Introduction in Machine Learning - WBAI056-2023

Assignment 2

Total points:	**40**
Starting date:	25 September 2023
Submission deadline:	**23:59, 8 October 2023**

Lecturers Dr. Matias Valdenegro, Dr. Andreea Sburlea, Dr. Marco Zullich, Dr. Tsegaye Tashu, Juan Cardenas.

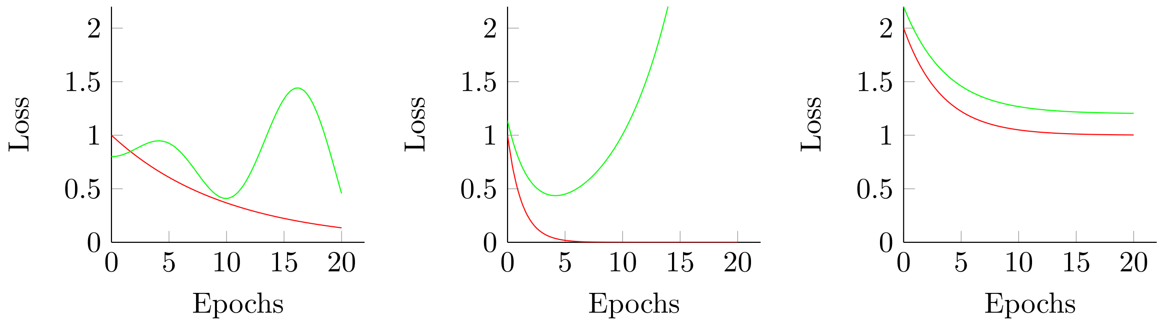
General guidelines:

- The tasks are targeted at groups of three students. Please make sure that the load is well divided: every student should contribute.
- Please take advantage of the practical sessions to ask your questions about the tasks.
- Provide a (short but comprehensive) explanation of what you are doing for each task.
- A reviewer should be able to understand plots independently; be sure to label axes, a legend for colors, use an easily readable font size, etc.
- Refer to all plots, tables, code blocks, etc. in your report.
- For the report: you can use a jupyter notebook or write a PDF in a word processor of your preference. Please include code as `.py` or `.ipynb` files as attachments in Brightspace.

Part I - Overfitting Detection

Maximum obtainable points: **10**

Consider the following plots of train (red) and validation (green) losses.



(a) !?

(b) ??

(c) ???

For each of them, argue if they show underfitting, overfitting or neither, with a clear explanation.

Part II - Splitting is Eternal? or Infinite?

Maximum obtainable points: **10**

Given knowledge from the lectures, answer the following two questions.

> You intend to train a machine learning model, is it always necessary to split your dataset? Is it always necessary to normalize your data?

Describe advantages and/or disadvantages of doing or not doing this.

Part III - Model Selection and Generalization

Maximum obtainable points: **20**

Learning Objective. Students will learn about the effect of training set size on the performance of their models.

For this exercise please choose one of the datasets mentioned below from this link: `https://scikit-learn.org/stable/datasets.html`

- California housing
- Olivetti faces
- Iris
- Diabetes
- Digits
- Linnerrud

Ideally between all student groups, all datasets should be chosen. The choice does not affect your grade.

Select a machine learning model of your choice (does not affect grade either), but appropriate for the task (classification or regression).

Train a model first on 10% of the training dataset, then on 30%, then on 50%, and finally on the entire training dataset. The smaller training sets can be obtained by sampling the original training set, which you can do using numpy.random.choice (without replacement).

Hence, train four models in total. Make sure each of your models obtains the best possible performance (hint: tune hyper-parameters). In order to compare performance of the models, the final evaluation has to be done on the same data across models[1].

Plot the performance of your four models. What do you observe in this plot? Is there a relation between the evaluation performance and the size of the training set? Write in your own words the observations that you made.

[1] Note that here we clearly say evaluation, not training.