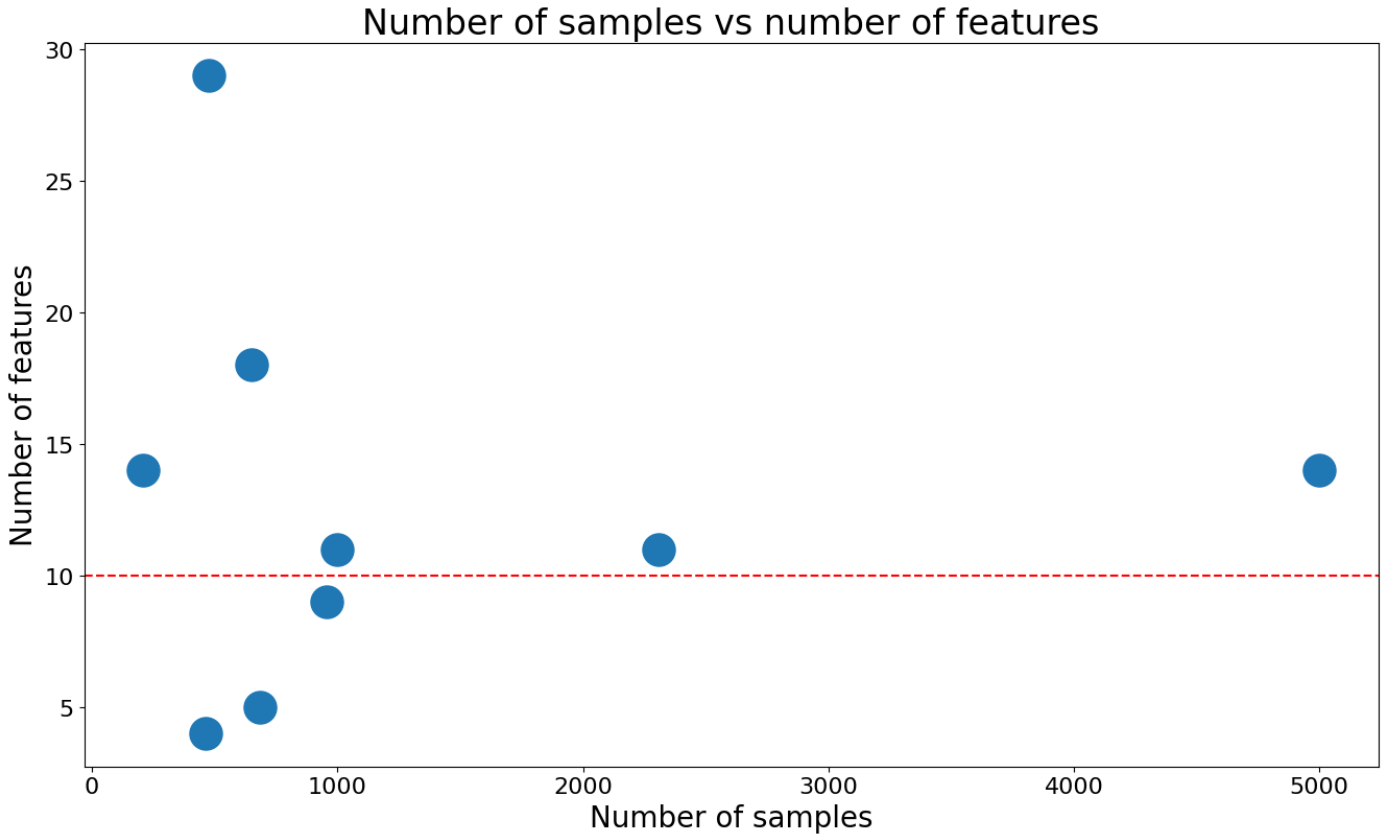
Advanced Machine Learning

MICHAŁ GROMADZKI

Datasets


Number of samples vs number of features
Number of features
Number of samples
30
25
20
15
10
5
0
1000
2000
3000
4000
5000

Remove collinear variables

- Algorithm for removing collinear variables:
- Algorithm is iteration-based, in each iteration 1 variable can be removed
- In each iteration:
  - Calculate variance inflation factor (VIF) for all variables
  - Select all variables with VIF higher than TH = 5
  - Delete variable with the highest VIF
- Algorithm stops when there is no variables with VIF over TH

Implementations

All optimization algorithms for parameter estimation in logistic regression have been implemented as classes in Python

Attributes:

- **max_iter** (int) - Maximum number of iterations of the optimizer
- **tol** (float) - Minimum value of Frobenius norm of difference in parameters between iterations, used to determine convergence
- **coef_** (array) - Array containing parameters
- **intersections** (bool) - Whether to also include intersections of provided variables

Methods:

- **_add_intersections()** - Adds intersections to the provided variables
- **fit()** - Trains the model, return history of training and number of iterations
- **predict_proba()** - Predicts probabilities based on the provided data
- **predict()** - Rounds the probabilities to class labels

Stopping rule

All algorithms have the same stopping rule to ensure fair comparison. At the end of each iteration of the algorithms the following value is computed.

$$diff = norm(coef - prev\_coef)$$

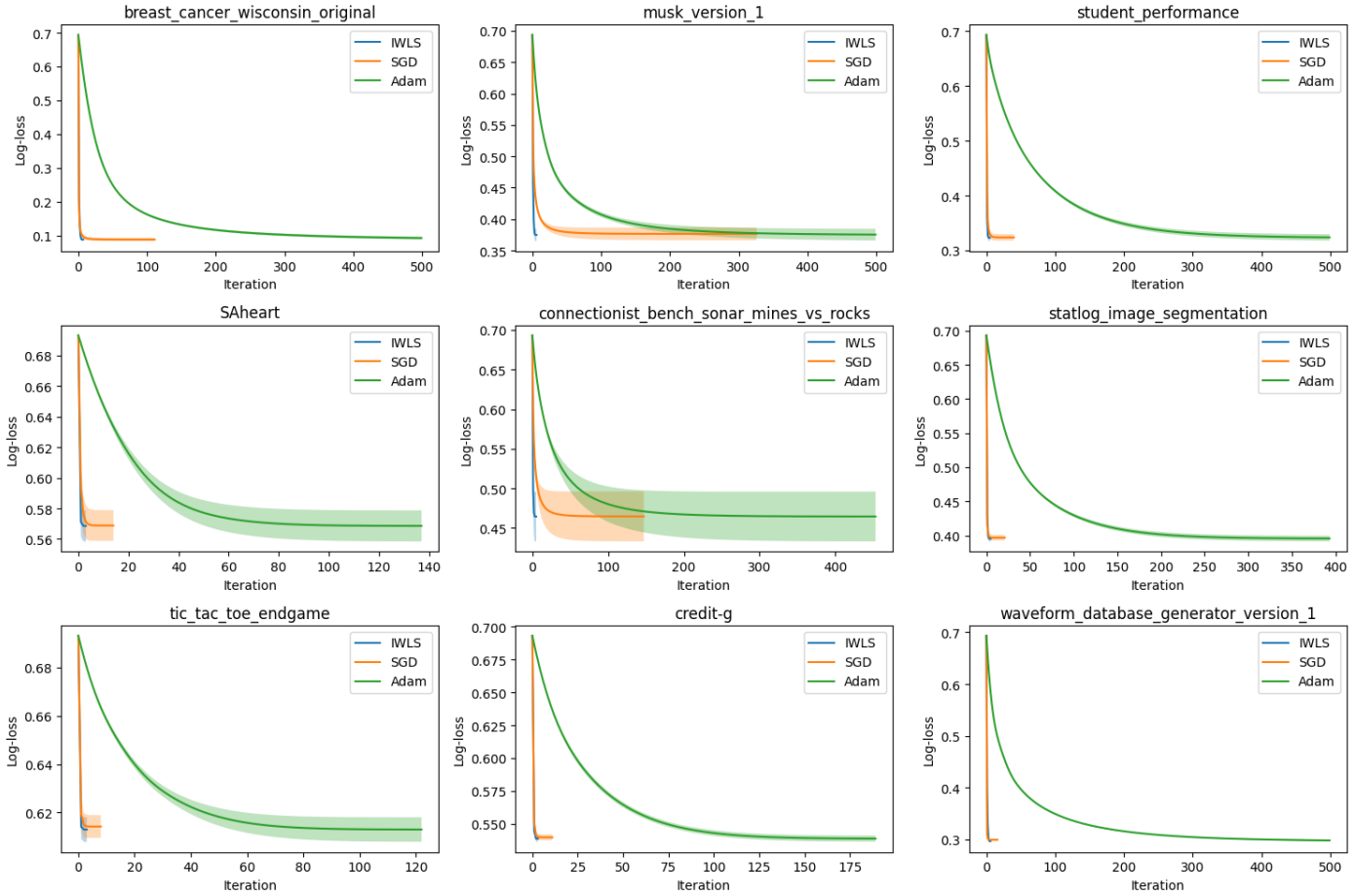
where:

- norm - The Frobenius norm
- coef - Parameters in the current iteration
- prev_coef - Parameters in the previous iteration

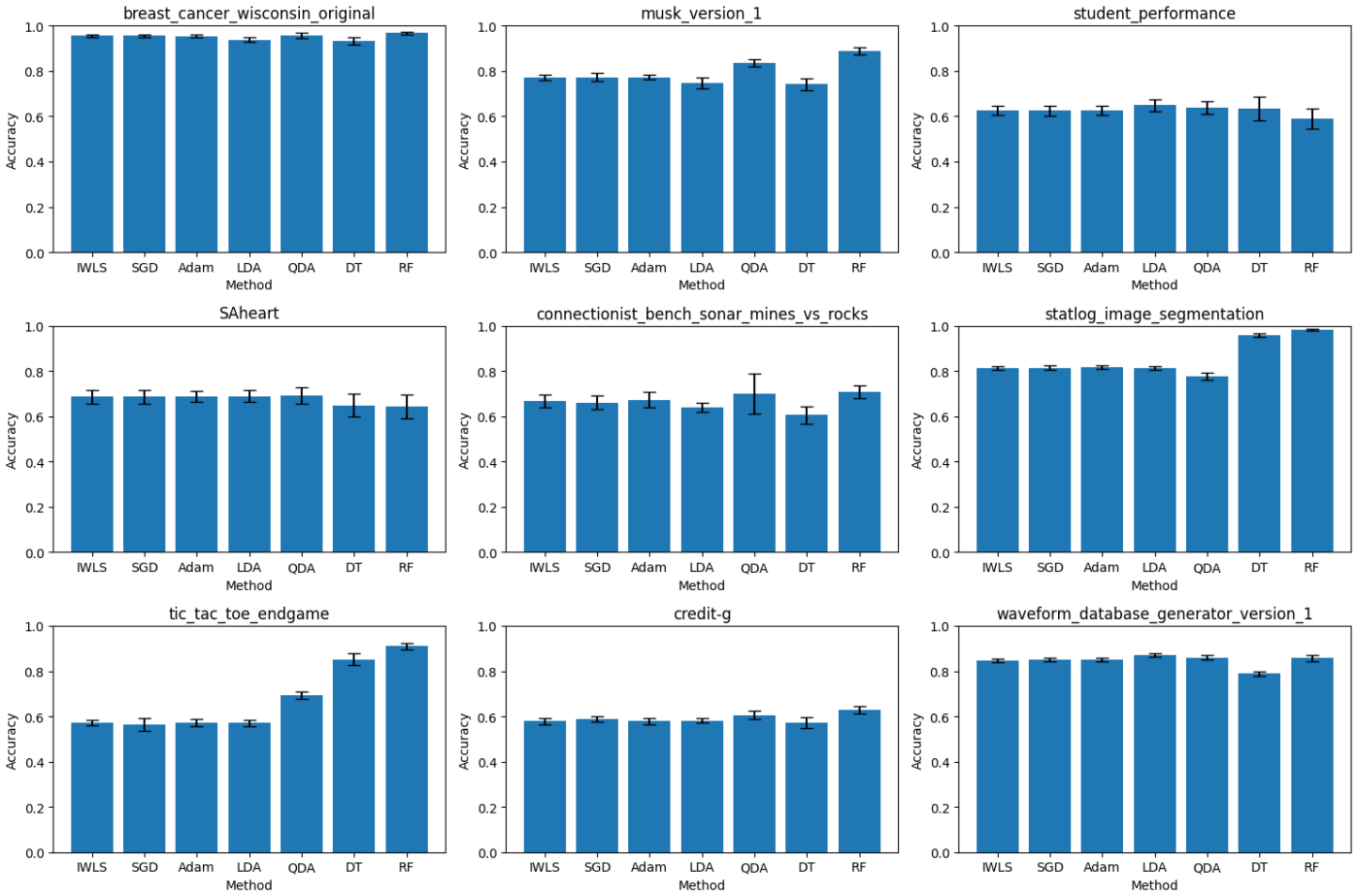
$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i}^{m}\sum_{j}^{n}|a_{ij}|^2}$$

If the $diff$ is smaller than the pre-set tol value the training stops.

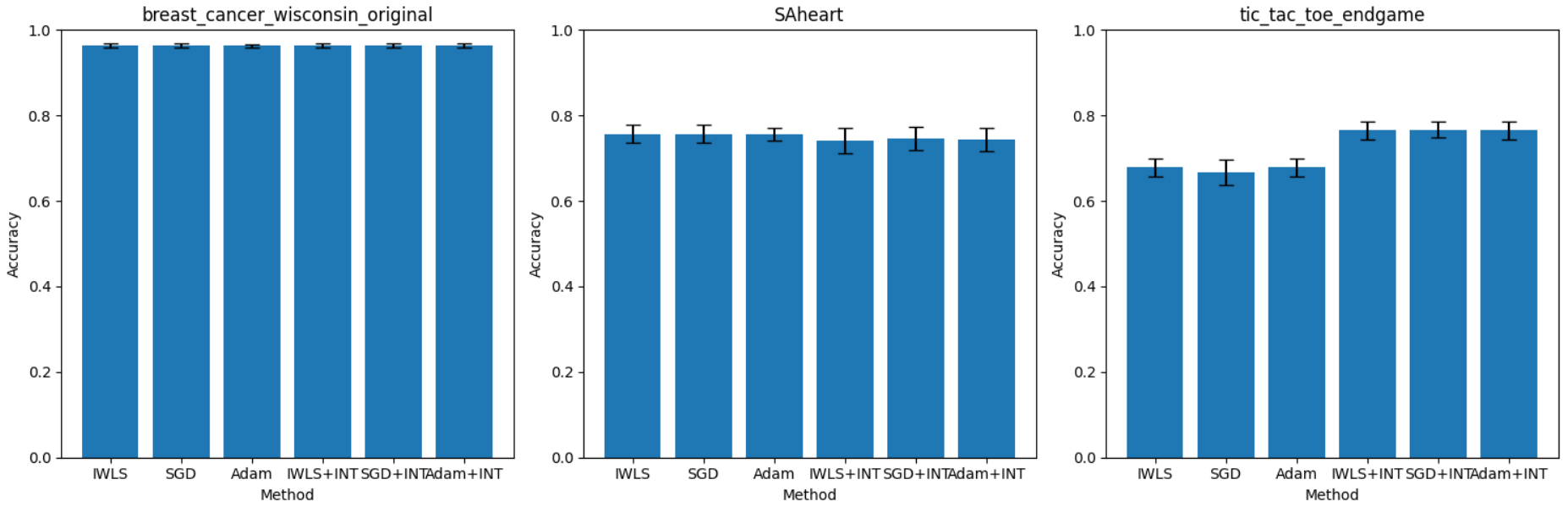
Convergence analysis



Classification performance



Classification performance - intersections



The End