Advanced Machine Learning - project 1

Klaudia Gruszkowska, Zofia Łągiewka, Jacek Zalewski

Warsaw University of Technology

April 2024

Presentation Overview

Introduction

The main goal of this project:

- implement optimization algorithm for parameter estimation in logistic regression
  - IWLS
  - SGD
  - Adam
- perform few comparisons:
  - with baseline models
  - with and without interactions
  - convergence analysis

Datasets

Table 1: Details of small datasets

Name	**Features**	**Rows**
banknote_authentication	4	1372
fertility	10	100
magic_gamma_telescope	10	19020
spambase	57	4601
taiwanese_bankruptcy_prediction	95	6819
connectionist_bench_sonar_mines_vs_rocks	60	208
ionosphere	34	351
algerian_forest_fires	14	244
waveform_database_generator_version_1	21	5000

Preprocessing

The selected datasets were preprocessed. The following operations were performed:

- Deletion of null values
- Mapping categorical values
- Removing collinear variables

Additionally, one dataset requires more preprocessing because of not well-prepared labels, and for Waveforms dataset we choose to use only 2 classes from the target table.

Stopping condition

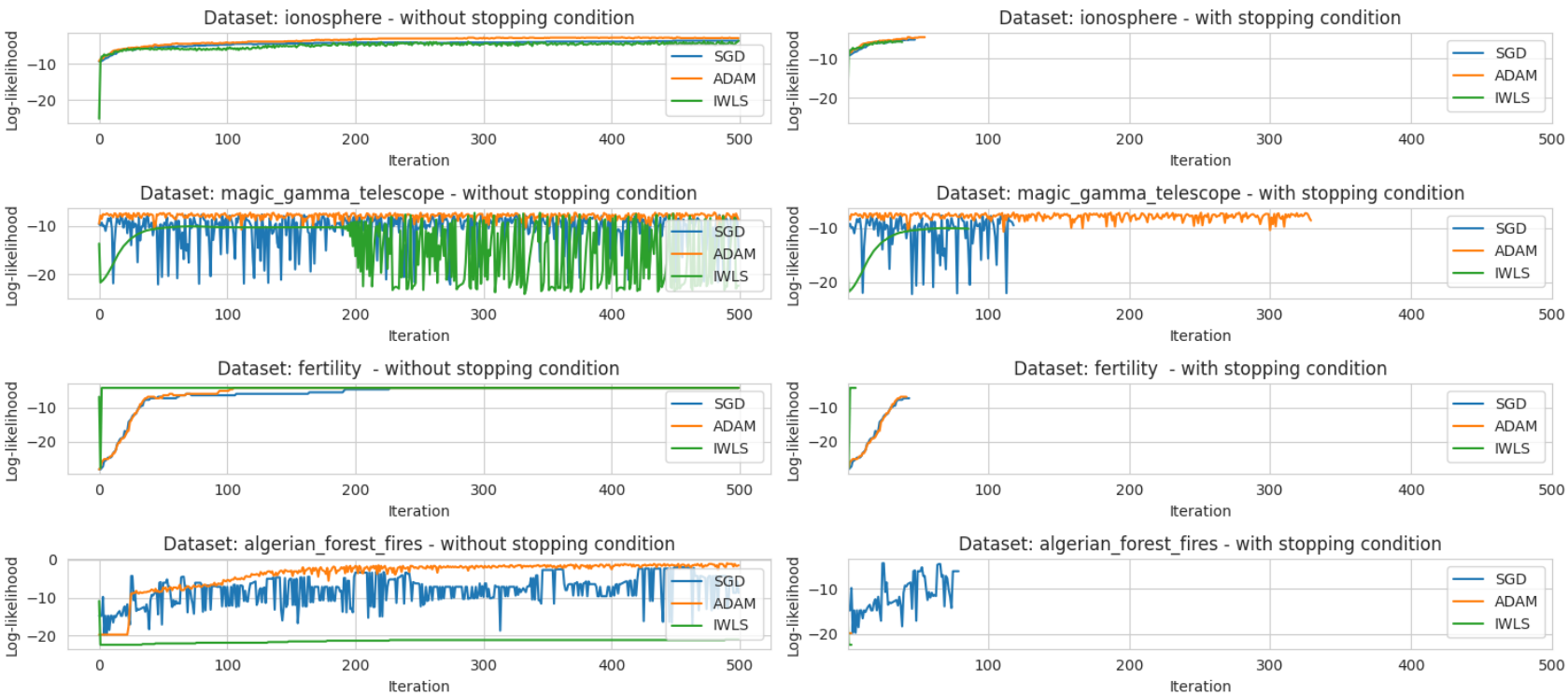
The main stopping condition is:

- number of iterations - by default set to 500

Moreover, the training process is stopped earlier if:

- the log-likelihood does not increase in some given number of next iterations - by default set to 5

Convergence analysis



Performance measure

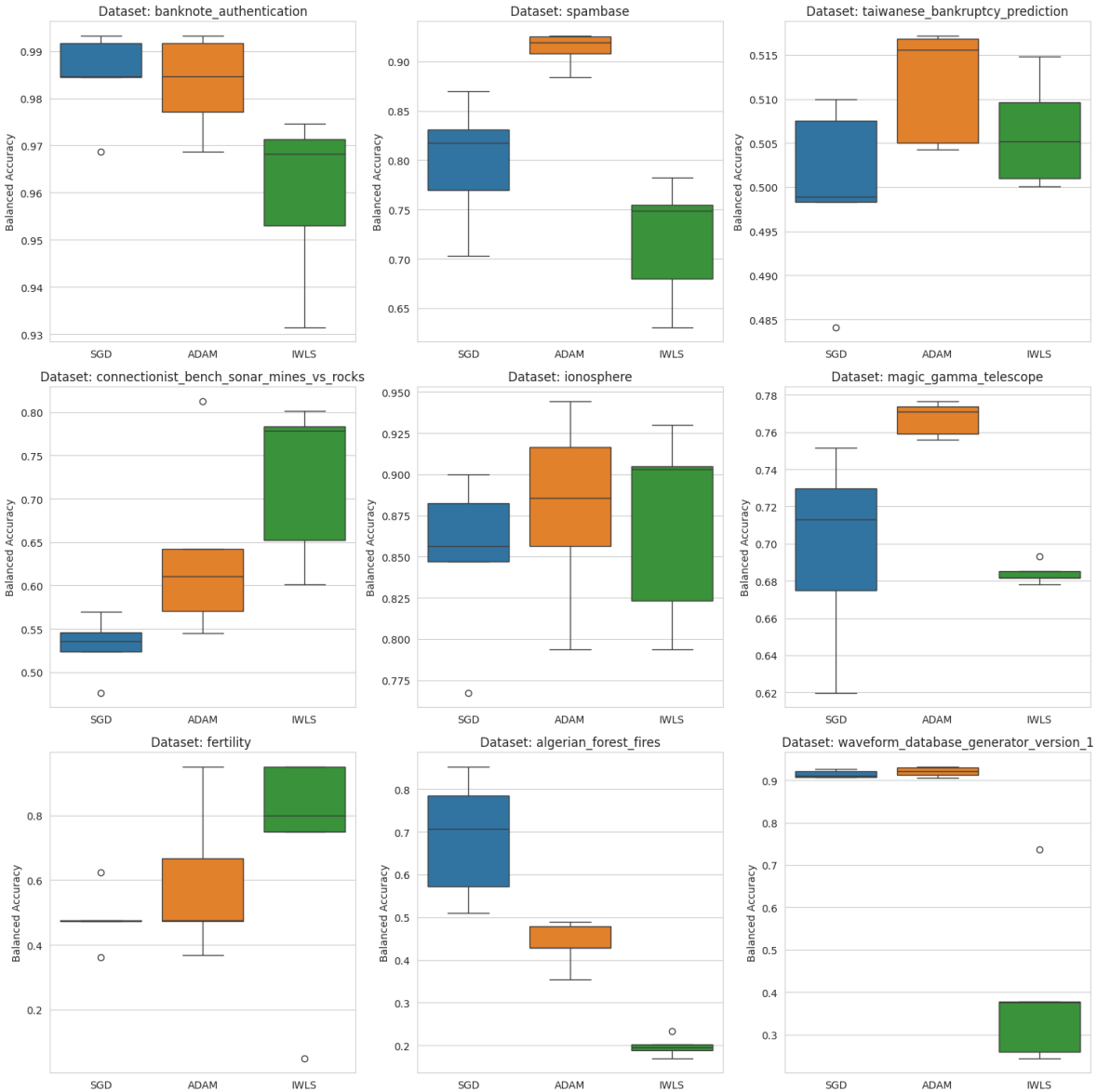
$$Balanced\_accuracy = \frac{(Sensitivity + Specificity)}{2}$$

where:

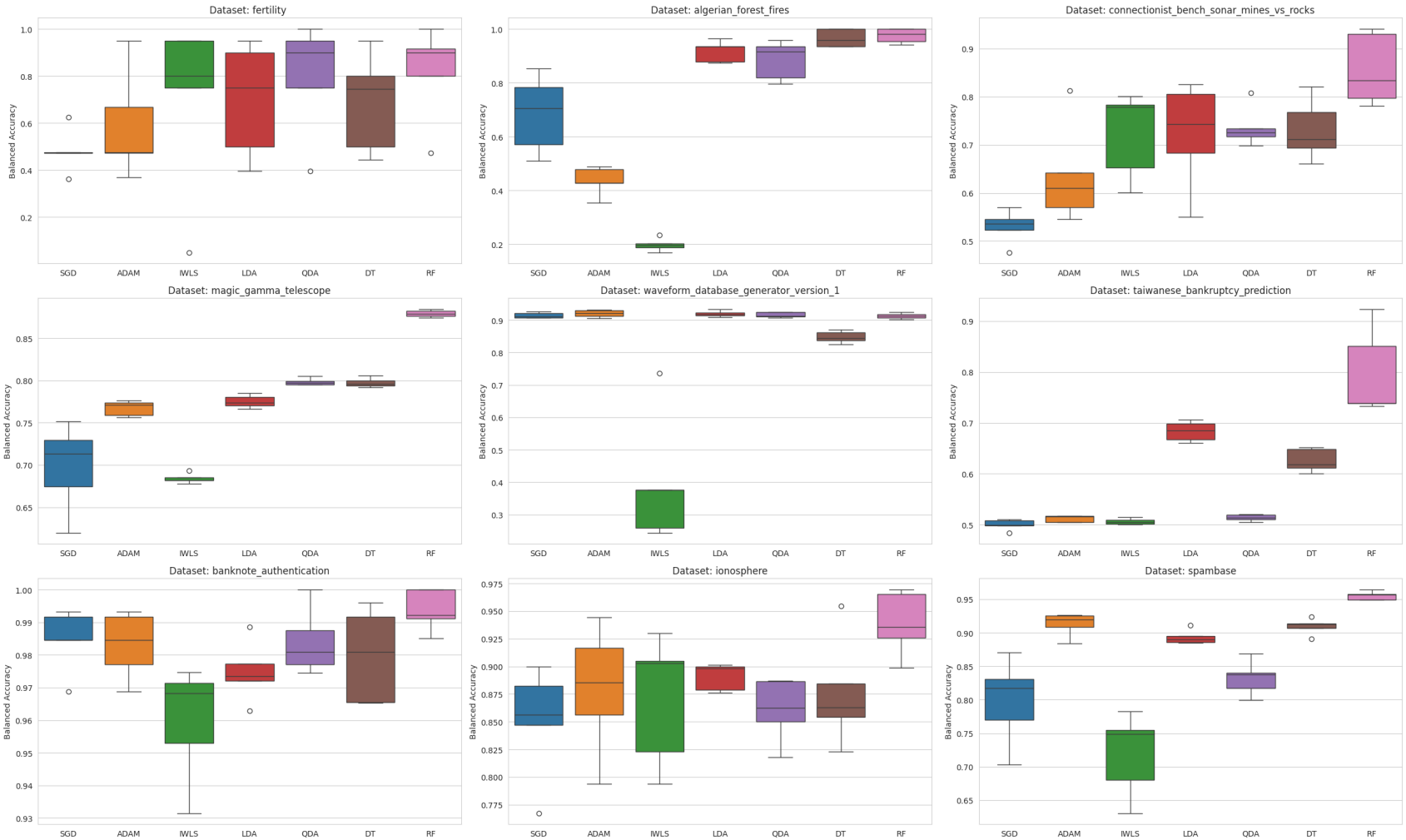
- Sensitivity - The "true positive rate" – is the percentage of positive cases the model can detect
- Specificity - The "true negative rate" – the percentage of negative cases the model can detect

Each model was evaluated with an average of 5 train_test splits with the maximum iterations set to 500.

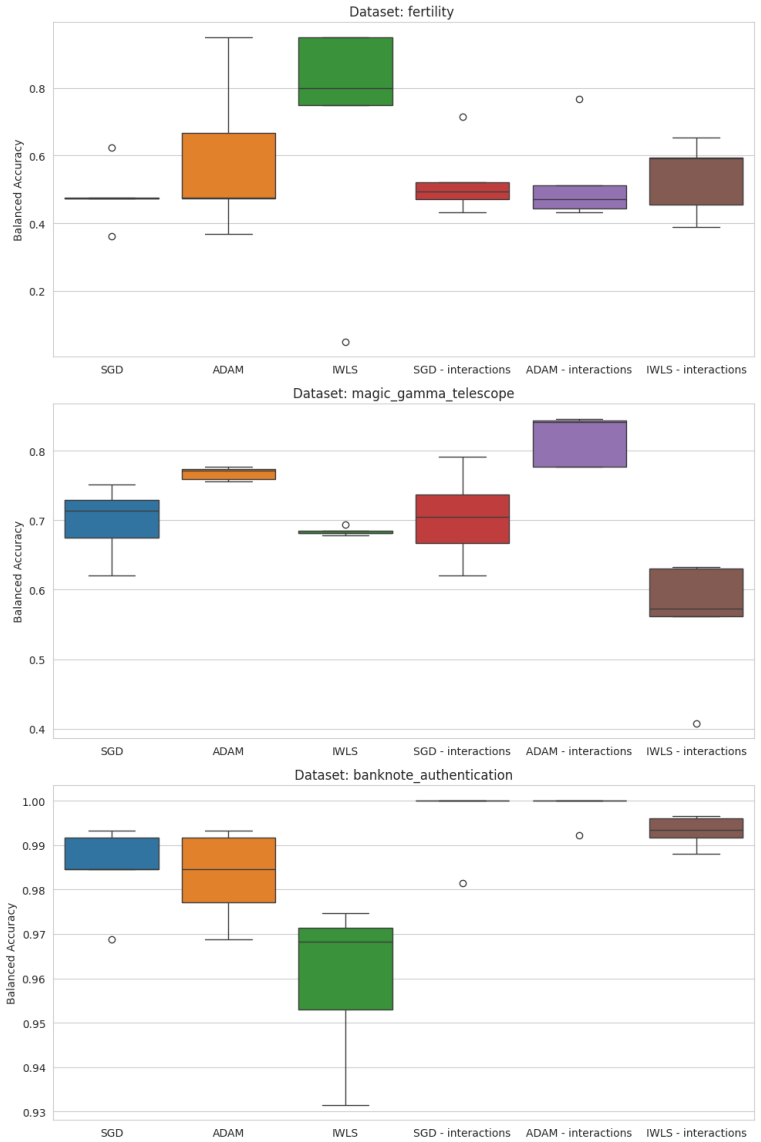
Comparison of optimizers



Comparison of classification performance



Comparison of models with and without interactions



The end

Thank you for your attention!