Project 1

The aim of the project is to implement different optimization algorithms for logistic regression and compare their performance.

Part 1

- Find 5 different datasets corresponding to classification problem with binary class variable. You can use UCI repository: https://archive.ics.uci.edu/ or other sources. Non-standard, interesting datasets will be appreciated.
- Prepare datasets to run logistic regression algorithms. This point includes:
    - dealing with missing values,
    - converting categorical variables into numerical ones (use e.g. one-hot encoding),
    - removing collinear variables
    - split data into training and testing sets

Part 2

- Implement measures of classification performance: accuracy, recall, precision, F measure (harmonic mean between recall and precision).

Part 3

- Implement optimization algorithms for parameter estimation in logistic regression:
    - IWLS (Iterative Reweighted Least Squares)
    - GD (Gradient Descent)
    - SGD (Stochastic Gradient Descent)

Part 4 (experiments)

- Convergence analysis: check how the value of log-likelihood function depends on the number of iterations for 3 above algorithms. Check how the value of learning rate in GD and SGD affects the results.
- Propose the stopping rule for the above algorithms. Please remember to use the same rule in all three algorithms to make a comparison fair.
- Compare the classification performance of logistic regression (try all 3 methods: IWLS, GD, SGD) and 3 popular classification methods: LDA, QDA and KNN. Use the performance measures implemented in Part 2 and datasets prepared in Part 1. Use also R2 measure which is defined as 1-log-likelihood(model)/log-likelihood(null model), when null model contains only intercept. The models should be trained on training set. The performance measures should be calculated on test set. If the given algorithm does not converge, within 1000 iterations, stop the algorithm and use the solutions from the last iteration.
- Extra: include other optimization algorithms for logistic regression (you can use implementations available in libraries).

Additional remarks:

- The projects are implemented in teams of 2 students.
- The projects can be implemented in R or Python.
- The final grade will be based on: (1) source codes, (2) reports (max 4 pages A4) summarizing experiments, (3) presentations (max 5 minutes).

- Please send zip file (name of the file: surname1_surname2.zip) including 3 folders: code, presentation, report. My e-mail address: teisseyrep@ipipan.waw.pl
- Deadline: group 1: April 7 2021; group 2: April 14 2021.
- Presentations: group 1: April 14 2021, group 2: April 21 2021.
- If you have any questions, please send my an e-mail: teisseyrep@ipipan.waw.pl