Machine Learning I (CC2008)
2024/2025

U.PORTO

Practical Assignment

26th March, 2025

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 use different strategies to approximate the function that maps a set of predictors to a target variable. The effectiveness of these models depends on how well the chosen algorithm fits the data. However, real-world datasets often contain challenges that impact model performance, such as noise, outliers, class imbalance, and multiclass classification.

Objective: In this assignment, you will modify a classification algorithm of your choice and evaluate its impact on performance using benchmark datasets.

To accomplish this, you will need to:

- Implement the classification algorithm from scratch (not using scikit-learn). You may base your implementation on existing open-source code, such as the one available at https://github.com/rushter/MLAlgorithms, but you are free to use other sources as long as they comply with the constraint above and you make reference to them.

- Select one of the following challenges to address and test your modified algorithm on the corresponding set of benchmark datasets:
  - Dataset Group 1: Noise or outliers
  - Dataset Group 2: Class imbalance in binary classification
  - Dataset Group 3: Multiclass classification

Tasks

Phase 1

- Select a classification algorithm and find code that implements a standard version of that algorithm.
- Understand the algorithm and hypothesize which of the suggested data characteristics affects it the most.
- Choose a data characteristic to tackle.
- Empirically evaluate the performance of the chosen algorithm on the benchmark data sets. It may be necessary to remove some data sets or modify them. This should be done with support from the instructors and documented in the report.

Phase 2

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

Deadlines

- **Checkpoint:** During the **practical classes from 24th April to 2nd May**, the professor will have a checkpoint with you. By that time, you should have completed Phase 1 and begun formulating ideas for Phase 2.

- **Final deadline:** You should submit the complete assignment by **25th May at 23:59.**

Groups

The practical assignment is mandatory and should be performed by groups of three students. Please select your group in Moodle by enrolling in the available groups for the lab class you attend: PL1_GX for PL1, PL2_GX for PL2, PL3_GX for PL3, PL4_GX for PL4 and PL5_GX for PL5.

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 min 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

- **Checkpoint** - mandatory (10%)
- **Comprehension of the selected algorithm** (10%)
    - general (5%)
    - effect of data characteristics (5%)
- **Proposal for handling the selected data characteristic** (30%)
    - theoretical and empirical motivation (10%)
    - description (10%)
    - originality (10%)
- **Empirical study** (15%)
    - experimental setup (5%)
    - analysis of results (10%)
- **Notebook** (10%)
    - organization and python implementation
- **Presentation** - mandatory (25%)
    - 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.