Machine Learning I (CC2008)
2022/2023

U.PORTO
Rita P. Ribeiro, Carlos Soares, Moisés Santos

Practical Assignment

Learning Goals

1. Understand how a selected Machine Learning (ML) algorithm works in detail, both theoretically and empirically;
2. Understand how benchmarking of ML algorithms is carried out;
3. Understand the difference between ML research and the use of ML to solve a specific application.

Description

Standard supervised machine learning algorithms follow different approaches to search for the best approximation to the function that maps a set of predictors to a target variable. Therefore, the ability to build accurate models depends on how well the selected algorithm fits the data. Furthermore, real-world data typically contain issues that challenge those algorithms with different levels of magnitude. Some of those issues in the case of classification tasks are:

- qualitative attributes with a large number of possible values;
- noise or outliers;
- class imbalance in binary problems;
- multiclass classification;
- class overlap.

Description

You will develop a variation of a classification algorithm of your choice and assess how that impacts the performance over the benchmark data sets. We suggest that the data sets used are the ones available in the OpenML-CC18 Curated Classification benchmark (How to access to OpenML datasets). We also suggest that the implementation is based on existing code, such as the one available at: https://github.com/rushter/MLAlgorithms.

Suggested Work Plan

1. Select a classification algorithm and find code that implements a standard version of that algorithm.
2. Understand the algorithm and hypothesize which of the suggested data characteristics affects it the most.
3. Choose a data characteristic to tackle.
4. Empirically evaluate the behavior of the selected algorithm on the set of benchmark data sets. It may be necessary to eliminate some of the data sets or adapt them. This

should be done with the support of the teachers and documented in the report. The analysis of the results should focus on the selected characteristic.

5. Propose a change to the algorithm that is expected to make it more robust to the selected data characteristic.
6. Implement the proposed variant.
7. Empirically evaluate the behavior of the proposed variant on the same set of datasets and compare the results with the original version.

Deadline

The deadline for submitting your assignment is **May 19th at 23:59**.

Groups

The practical assignment is mandatory and should be performed by groups of three students. You should constitute your group in moodle by enrolling in the groups available for the lab class you attend. That is, G1_X, G2_X and G3_X, for PL1, PL2 and PL3, respectively.

Deliverables

Your assignment should be submitted in moodle with a compressed file containing the following items:

1. the source of a ready-to-execute notebook with all the code necessary to run to obtain the presented results, including any complementary files needed to execute your notebook (e.g. data files, data objects);

2. slides for presentation (PDF format) focusing on the main issues of the assignment for a 12 minutes presentation; any additional information that cannot be presented in that time slot can be included as annexes to the presentation; see the presentation guidelines for further details.

Grading

- Comprehension of the selected algorithm
    - general (5%)
    - effect of data characteristics (5%)
- Proposal for handling the selected data characteristic
    - theoretical and empirical motivation (10%)
    - description (15%)
    - originality (15%)
- Empirical study

    - experimental setup (5%)
    - analysis of results (10%)
- Notebook (10%)
    - organization and python implementation
- Presentation - mandatory
    - slides (10%)
    - presentation (5%)
    - discussion (10%)

Presentation Guidelines

Suggested organization:

- cover slide (1 slide)
    - with names and numbers of all members of the group
- executive summary (1 slide)
    - goals
    - outline of the approach
    - summary of results
- selected algorithm and data characteristic (2 slides)
    - description
    - discussion of the behavior of the algorithm concerning the selected data characteristic
- proposal (1-2 slides)
    - motivation
    - description
- empirical study (3-4 slides)
    - experimental setup
        - datasets and their characteristics
            - focusing on the data characteristic of interest
        - hyperparameters of the algorithm
        - performance estimation methodology
    - analysis of results
        - presentation of results
        - discussion, in light of hypothesized effect
- conclusions and future work (1 slides)

Please note that the number of slides for the presentation is merely an indication.

The total number of slides, including annexes, should not exceed 40.