Deep Learning (Foundations): Course Project

Master's Degree Program in Data Science and Advanced Analytics,

with a specialization in Data Science (A.Y. 2022/2023)

Illya Bakurov, **Mafalda Sá Velho (project assistant)**
ibakurov@novaims.unl.pt, mvelho@novaims.unl.pt

15/03/2023

Abstract

This document describes the group project of the curricular unit Deep Learning (Foundations) and corresponds to 100% of the first epoch's evaluation. The objective of this project is to apply the knowledge acquired throughout this course to tackle a given Supervised Machine Learning (SML) problem. Problem definition and data acquisition are the groups' responsibility. Each group must register their project in a shared spreadsheet to avoid topic overlapping. The project must be delivered via moodle, without delay, until 23:59 of 08/04/2023 (Lisbon time). The defences will take place on Friday, April 14. Do your best at proving that you have strong foundations of Deep Learning.

Objectives

1. Define a Supervised Machine Learning (SML) problem of your interest that can be approached using Deep Learning (DL) techniques covered during the classes
2. Find the data
3. Perform data exploration. Relevant aspects of data distribution must be highlighted and addressed appropriately
4. Perform the necessary preprocessing based on data exploration
5. Perform sequential model construction to solve your problem in a greedy procedure that iteratively finds/builds/constructs the best model architecture starting from a simple baseline model (forward sequential model construction). Concretely:
   (a) Define a simple baseline model B
   (b) Reasonably identify an architectural component to improve upon the baseline model
   (c) Extend your baseline model E
   (d) Perform a fair comparison between B and E on the underlying problem, and identify which generalizes better
   (e) repeat
6. Perform an automatic hyperparameter search and fairly compare it with your *hand-crafted* model
7. Apply transfer learning technique (if possible) and fairly compare it with your best model
8. Optionally, try some techniques/approaches that were not covered during the lectures. If you decide on that, provide a clear description of the *out of the class* techniques in your report (includes referencing the relevant scientific literature).

Ask yourself the following questions to identify the challenges of (i) building the DL models and (ii) capturing the phenomena in the data: how to model the underlying task? Which DL architectures are appropriate? What is the trade-off between effectiveness and efficiency? Are there any problem-specific implementation choices? Which features did you identify to be the most appropriate to solve the task at hand? Etc.

All decisions made throughout the project by the group must have *a priori* rationale clearly explicit in the report. The obtained experimental results must be commented on and justified.

Evaluation Guidelines

Deliverables

The project must be delivered via moodle as a collection of `.ipynb` notebooks files in a `zip` folder. Each file has to be related to a specific step of the project (enumeration is provided in below). Before the delivery, each cell in each file must be executed to exhibit the respective (clean) output, if any. Consider the following blueprint based on the Section Objectives):

1. `explore.ipynb`: data exploration
2. `preprocess.ipynb`: necessary preprocessing of the data (if any)
3. `model_handcrafted.ipynb`: sequential model construction in a greedy iterative manner
4. `model_hyper_search.ipynb`: hyperparameter search and comparison with the *handcrafted* model
5. `transfer.ipynb`: transfer learning use case(s) and comparison with the best model
6. `optional_1.ipynb`: first optional technique
7. (...)
8. `optional_N.ipynb`: Nth optional technique.

Every `.ipynb` notebook must start with an enumeration of the group's members (names, surnames and student e-mails), (ii) a data source (a web link to access the clean and ready-to-use dataset), and (iii) a summary of the main steps and findings regarding the underlying notebook.

Evaluation criteria

This group project corresponds to 100% of the first epoch's evaluation and is given by the following formula: $0.8 * project's\ grade\ +\ 0.2 * defense's\ grade$. The groups should be made up of 4-5 students. **The *default* grade upper bound is 19, while the remaining 1 value is a function of the optional *out of the class* techniques' evaluation component**. The following (unsorted) bullet list provides details about each major criterion:

- Report-quality and language adequacy: each `.ipynb` notebook must start with (i) an enumeration of the group's members (names, surnames and student e-mails), (ii) a data source (a web link to access the clean and ready-to-use dataset), and (iii) a summary of the main steps and findings regarding the underlying notebook. The necessary references must be properly included in the document (you can use the class notebooks' reference style). The code must be *clean* and run without errors, the code cells must be preceded and followed by a text cell - the former providing an accurate and concise description of cells' content, the latter providing a comprehensive and succinct description of the output (if any). Clarity, synthesis, and objectiveness are highly appreciated
- Ability to formalize an optimization problem
- Ability to explore and prepare the data
- Ability to correctly design and implement DL models
- Ability to justify your decisions with *a priori* reasoning and comment upon the obtained experimental results
- Ability to analyze the obtained results and relate them correctly with DL concepts
- Code: the code must run without errors
- *Out of the class* (optional): if the group decides to use techniques that were not covered during the classes, the correctness of their application and the rationale behind them will be evaluated. This specific evaluation component will add up to the project's final grade a maximum of 3 values (if implemented and justified properly). [1]

Note that the set of optional *out of the class* techniques must be formalized, discussed and accepted with the project grading assistant through email in order to be considered for the evaluation (mvelho@novaims.unl.pt). The following techniques can be considered (among many others): ensemble learning, autoencoders, TensorBoard, etc.

[1] If the project will be graded with 19 and the *out of the class* component with 3, then the final grade rounds to 20.

Defense

- 15 minutes per group
- QA format (i.e., there is no need for any kind of slides for the defense)
- The absence of a group element implies 0 on his/hers defense's grade
- The defenses' schedule will be published closer to the date by the prof. Mafalda Sá Velho.

Grade Discussion (i.e., aftermath)

Students will be provided with an opportunity to optionally discuss their grades; for that, all the group members must manifest their respective intention and appear (all) in the scheduled time slot. The discussion will involve an explanation of how the evaluation criteria were applied. The outcome of the discussion can be threefold:(1) the grade is maintained, (2) the grade is increased or (3) the grade is decreased. The grade can be either increased or decreased iff it will be evident that the evaluation criteria were applied incorrectly. Additionally, the grade can be decreased (individually) if some of them group members exhibit a clear absence of knowledge regarding the project and the course.