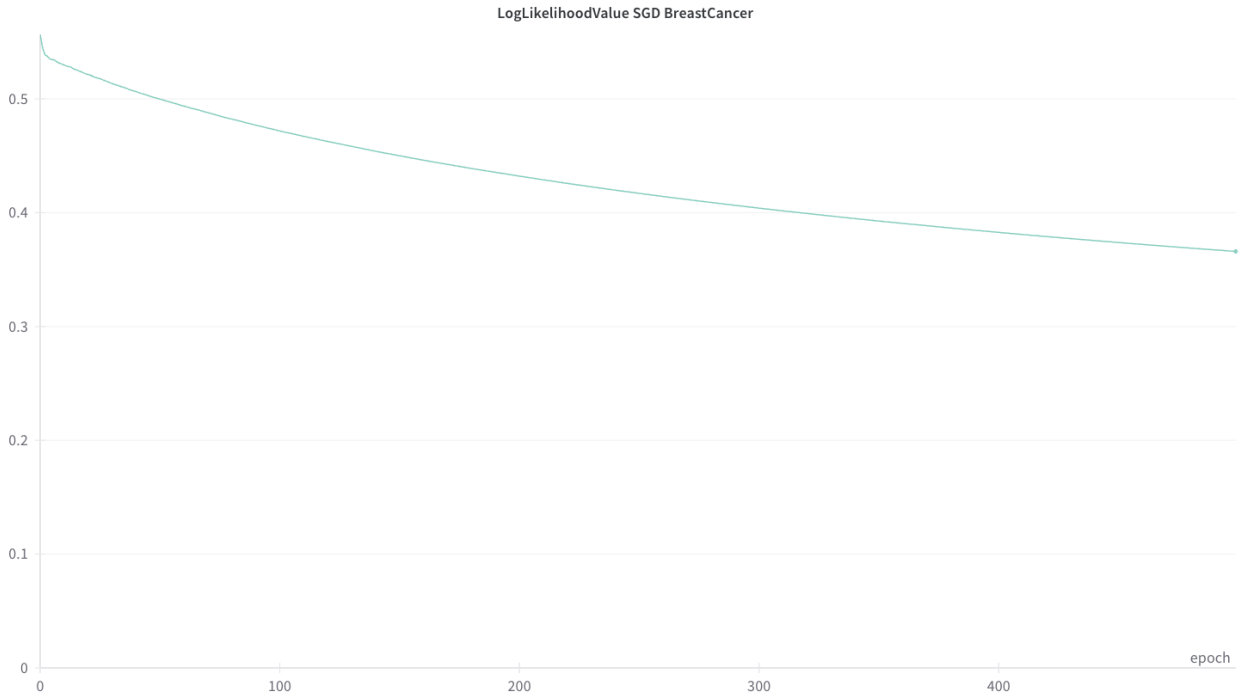

Figure 3: SGD LogLikelihood - Breast Cancer Results

Dataset	**IWLS**	**SGD**	**ADAM**	**LDA**	**QDA**	**Decision Tree**	**Random Forest**
HeartDisease	0.50	0.82	0.56	0.87	0.88	0.77	0.87
AppleQuality	0.77	0.75	0.74	0.75	0.86	0.80	0.89
Fertility	0.50	0.90	0.50	0.90	0.85	0.90	0.90
BreastCancer	0.92	0.89	0.72	0.96	0.96	0.89	0.93
Airline	0.87	0.85	0.77	0.87	0.85	0.95	0.96
WaterQuality	0.65	0.91	0.64	0.91	0.88	0.96	0.97
Dropout	0.77	0.71	0.66	0.81	0.78	0.75	0.83
Fires	1.00	0.65	0.88	0.90	0.44	0.98	0.96
Sonar	0.63	0.39	0.48	0.61	0.56	0.63	0.66

Table 2: Comparison of test balanced accuracies

4 Comparison of Logistic Regression Models with and Without Interactions

Dataset	IWLS	IWLS+INT	SGD	SGD+INT	ADAM	ADAM+INT
Heart Disease	0.50	0.54	0.82	0.85	0.56	0.5
Apple Quality	0.77	0.83	0.75	0.83	0.74	0.84
Fertility	0.5	0.65	0.9	0.9	0.50	0.55

Table 3: Dataset Method Accuracy Comparison

We can see that adding interactions improves the obtained accuracies for almost every scenario. The only one that doesn't fit this description is the ADAM for Hearth Disease dataset. It's probably because the accuracy is about 0.5 so we got a random-choice accuracy anyway, and slightly better results for interactions don't change anything because it's still about the same value. For other datasets and optimizers adding interactions works really well and improves the accuracy of the given optimizer.

5 Summary

The main conclusion from this task is that the results are very dependent on the chosen dataset. Generally despite ADAM being a combination of other optimizers, it doesn't perform better than IWLS and SGD. The `IWLS` algorithm has a notably wide range of performances, from as low as 0.50 in HeartDisease and Fertility to a perfect 1.00 in Fires. Such variation suggests that the suitability of `IWLS` highly depends on the specific characteristics of the dataset.

`SGD` is generally more consistent across the datasets but has a dip in performance in Sonar.

`ADAM` does not excel in these datasets as it's generally outperformed by `SGD`, especially on small datasets, and doesn't achieve the highest accuracy on any dataset.

For these methods, the results were usually quite similar. We can't say that one method outperformed the others in every possible field, but IWLS performed statistically the best.

No algorithm's the best for every task and every dataset.

References

[1] Airline Passenger Satisfaction. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.

[2] Algerian Forest Fires Dataset. https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++.

[3] Apple Quality. https://www.kaggle.com/datasets/nelgiriyewithana/apple-quality.

[4] Breast Cancer Wisconsin (Diagnostic) Data Set. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

[5] Fertility Data Set. https://archive.ics.uci.edu/ml/datasets/Fertility.

[6] Heart Disease Data Set. https://archive.ics.uci.edu/ml/datasets/heart+disease.

[7] Mine vs. Rock Dataset. https://www.openml.org/d/40.

[8] Predict Students Dropout and Academic Success. https://archive.ics.uci.edu/ml/datasets/predict+students+dropout+and+academic+success.

[9] Water Quality. https://www.kaggle.com/datasets/mssmartypants/water-quality.