MATH70103: Unstructured Data Analysis – MLDS
Autumn 2023-2024

Module Developer: Dr. Anthea Monod
Imperial College London

Final Project

Information and Technical Instructions Please select your own imaging, network, or text dataset(s) and perform a data analysis by applying the suitable techniques that you've learned from the relevant topic via the lecture notes, videos, and homeworks to your chosen dataset(s). The objective of the final project is to test your knowledge about the analysis of a real world unstructured dataset.

- This final project is worth 40% of your final score for the module. It is released on Monday, 11 December 2023 at 09:00 UK and due by **Friday, 5 January, 2024 at 23:59 UK**.
- You are expected to spend 20-25 hours on the project. You are given 4 weeks to submit the project, to allow for access to technical support for submitting your work and, scheduling/flexibility around the holidays, professional/personal commitments, as well as assessments for other modules you are simultaneously taking.
- Please submit your project as a PDF document (no more than 12 pages long, font size 11) written usign either word or latex. The report should contain all the relevant information about your project. Please also submit your dataset and accompanying Python code as a Jupyter notebook with clear code comments; your code does not count towards the 12-page limit on your project submission. Please give appropriate file names to your project submissions, e.g., `UDA_FinalProject_Lastname.pdf`.
- The preferred method for submitting your data and code is via a GitHub repository with the link included in your report. Alternatively, you can also upload code and data to Coursera.
- Please submit code written only in Python for the coding requirements of the project. You may use whichever packages you wish to complete the assignment, however, please clearly indicate which packages and versions you have used in your notebook.
- Please include a statement indicating that you have worked independently.
- Both the office hours and the live session of week 11 are fully dedicated to questions about the choice of the dataset and the final project in general.
- **Module lead and GTA will only be available to respond to questions about the project until the end of term on Friday, 15 December, 2023 at 17:00 UK.**
- **Please set the random seed at the beginning of your code implementations.**

```
import numpy
numpy.random.seed(030224)
```

Project Criteria and Mark Allocation

1. **Data Selection (15/100 marks):** You may choose any data type and real dataset(s) you wish, from work, a hobby, or the internet. You may choose several datasets of the same kind (image/network/text) but please ensure that each dataset does not exceed 100 MB in size in order to be able to submit your dataset along with your project report. Simulated datasets will not be accepted.
   You will be given credit for originality and complexity of your dataset (5/100 marks). For example, if you are working with a digitized image of an old black-and-white photograph from your family valuables, you will get more credit for originality and complexity rather than using a famous benchmarking dataset from a publicized challenge.
   Please be sure to properly source your data and give a thorough description and visualization of your data (10/100 marks).
   It is your full responsibility to ensure that you have permission to use your chosen dataset. We will not take any responsibility for permission-related oversights and their consequences. We will also not assist with setting up any data sharing agreements. For this reason, if you choose to use a dataset from online, we very strongly recommend that you use open access data.

2. **Problem Statement (15/100 marks):** Given your chosen dataset and its characteristics, please be sure to clearly and concisely state the problem you aim to address. Depending on the complexity of your chosen data, this may be a single task, but it may entail more. For example, if you have a very large and challenging network, you may choose to do only community detection. If you have a very noisy image, we have seen that the tasks of image denoising, edge detection, and segmentation are all closely related so you may try to take on all three tasks. The goal here is to choose task(s) of suitable complexity for the data you have chosen; there should be a clear desire to "get through the thick" of your data and uncover what you can from them, rather than just choose an easy task to "get it over and done with."

3. **Description and Justification of Methods and Analysis (20/100 marks):** Reflect on the most appropriate techniques and approaches from what we've seen in the module to apply to your data, given the problem you want to study. You may certainly use any technique or selection of techniques we've seen in the content, taken from the lecture notes, videos, and homeworks. You are also welcome to get creative and look up other approaches that we did not explicitly cover in the module, but if you do this, be sure to give an introduction and outline to the approach. For any technique you use, be sure to motivate its use properly and justify it as an appropriate approach.
   If you used a benchmarking or challenger dataset from the internet, you will also need to do a literature review on other approaches used and their results on these same data. Specifically, if you are straightforwardly applying one of the methods we've seen in the module to a famous dataset, it's extremely likely that this same analysis has been done previously by others. You should cite these works and summarize the results that others have achieved.

4. **Interpretation and Reflection on Output (20/100 marks):** Describe and interpret your findings from having applied your chosen techniques to your dataset. Are the results what you expected to see? Why or why not? Did you try several techniques? Which ones worked better, why, or why not? Are there any characteristics of the data that made one approach better than others?
   If you used a benchmarking or challenger dataset from the internet, you will also need to do a comparison on your results to those established by others who worked on the same data. How does your implementation compare to theirs and what can you say about that?

5. **Report Presentation and Clarity (15/100 marks):** Please pay close attention to detail as well as overall structure and format of your report. Please use appropriately-titled sections and subsections, references, clearly-labeled mathematical statements (e.g., definitions, propositions, etc.) where appropriate. Writing style and grammatical correctness are important.

6. **Code Presentation and Clarity (15/100 marks):** Similarly, a well-organized Jupyter notebook with clear code comments and a description of all functions and purposes for cells will be very important. Open and transparent research and reproducibility are becoming increasingly important; in fact, the top

statistics and data science journals are now requiring code to be submitted with manuscript submissions and dedicated reproducibility editors are being engaged on editorial boards to ensure reproducibility of results. **We will run your code to check for reproducibility of your results,** so please give clear instructions on how to run your code (e.g., in terms of a README markdown on your GitHub repository). Clearly label which functions and cells produced which figures in your report. Please also give clear indications on the hardware and software used in your analyses (what resources did you use; did you run the code on your laptop, a desktop, a cluster? Did you run parallel jobs, using e.g., SLURM or OpenPBS? How many nodes and cores, CPU/GPU, memory per CPU?) and an indication of runtime.

General Advice

- Week 11 is a good opportunity for you to get feedback from the teaching team on the suitability of your chosen dataset and the appropriateness of your proposed problem statement. Remember, your project will be assessed as a whole and the appropriateness of your proposed problem statement and suitability of your chosen dataset go hand-in-hand. They will also largely depend on the analyses you ultimately carry out. This means that although the module lead and GTA can give support and may be able to give some indication on whether the data and/or problem seem reasonable, this is by no means meant to be considered as a "pre-approval" of your data and/or problem.
- You are expected to carry out the actual analyses and write the report individually, i.e., without support from the module lead or GTA. Although you might want to brainstorm some approaches you are considering with the module lead and GTA during Week 11, remember that this plan can (and probably will) change as you carry out the analysis. Again, this is an indication that any feedback you receive from the module lead and GTA during Week 11 is only indicative and does not constitute a "pre-approval" and promise of high marks for Criteria 1-3 given above.
- **The module lead and GTA will not look at or help you with your code, or read any drafts of your project report before submission.**
- Try to think of this project holistically and not as a "box-ticking" exercise to fulfill all the criteria listed above. While they are required components of your final report and marks will be allocated according to those criteria, the overall structure and approach is quite a good simulation of how real research problems look like and how to approach real data-centric problems, both in academia and industry. This is ultimately what you will have to encounter, certainly in the near future with the project component to the program, and most likely in your future professional and career plans upon completion of the program as well.