Coursework 2: Machine learning prediction of patients with schizophrenia from anatomical brain imaging

The aim of this course work is to investigate the performance of different feature sets, models and cross-validation strategies in a machine learning challenge and write an individual report describing the project.

Submission: You should submit two files to the course work submission – an individual scientific report (pdf file) and a Jupyter notebook with your code (with comments throughout). In addition, you should submit your best pipeline (estimator) to the challenge website.

Grading: The content as well as the clarity (typos, structure, etc) of the report will be considered for marking. We will focus grading mostly on the report, but the Jupyter notebook might be used as supporting material. Therefore, it is important that the analysis steps, results, etc are described in the report.

Note: This is an individual coursework, and we will investigate any submissions that seem to have been plagiarised.

Individual Scientific Report

The scientific report should be independently completed. The Lecture Notes in Computer Science template is recommended (https://www.springer.com/gp/computer-science/lncs/conference-proceedings-guidelines). The report should not exceed 4 pages with maximum 2 pages references. The report should include the following sections to describe the different aspects of the project as described below.

- Introduction
- Methods
- Results
- Discussion
- Conclusion

Project: Predict schizophrenia from brain grey matter (classification task)

Background

Schizophrenia is associated with diffuse and complex patterns of brain atrophy. In this project we will learn predictors of the clinical status (patient with schizophrenia vs. healthy control) using grey matter (GM) measurements from the participants' brain.

Dataset

The dataset for the project can be found in the link https://ramp.studio/problems/brain_anatomy_schizophrenia and consists of pre-processed structural Magnetic Resonance Imaging (MRI) data (voxel-based morphometry, VBM). Two versions of the data were extracted from the pre-processed structural MRI data: (1) a low dimensional version consisting of 284 brain regions of interest (ROIs); (2) a high dimensional version consisting of 331695 grey matter voxels. The dataset can also be assessed through the challenge page as described below.

Project aims:

(1) Investigate the performance of three different machine learning models (including a regularized linear model, a tree-based/ensemble model and a non-linear model) for predicting schizophrenia from brain grey matter using the high dimensional and low dimensional features. The three different models should be applied to the high and low dimensional versions of the data unless it is computationally not possible.

(2) Compare the performance of a common cross-validation versus a group stratified cross-validation, considering sex as group.

(3) Submit the best pipeline (estimator) to the RAMP platform challenge: Predict schizophrenia from brain grey matter.

Comment: Since one of the project aims is to compare the performance of the models using the high versus the low dimensional data at least one of the models should be applied to both versions of the data without dimensionality reduction or feature selection.

What should be covered in the report?

- **Introduction:**
  - Description of the motivation and background of the project. (10%)
- **Methods:**
  - Description and justification of the models used for comparison. (10%)
  - Description of the pipelines used and of the pipeline submitted to the challenge (e.g. cross-validation strategies, hyper-parameter optimization). (20%)
  - Description and justification of the choice of comparison metrics. The comparison metrics should include metrics that measure both the performance and the computational cost of the models. (5%)
- **Results**
  - Description of the results for different models and cross-validation strategies applied to the low and high dimensional versions of the data (e.g. cross-validation performance, test performance and computational cost). The results should include tables and/or figures. (20%)
- **Discussion**
  - Discussion of the results for the different models, explaining which models perform better (in terms of performance and computational cost) in the low and high dimensional data. Discussion of the impact of the different cross validation strategies on the results. Discussion of the implication of the results from an applied perspective. (15%)
- **Conclusion**
  - Description of the main conclusions of the project. (10%)

Examples of report can be found in the Moodle page.

Machine learning challenge (10%)

The challenge will be done on the RAMP platform https://ramp.studio/.

1. Register on the platform.
   - Choose as login your last name.
   - **Use your UCL e-mail**.

2. Join the "join event" challenge on the page:
https://ramp.studio/events/brain_anatomy_schizophrenia_UCL_2024

3. Download the kit " Download kit ".

4. Follow the "Installation" instructions on the challenge page.

5. Apply the "Getting started" to:
   - download the data,
   - run the data notebook,
   - create a new model and test it locally and,
   - make a submission to RAMP Test submissions locally

The results of the challenge will be released in Moodle after the coursework deadline. The top three winners of the challenge will be invited to present their Jupyter Notebook on the Applied AI practical session on March 19th 2024.