Portfolio-Exam

This is the main document (6 pages) of the portfolio exam for the Data Science module MADS-EMDM – Advanced Topics of Data Mining. In this document, the structure of the exam, 8 out of the 9 portfolio tasks, several rules, and evaluation criteria are described. The last task will be revealed at the end of the term in the document `portfolio_standalone.pdf`.

1 The Portfolio Structure

The portfolio consists of 9 tasks.

1.1 Graph Tasks

Tasks 1–8 together resemble a term paper including story, experiments, and conclusions. The main idea is to conduct a comprehensive experiment on a real-world graph dataset of your choice (see criteria below) using various graph algorithms.
The description of the graph tasks can be found in Section 4 of this document. The tasks are supposed to be completed in order. The result will be one Jupyter Notebook containing comprehensive experiments. Think of this notebook as a report of a (small) project which you – as a data scientist in a data science company – conduct for a customer.

1.2 Standalone Task

The last task, Task 9, is a separate experiment. Its description will be published at the end of the term in the document `portfolio_standalone.pdf`.

2 Rules and Conditions

Programming Language Use Python and networkX for the submission.

Text Language Choose either English or German for all textual content.

Dataset The dataset may be chosen according to the following three requirements:

- Use REAL data from real applications/sources, no artificial datasets!
- Do NOT use graph datasets from the EMDM module. All graph datasets from the lectures and exercises of the graph chapters as well as derivatives thereof are NOT allowed for this exam.
- The data must be available: There are two possible availability scenarios:
  - It is available online with proper license. In that case indicate in your notebook where the data can be downloaded.

– You (legally!) obtain a dataset and share it with me via Moodle. Such data should not come with any form of NDA or other obligations. If you are not sure about these criteria regarding the dataset you consider, please contact me before you invest too much time in the experiments. If you write your own code to acquire data (e.g. querying an API), create a **separate notebook** for that purpose only and submit everything in a zip file.

No Teamwork This exam is supposed to be done during the self study time of the module. Students are allowed to exchange ideas. However, this is **NOT a teamwork exercise**. Every student must derive and write up their own solutions in their own words and programming style.

Code from other Sources You may reuse all the code from this module's lectures and exercises. Copying (and adapting) from other sources is allowed in small quantities – e.g. a function from stackoverflow. Quote the respective source. WARNING: Copying code in large quantities will be treated as intent to deceive and result in a score of zero points.

3 Submission

All parts of the portfolio should be contained in two Jupyter Notebooks. The first notebook called `experiments_graph.ipynb` should contain your answers to Tasks 1–8 (graph tasks). The second notebook called `experiments_standalone.ipynb` should contain your answer to Task 9. You may submit either just the notebooks or a zip file containing

- the notebooks,
- a resources folder **if** you have additional resources (e.g. images),
- an additional notebook **if** code is required to assemble the dataset (cf. Section 2), and
- a data folder **if** your data is not simply available online.

Upload your results to Moodle **before 11:59 pm (23:59 o'clock German time) January 08, 2025**. **Submissions after the deadline or via means other than Moodle will not be considered!**

4 Graph Tasks

Here, you will find Tasks 1–8 of the portfolio exam. Think of these first parts of the portfolio exam as a report of a (small) project which you – as a data scientist in a data science company – conduct for a customer (a company, an institute, some organization).

Task 1 – Context

This task sets the stage for the subsequent graph tasks as it determines the context for the experiments. For that purpose:

- You present a real scenario or a (realistic) fictitious situation, in which you are the data scientist. You propose to analyze and use an **undirected graph** dataset to learn about the entities of the graph – using the methodology described in the following tasks.

- Explain the context – project or task – for your experiment.
- Explain the plan of the experiment and the purpose – the value (monetary or otherwise) you expect to create for the organization.

Task 2 – The Data

Load and present an undirected graph dataset (respect the conditions in Section 2). This will be the raw data, that you'll exploit in the subsequent tasks.

- Explain the dataset itself (e.g., what do the nodes and edges represent?).
- Explain how the dataset is suitable for the project from Task 1. Show a helpful visualization of (part of) to support your explanation.

Task 3 – IDA and Preprocessing

Conduct an initial data analysis and bring the graph into the form for the analysis.

- Present basic properties of the graph.
- Present and discuss the degree distribution in the graph.
- Conduct necessary preprocessing. If you change the dataset, show and explain how that changes the basic properties and the degree distribution.

Task 4 – Graph Properties

- Compute and explore various graph properties.
- Show a table in which you list the graph properties and additional data that helps to interpret the results.
- Discuss your findings.

Task 5 – Central Nodes

Determine central nodes.

- Compute several useful centrality measures.
- Compare nodes according to different measures.
- Draw conclusions about their relevance/importance/position/role in the network – depending on what fits to the context of your experiments.

Task 6 – Node Roles

Assign roles to the nodes in the graph.

- Choose a node role assignment (it does not have to be the one from the lectures).
- Explain what the roles mean within the context of your project. – If you think that the role assignment does not make sense in your context, then argue this point in the notebook, but still do the experiment.
- Analyze the assignment of roles.

Task 7 – Community Discovery

Compute community clusterings of the graph.

- Run at least two community discovery algorithms.
- Compare and interpret results.
- Investigate the community membership of the central nodes (Task 5).

Task 8 – Conclusions and Future Work

Please address the following points separately an in that order.

1. Summarize and interpret the achieved results.
2. Compare your results to the expected or desired outcomes in the original plan (Task 1).
3. Explain the generated value! How do the analysis and the selected prediction algorithms help the organization?
4. Recommend a course of action for the organization in your story based on the results.
5. Reflect on limitations and possible pitfalls of using these results.
6. Critically discuss the employed methodology (your choices as well as the choices given in these tasks). What could or even should have been done differently?
7. Propose ideas for future work (a short sketch or enumeration of ideas is sufficient, no further experiments). The ideas should not be too general (e.g., "try further algorithms") but be specific to the project (e.g., "try Algorithm X, as because of Property Y, it might work specifically well on this dataset").

Note: One further task of the portfolio will be published at the end of the term. This will, however, be a standalone task.

5 Expectations

Your final score will be composed of 15 points for the context (set in the first task and followed throughout the report), 70 points for the actual experiments, and 15 points for presentation. Note,

- that poor presentation can lead to loss of points in the other two categories as well, e.g. if it's too confusing!
- that an unreasonable or trivial data science problem can lead to loss of points in other categories or even to a rejection of the portfolio!

The points for context are achieved with the graph tasks only, the other points result from all tasks.

5.1 Context

The (possibly fictitious but) realistic context of the experiments is set in the first task. Especially the last graph task should relate to the goals and compare the achieved results against them. The report should follow the story of Task 1 in a straightforward, comprehensive way. The experiments and their results' interpretation should fit to the goals and values proposed.

5.2 Experiments

When grading your experiments, I consider technical soundness, completeness, and fit to the tasks. I expect (among others):

1. The task is a non-trivial data science task.
2. Your code is executable and yields reasonable and reliably replicable results.
3. All cells of the notebook have been executed.
4. The choices of methodology (different approaches, dataset specific choices, settings of hyperparameters, etc.) make sense and are reasonably explained.
5. You have experimented with a reasonable selection of hyperparameters (just one selection is not sufficient).
6. The experiments address the influence of random choices and use appropriate steps to mitigate them.

5.3 Presentation

When grading the presentation, I will put myself into the position of your project's customer. I expect (among others)

1. that the code is structured (DRY principle, organized imports, . . .) and commented and free of errors or distracting outputs (e.g., no debug messages).
2. that the code is documented using appropriate means of Jupyter.

3. that results are presented in helpful numbers, in tables, and customized diagrams which are referenced and **interpreted** in the text.
4. that diagrams are easy to understand (appropriate colors, tics, scaling, labelling, legends, ...).
5. that the text is easy to understand (short, concise sentences, proper references to (previous) results, diagrams, ...).
6. that I am guided through the different parts of the experiments and told what the purpose of upcoming code blocks will be.
7. that the **visualizations are actually visible in the uploaded notebook**! For example, there are known issues with `plotly` graphics, which only show in notebooks when executed directly, but not afterwards, unless explicitly configured.

6 Advice

- Depending on the project and dataset you choose, some of the portfolio tasks may be more or less extensive. E.g. if your dataset is already cleaned up, preprocessing might be very quick. However, in other situations you might want to make use of various preprocessing steps and iteratively refine the dataset before you run algorithms on it.
- Your notebook should contain **much more text**, introduction, comments, etc. than the notebooks we use to demonstrate methods in the lectures or the exercises!
- **You may use bullet points** instead of continuous text!
- Before you submit, re-read the exam description and make sure that all points are addressed.
- Before you submit, re-run your notebook and check that the results are still as expected.
- Keep in mind that this is NOT your Master Thesis. The experiments should be comprehensive and created and conducted solely by you. However, limit your work to the portfolio tasks, solving ONE problem – even though along the way you might recognize other interesting angles to follow up on.