The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

COMP3009/COMP4139 Machine Learning 2024-25

Assignment 1

Machine Learning for Classification and Regression

Xin Chen

1. Introduction

This assignment assesses your practical data processing skills and the capability of applying machine learning methods to real-world problems. This assignment contains two tasks: regression and classification. The implementation will be based on Python and third-party Machine Learning libraries. From here on, you will **have to** work as a group, submitting a single report and code by **7th Nov, 2024 at 3 pm** on Moodle by **member 1** of each group. You can split and distribute the work to individual members, but each individual is expected to understand every aspect of the work.

2. Data

As a group, you will have to select **two** datasets: one for the task of regression, and one for the task of classification. You have to choose your dataset from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). You cannot choose WDBC dataset that was used in the lab sessions. The regression task is to predict an output with continuous values (e.g. house price, age, etc.). The classification task is to predict a categorical output (e.g. diseased/non-diseased, product quality: low/medium/high, etc.). You may consult the lab assistant on selecting datasets. You are allowed to clean the data either manually or through coding (preferred).

3. Implementation Requirement

For each of the datasets that you selected in section 2, implement a machine learning solution to achieve the specified task (regression or classification). You must implement **four** methods for each task, **including linear regression (logistic regression), support vector machines, decision trees and multi-layer perceptron neural network.**

Apply K-fold cross-validation (K is determined by yourselves) to evaluate each method and compare their performances. You may use classification accuracy as the evaluation metric for classification, and mean squared error (MSE) for regression.

All implementations need to use Python programming language. Any machine learning libraries are allowed (e.g. Scikit-learn, Scipy, Pandas, Tensorflow, Pytorch, etc.)

4. Assessment

The purpose of assignment 1 is to make the students familiarise themselves with the general ML pipeline and how to work in a group. Assignment 1 will not be marked against either the quality of the code and report or the method performance. It will be marked as either **pass or fail**. To pass it, you only need to submit a valid code and a report that demonstrate your group has done both a classification and a regression task using the four methods mentioned previously. No submission or inadequate quality of code and report (e.g. code and report do not match, the work does not accomplish the two tasks using the four methods) will receive a fail. By receiving a pass, each group member will get 20% out of 100 of the coursework mark (i.e. 6 marks of the total module mark). This assignment will also help to identify any problems in group work. **Please make the lab assistant and module convenor aware of any inactive members in the group, a mark of zero for assignment 1 will be given to the inactive member as a warning.**

5. Deliverables

For the completion of Assignment 1, the following have to be submitted on Moodle:

1. The Python code (.py or .ipynb) for implementing the two tasks and the associated datasets (spreadsheet). Store all files in one folder and compress them to a .zip file.

2. A report of up to 1000 words containing: a cover sheet (names of the group members), introduction, description of methods, parameter settings, evaluation method, results and a conclusion.

6. Marking Criteria

Criteria for Pass (20% of coursework mark):

- Submit both code and report on time.
- Completed both tasks (i.e. classification and regression).
- Applied all the four suggested methods
- The report covers the specified contents described in section 5.

Criteria for Fail (0 mark for Assignment 1):

- Either code or report is missing.
- Completed only one of the tasks.
- Didn't implement all four methods.
- The results in the report do not match the results of the code.
- No contribution to the group work.

Plagiarism check will apply, meaning that high similarities across different groups are not expected. Late submissions in each assignment will result in a 5% penalty per day (days rounded up to the next integer). Only one report and one code implementation need to be submitted per group.