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.)

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

COMP3009/COMP4139 机器学习 2024-25

作业 1 用于分类和回归的机器学习

Xin Chen

一、简介

这项作业评估您的实际数据处理技能以及将机器学习方法应用于现实问题的能力。这
作业包含两个任务：回归和分类。 这
实施将基于Python和第三方机器学习库。从现在开始，您必须以小组形式工作，由每个小组的 1 名成员于 2024 年 11 月 7 日下午 3 点之前在 Moodle 上提交一份报告和代码。您可以将工作拆分并分配给各个成员，但每个人都应该了解工作的各个方面。

2. 数据

作为一组，您必须选择两个数据集：一个用于回归任务，一个用于分类任务。您必须从 UCI 机器学习存储库 (https://archive.ics.uci.edu/ml/index.php)
选择数据集。您不能选择实验室会话中使用的 WDBC 数据集。回归任务是预测具有连续值（例如房价、年龄等）的输出。分类任务是预测分类输出（例如，患病/未患病、产品质量：低/中/高等）。您可以咨询实验助理选择数据集。您可以手动或通过编码（首选）来清理数据。

三、实施要求

对于您在第 2 部分中选择的每个数据集，实施机器学习解决方案以实现指定的任务（回归或分类）。你必须为每个任务实现四种方法，包括线性回归（逻辑回归）、支持向量机、决策树和多层感知器神经网络。

应用K折交叉验证（K由您自己确定）来评估每种方法并比较它们的性能。您可以使用分类准确率作为分类的评估指标，并使用均方误差（MSE）作为回归的评估指标。

所有的实现都需要使用Python编程语言。允许任何机器学习库（例如
Scikit-learn、Scipy、Pandas、Tensorflow、Pytorch 等）

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.

4. 评估

作业 1 的目的是让学生熟悉一般的 ML 流程以及如何在小组中工作。作业 1 不会根据代码和报告的质量或方法性能进行评分。它将被标记为通过或失败。要通过它，您只需要提交有效的代码和报告，证明您的团队已经使用前面提到的四种方法完成了分类和回归任务。未提交代码和报告或质量不合格（例如代码和报告不匹配、工作没有使用四种方法完成两项任务）将失败。通过考试后，每个小组成员将获得课程作业分数的 20%（满分 100 分）（即模块总分数的 6 分）。这项作业还将有助于发现小组工作中的任何问题。请让实验室助理和模块召集人了解小组中任何不活跃的成员，作业 1 的零分将给予不活跃成员作为警告。

5. 可交付成果

为了完成作业 1，必须在 Moodle 上提交以下内容：

1. 用于实现两个任务的 Python 代码（.py 或 .ipynb）以及相关数据集（电子表格）。将所有文件存储在一个文件夹中并将它们压缩为 .zip 文件。

2. 1000字以内的报告，内容包括：封面（课题组成员姓名）、简介、方法说明、参数设置、评估方法、结果和结论。

六、评分标准

通过标准（占课程成绩的 20%）：

- 按时提交代码和报告。
- 完成两项任务（即分类和回归）。
- 应用所有四种建议的方法
- 该报告涵盖第5节中描述的具体内容。

失败标准（作业 1 为 0 分）：

- 代码或报告丢失。
- 只完成了其中一项任务。
- 没有实现全部四种方法。
- 报告中的结果与代码的结果不匹配。
- 对小组工作没有贡献。

剽窃检查将适用，这意味着不同群体之间不会有高度相似性。每项作业的逾期提交将导致每天 5%
的罚款（天数四舍五入到下一个整数）。每组只需提交一份报告和一份代码实现。