Project 1

Deadline: 03.04.2024

Goal

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

1 Task 1

The aim of the first task is to collect and prepare data sets for conducting experiments.

1. Find **9 different datasets** corresponding to classification problem with binary class variable containing numerical variables. You can use repositories: https://archive.ics.uci.edu/, https://www.openml.org/ or other sources. Please choose **three small** datasets containing at most 10 variables and **six large** datasets containing more than 10 variables. In the case of all datasets, the number of observations should be larger that the number of variables.

   Remarks about dataset selection:

   - Non-standard, interesting datasets will be appreciated – datasets are interesting when 50% of them are different from used ones during previous courses.
   - You can convert multi-class datasets to binary datasets by combining classes.

2. Prepare datasets to run logistic regression algorithms. This includes:

   (a) filling in missing values (Datasets must have less than 10% missing data per variable.),

   (b) removing collinear variables.

2 Task 2

The aim of the second task is to implement optimization algorithms.

1. Implement optimization algorithms for parameter estimation in logistic regression:

   (a) IWLS (Iterative Reweighted Least Squares)

   (b) SGD (Stochastic Gradient Descent)

   (c) ADAM (Adaptive Moment estimation)

   Using implementations available on the web is not allowed, the idea is to write your own functions.

2. Each algorithm should have the option to include, in addition to the original input variables, interactions between variables that are defined as the products of two variables. For example, having variables $X1$, $X2$, $X3$, the model without interactions is based on the variables $X1$, $X2$, $X3$, while the model with interactions is based on the variables: $X1$, $X2$, $X3$, $X1 \cdot X2$, $X1 \cdot X3$, $X2 \cdot X3$.

3. Use the default parameters recommended in available implementations in the above optimization algorithms.

3 Task 3

The aim of the third task is to perform experiments.

1. Propose the **stopping rule** for the above algorithms. Please remember to use the same rule in all 3 algorithms to make a comparison fair.

2. For all algorithms as **performance measure use Balanced Accuracy**. The models should be trained on training set. The performance measure should be calculated on test set. Please average the results over at least 5 train-test splits. If the given algorithm does not converge, within 500 iterations, stop the algorithm and use the solutions from the last iteration.

3. **Convergence analysis**: check how the value of log-likelihood function depends on the number of iterations for 3 above algorithms. Convergence analysis should be performed on the train data.

4. **Compare the classification performance** of logistic regression (try all 3 methods: IWLS, SGD and ADAM) and 4 popular classification methods: LDA (Linear Discriminant analysis), QDA (Quadratic Discriminant Analysis), Decision tree and Random Forest. Use available implementations, e.g. from scikit-learn library.

5. In the case of small datasets, please **compare the two versions of the logistic regression: model without interactions and model with interactions**. This gives a total of 6 variants of logistic regression (IWLS, SGD, ADAM, IWLS+INT, SGD+INT, ADAM+INT).

4 General additional remarks

- The projects are implemented in teams of 3 students.
- The projects should be implemented in R or Python.

5 Final grade

The total number of points to be scored is 40, including:

1. source codes (12 points)
2. report summarizing experiments (20 points = 5 points x 4 section)
3. presentation* (8 points)

* All teams must record a presentation (maximum 5 minutes). During the classes 50% of the group will additionally present their results live (maximum 10 minutes).

Live presentation dates: **Group 1** 04.04.2024, **Group 2** 09.04.2024, **Group 3** 16.04.2024.

5.1 Requirements for reports

Every point described below should be included in separate section. Each section should have at most two pages: one page for text and one page for figures. Every section should start on the new page.

1. Methodology (Task 1, Task 3.1, Task 3.2)
2. Convergence analysis (Task 3.3)
3. Comparison of classification performance (Task 3.4)
4. Comparison of classification performance of models with and without interactions (Task 3.5)

6 Solution

Please upload a directory `Surname1_Surname2_Surname3` to GitHub repository https://github.com/kozaka93/2024L-DSAdvancedML via Pull Request. The title of PR should be `[P1] Surname1, Surname2, Surname3`. Every group should upload files to the folder `/projects/project1/groupX`, depending on attending the project group.

The directory should include:

- `codes` directory include all code needed to reproduce results
- report **(.pdf file)**

The recorded presentation **(.mp4 file** entitled `Surname1_Surname2_Surname3`) should be uploaded to MS Teams to the folder `/Projects/Project 1/Presentation/Group X`.