Portfolio-Exam

Read this document (5 pages) carefully – it describes the conditions and criteria for grading!
This is the main document for the portfolio-exam of the Data Science course MADS-DL (Deep Learning). In this document, the structure of the exam, several rules, and evaluation criteria are described. Please also read

- the document `lecture_datasets.pdf` from September 22, 2024 and
- the document `portfolio_tasks.pdf` containing a series of tasks to fulfill.

The result will be a short presentation (pitch) and a Jupyter Notebook containing comprehensive experiments, conducted according to the following requirements.

1 The Portfolio Tasks

The full portfolio will resemble a term paper including project proposal (pitch), experiments, and conclusions. The main idea is to conduct a comprehensive experiment on a real-world dataset using Deep Learning methodology. In a first submission, a pitch presentation will be created. In a second submission, all experiments will be submitted together in one Jupyter Notebook. Think of the notebook as a report of a (small) project which you – as a data scientist in a data science company – conduct for a customer. The tasks are published in the separate document `portfolio_tasks.pdf` in Moodle and are supposed to be completed in order.

1.1 The Pitch

The deliverable of the first phase is a presentation in PDF format, named `pitch.pdf`. You may use any tool you like to create it, but you **must submit a PDF!** It should be 3-5 slides including a title card. The pitch will NOT actually be presented, thus the slides must be self-explanatory!
The pitch sets the stage for the subsequent phases as it determines the story for the experiments in the subsequent phases. For that purpose:

1. You present a fictitious situation or a real scenario, in which you are the data scientist. You propose to solve a **non-trivial data science problem** and want to convince a customer to greenlight your project.
2. You present and explain the goal of your experiments and the value that you expect to generate for the customer if those experiments are successfull.
3. You present the dataset that you are going to use and you explain why you consider it suitable to reach the intended goals.

1.1.1 Pitch Advice

- The proposal must make sense in the context and approach a meaningful problem that does or could exist in reality.
- Make sure that the proposed problem can be approached through the experiments described in the document `portfolio_tasks.pdf`. Especially, make sure that you propose a clustering task and that the described value could result from the models you build following those instructions.
- The expected value for the customer can but does not have to be monetary. It must, however, explain what specifically your experiments (if successful) will enable.

1.2 The Report

Use the dataset you proposed in your pitch to conduct the experiments described in the document `portfolio_tasks.pdf`. Over the course of the term, you will learn the necessary means to complete these tasks.

1.2.1 Report Deliverables

All parts of the report should be contained in one Jupyter Notebook called `experiments.ipynb`. You may submit either just the notebook or a zip file containing

- the notebook(s),
- an optional resources folder if you have additional resources (e.g. images, a separate data acquisition notebook), and
- a data folder **if** your data is not simply available online (cf. Section 3).

Please put everything into a zip archive and submit the latter to Moodle.

2 Submission

There are two mandatory submissions, the pitch and the report. Submit the respective deliverables (see above) to Moodle **before 23:59 o'clock (German time)** on the day of the deadline:

1. Deadline for the pitch: Oct. 24, 2024
2. Deadline for the report: Nov. 17, 2024

Submissions after the respective deadline or via means other than Moodle will not be considered!

3 Rules and Conditions

Programming Language Use Python (for everything) and PyTorch (for the deep learning architecture and training) 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 datasets from the lectures. The document `lecture_datasets.pdf` from September 22, 2024 contains a list of datasets that are used in the modules MMS, ML, and DL. **None** of these datasets nor derivatives are 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.

4 Expectations

Your final score will be composed of 20 points for the context (set in the pitch and followed in the report), 65 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!

4.1 Context

The (possibly ficticious but) realistic context of the experiments is set in the pitch. The report should follow the story of of the pitch in a straightforward, comprehensive way. The experiments and their results' interpretation should to the goals and values proposed in the pitch.

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, architecture of your models, settings of hyperparameters, etc.) make sense and are reasonably explained.
5. The experiments use a proper, clean machine learning setup that addresses the influence of random choices and uses appropriate steps to mitigate it. In particular, note that many of the notebooks from the lecture focus only on certain aspects of the process and did NOT use proper evaluation setups (as discussed in the lectures).
6. Relevant issues – like choice of hyperparameters, overfitting, class imbalance – are addressed and handled properly.

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.

5 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!

- Before you submit, re-read the exam description and make sure that all points are adressed.
- 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.