Portfolio-Exam Tasks

Read this document (3 pages) carefully after having read the document `portfolio.pdf`!
Think 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 – Story

In the first part, explain the story and circumstances of the experiments.

- The focus of the tasks in the exam is a deep learning experiment for a **classification task**. Therefore choose and plan a story in which classification is the means to solve some problem.
- Imagine and describe a fictitious situation or describe a real scenario in which you are a data scientist working for some organization.
- The story must have a meaningful context that does or could occur in real life.
- Explain the context and the plan of the experiment and its purpose – the value you expect to create for the organization.
- The value can but does not have to be monetary. It must, however, explain what specifically your classifier (if successful) will enable.

Task 2 – The Data

In the second part, load and present a dataset.

- Explain the dataset itself (e.g., what do the features represent?).
- Explain how the dataset is suitable for the project from Task 1.
- Specifically keep the conditions for selecting datasets in mind!

Task 3 – IDA

Conduct an initial data analysis.

- Present some distributions and statistical properties that inform the reader about the dataset or that are relevant for your project.

Task 4 – EDA, Preprocessing

Bring the dataset into the form that you need for the experiments.

- Explore the data and conduct necessary transformations.
- If necessary, use different means of preprocessing until the dataset is suitable.
- If you change data, do not forget to present and summarize properties/distributions of the result.

Task 5 – Baselines

As a comparison for the upcoming tasks, compute two baselines:

- Use a very simple suitable classification baseline that does not train a model.
- Run and evaluate a classical machine learning classification algorithm. In this exam, no hyperparameter tuning is required for this baseline. Use default values of the implementation.[1]

Task 6 – Deep Learning Experiments

This is the largest and most important task in the portfolio.

- Demonstrate, describe, and interpret all necessary steps for conducting the deep learning classification experiment!
- Conduct a reproducible experiment!
- Train exactly two different architectures of Deep Learning Models.
- Choose *one* of the two following alternatives (and indicate your choice in the notebook):
  1. Choose at least one of the architectures such that it contains layers or cells beyond simple linear layers and activation functions.
  2. Conduct an optimization of the following two *model* hyperparameters: number of layers, activation functions.

  In both cases, pick all (other) model hyperparameters reasonable for the usecase (without optimization).
- Optimize the number of epochs during training (early stopping). For all other *learning* hyperparameters choose defaults where available or use typical values.
- Track the training progress in Tensorboard, add respective figures to your Jupyter Notebook, and discuss them!
- Evaluate your models using at least two suitable metrics and compare to the results of your baselines.

[1] In a real experiment, of cause proper hyperparameter tuning would be part of the experiments.

Task 7 – Conclusions and Future Work

Please address the following points separately an in that order.

1. Summarize and interpret the achieved results.
2. Recommend a course of action for the organization in your story based on the results of Task 6.
3. Reflect on limitations and possible pitfalls of using these results.
4. 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?
5. 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").