UNIVERSITY OF OXFORD
SOFTWARE ENGINEERING PROGRAMME

Wolfson Building, Parks Road, Oxford OX1 3QD, UK
Tel +44(0)1865 283525 Fax +44(0)1865 283531
info@softeng.ox.ac.uk www.softeng.ox.ac.uk

Part-time postgraduate study in software engineering

Classical Machine Learning, CML

9th – 13th September 2024

ASSIGNMENT

The purpose of this assignment is to test the extent to which you have achieved the learning objectives of the course. As such, your answer must be substantially your own original work. Where material has been quoted, reproduced, or co-authored, you should take care to identify the extent of that material, and the source or co-author.

Your answers to the questions on this assignment should be submitted using the Software Engineering Programme website — `www.softeng.ox.ac.uk` — following the submission guidelines. When submitting the assignment online, it is important that you formally complete all three assignment submission steps: step 1, read through the declaration; step 2, upload your files; step 3, check your files. **Please ensure your submission is anonymous: do not include your name or any other identifying information on the assignment, nor in accompanying material such as source code, nor within the file names of anything submitted.**

The deadline for submission is 12 noon on Tuesday, 29th October 2024. You are strongly encouraged to submit a version well before the deadline. You may update your submission as often as you like before the deadline, but no submissions or changes will be accepted after the deadline.

We hope to have preliminary results and comments available during the week commencing Monday, 9th December 2024. The final results and comments will be available after the subsequent examiners' meeting. Exam Conventions can be found here `https://www.softeng.ox.ac.uk/handbook/`

ANY QUERIES OR REQUESTS FOR CLARIFICATION REGARDING THIS ASSIGNMENT OR PROBLEMS INSTALLING SOFTWARE SHOULD, IN THE FIRST INSTANCE, BE DIRECTED TO THE PROGRAMME OFFICE WITHIN THE NEXT TWO WEEKS.

Classical Machine Learning (CML) Sept 2024

1 Assignment Task

In the following scenario you will play the role of expert in Machine Learning working in the field of genetics research in the domain of agronomy (the science of crop production and soil management, including optimizing plant growth and yield maximisation). A commercial organisation has completed a genetic assay that focusses on genetic marker development. They have surveyed a large collection of Landrace Wheat[1] and need you to analyse the results.

You have been given a data file ('wheat_genome_1.csv') that contains the results from automated analysis of DNA from a historic Wheat collection. For reasons of commercial sensitivity, the feature names in the data file have been coded; you won't be told the exact definition of the feature names – but they can be understood in general terms:

- The first 40 features have been assigned coded, 4-character names ('ZYTI', 'WIBD', 'KVJI' etc). These features contain floating-point numbers representing the prevalence of specific genetic markers in that each sample (observation).
- The final 5 columns (called 'RESP__0', 'RESP__1' etc.) are 'response' measurements. These represent figures of merit for various properties of the wheat sample. You won't be told the exact interpretation of these response variables. This does not matter to your task, but you can imagine that they might refer to properties such as 'pest resistance', 'productivity' and 'disease immunity' etc.

The research scientists working on the wheat genetics have several questions about the data. Your task is to analyse the dataset, to answer these questions and to provide any other insights that you may discover when reviewing the dataset. The questions are as follows:

1. Is the wheat in this dataset homogonous, random or does it fall into distinct 'types'?
2. If the wheat is homogonous or random, then identify the overall characteristics of the complete sample.
3. If there are distinct types of wheat in the dataset, then identify the characteristics of each type of wheat.
4. What is the relationship between the input variables (first 40 features) and the 5 response variables?
5. Are there specific features which 'drive' specific responses?
    a. If so, how?
    b. What are the important features and what is their relative impact?
6. If there are different types of wheat present I the dataset, then which type of wheat has the greatest impact on each response variable?

[1] Landrace wheat refers to traditional, locally adapted varieties of wheat that have evolved through natural and farmer-driven selection over centuries. These varieties are typically well-adapted to specific local environments, showing resilience to local pests, diseases, and environmental conditions.

You should present your findings both numerically and graphically (in the form of charts) as appropriate.

You should document your work within a Jupyter notebook: Analyse the data, make sense of it, build models, draw conclusions, assure the quality of the models, make appropriate predictions, document any interesting discoveries and report on your work. You should experiment with different types of model to discover which are most suited to your analysis and modelling task.

You are required to submit a Jupyter Notebook that contains: (but also carefully note the Submission Requirements below):

- All executable software code (in Python, using where required appropriate libraries);
- Explanatory text that documents your code : That is, explains it so that another ML expert with broadly similar knowledge and experience could understand and reproduce your results;
- Your explanatory text should include rationale for each choice you make.

Additionally, following company standards, your analysis and modelling is to be completed using the Python Language. You must include a section within your Jupyter Notebook that identifies which libraries you have used with references identifying their source.

2 Submission Requirements

You are required to submit **two** versions of any Jupyter notebooks you submit:

- A fully-executed version of .ipynb file (that is, containing all outputs) : so that the examiner can execute the code if required;
- A **pdf version** of the fully-executed notebook : if the examiner cannot execute any cell within your .ipynb file it will still be clear how that code operated on your machine

You should not include any analysis employing members of the family of Neural Networks (including large language models). Any such analysis is outside of the scope of this assignment will not attract any credit.

Do not submit information other than the above.

3 Assessment Criteria

This assignment is intended to assess your ability to:

- Employ the tools and techniques presented as part of the Classical Machine Learning course in novel situations;
- Make critical judgments regarding the applicability and suitability of Classical Machine Learning algorithms to specific problems;
- Provide rationale, explanation and documentation for any decisions you make whilst engaged in a Classical Machine Learning project; and,
- Critically assess the quality and limitations of any models you build and/or analysis that you complete.