Project

Foundations of Data Science – TA team

March 2023

Introduction

The goal of the project is for you to experience an entire data science project from start (choice of dataset, formulation of research hypothesis) to end (presentation of results). You will have to apply the methods and tools you were introduced to in the lecture and had a chance to practise with during tutorials and in homeworks.

1 Groups

Form a team of 4 Students to be able to work together during the tutorials and for the final presentation of results. We cannot accommodate smaller groups. You need to register your group by 06 April 2023 via moodle.

2 Datasets

We provide you with a closed list of datasets, from which your group should choose one. The datasets are available to you on the moodle page of the course together with brief descriptions and potential additional references. Once groups have been registered through moodle, there will be another option available to choose the dataset. Each group has to submit a choice of dataset(s) by 16 April 2023.

To have sufficient diversity in projects within tutorial groups, each dataset can only be used at most in five projects. We will ask you to submit your group registration together with the dataset you would like to use and confirm the availability of the dataset to you. Please do not start working on a project with the dataset you initially submitted unless you receive a confirmation from your tutor.

For the selected dataset you should perform:

- a brief literature review motivating your project,
- data preprocessing and visualisation,
- training of 3-4 different machine learning models (one per group member),
- evaluation and comparison of your models' performance (visualisation of results),
- discussion of the biomedical and literature context of your results.

3 Deliverables

3.1 Presentation

You will present your projects in the tutorial in last week of the semester, 01 June 2023 as a team. The presentation should be 10 minutes long and will be followed by a brief question and answer session (two to three questions).

3.2 Report

Each team is asked to submit one written report. This is is due on 19 June 2023 at 23:59.

Please use the NeurIPS template for the report. You can write the report using Overleaf, of which the full version is available to you for free through ETH and which has the NeurIPS template available, or a local LaTeX installation. You can recreate the same document style (font size, margins, etc.) in Word but we suggest you write the report in LaTeX.

The report should contain the following sections:

1. Abstract (150 words)
2. Introduction: this should end with the aim(s) of the project and clearly state the scientific hypothesis addressed.
3. Methods: this should describe the source of the data, preprocessing performed, potential feature selection, implementation of model(s), model training and evaluation.
4. Results: this should contain a description of the data and results indicating model performance.
5. Discussion: this should relate your results to your initial hypothesis and contextualise your results with respect to the literature.
6. References: a minimum of 10 references would be expected.

The report should be 5-10 pages long (including figures but excluding references). Any text exceeding 10 pages (excluding references) will not be considered.

3.3 Code/github repository

You will have to submit your code once for your team and make it available to the tutors through a github repository. We will provide you with more detailed instructions closer to the deadline. The code has to be submitted by 19 June 2023 at 23:59.

4 Timeline

We recommend to form project groups early on - by the week following Easter you need to register your group and provide a very brief outline of your anticipated project (which data set are you planning to use to address which research question). We will then make time in the following tutorials for you to work on the project and ask specific questions to the tutors.

4.1 Deadlines

Please note that deadlines are hard deadlines and deliverables cannot be submitted late. The presentation is an essential part of the project and all team members are expected to attend.

1. group registration: 06 April 2023
2. data choice: 16 April 2023
3. presentation: last tutorial (01 June 2023)
4. report and code: 19 June 2023 at 23:59

5 Grading

We will account for the following aspects and questions when grading:

- the difficulty of the task chosen (for example: binary or multi-class classification; dimensionality of dataset chosen)
- the visual presentation of the raw and/or preprocessed data
- the choice and methodology and used ML models (the student driving a particular model should be highlighted in the methods)
- Do your comparisons and conclusions make sense?
- Did you consider the current state of the art in terms of biomedical and ML context motivated in the discussion/introduction through relevant references?
- Are potential limitations or difficulties highlighted?

Please note that the final performance of your model will not influence the grade per se, *i.e.* if your methodology is sound and you have performed all relevant aspects needed but your model still does not perform very well that is okay. We would like you to provide a realistic reflection of what is possible with a specific model. Of course a discussion and how you can explain the observed performance and differences between your models and the literature is key here.

The grade for the project will be a composite of all these aspects.

6 Python

You have to complete all coding for this project in Python. Please use a Python version ≥ 3.7 and provide a list of used packages (and their versions) with your submission in the appendix of your report (excluded from page limit).