Introduction in Machine Learning - WBAI056-2023

Assignment 2

Total points:	**40**
Starting date:	10 October 2023
Submission deadline:	**23:59, 23 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 - Curse of Example Dimensionality

Maximum obtainable points: **10**

Give two real or conceptual examples/cases that show the curse of dimensionality. These examples/cases can be made by you, or found on the internet, in the latter case you should cite and quote the example you found.

Then, argue why each of the two examples are a valid case of the dimensionality curse.

Part II - Cursed Featureless Birds

Maximum obtainable points: **10**

You have a dataset that represents types of birds. The dataset has 150 features and 100 data points. Your task is to categorize the type of birds using a clustering method (such as k-means or DBSCAN).

Explain your reasoning and possible issues for the following steps that you could take into solving your task.

- You would apply dimensionality reduction and then attempt to cluster the principal components.
- You would apply clustering directly on the raw data.

Which one of these steps makes more sense according to what you learned in the Unsupervised learning lecture?

Part III - Fun with Dimensionality Reduction

Maximum obtainable points: **20**

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
- Breast cancer
- Linnerrud
- Diabetes
- Digits
- Iris
- Wine recognition

Ideally between all student groups, all datasets should be chosen. The choice does not affect your grade, and your choice of dataset in the previous assignment is not related to this question, but we would recommend that you select a different dataset.

Discard the labels (the y array), only using the features x, use t-SNE and Multi-Dimensional Scaling from scikit-learn (package https://scikit-learn.org/stable/modules/classes.html#module-sklearn.manifold) to reduce the dimensionality of your selected dataset to two or three dimensions (choose one of them for all comparisons), and compare the structure obtained by t-SNE and MDS.

Additionally, train an autoencoder on your data with a two or three dimensional code/feature (as previously chosen), and make the same comparison with the structure obtained with t-SNE and MDS. Train your autoencoder properly, with data normalization, and test for overfitting.

For this you could use libraries like Keras and PyTorch, there are examples available at https://blog.keras.io/building-autoencoders-in-keras.html and https://lightning.ai/docs/pytorch/stable/notebooks/course_UvA-DL/08-deep-autoencoders.html. Do not forget to cite any code that you did not make yourselves.

Visually describe and compare the structure obtained in each case (three plots in total), which algorithm (t-SNE, MDS, or Autoencoder) you think describes better the selected dataset features?