Machine Learning

Practical work 13 - Dimensionality reduction

Teachers: A. Perez-Uribe (Email: andres.perez-uribe@heig-vd.ch) & J. Hennebert
Assistants: Benoît Hohl (Email: benoit.hohl@heig-vd.ch) and Simon Walther (Email: simon.walther@heig-vd.ch)

0. Notebooks and libraries

- Notebooks were tested on Google Colab
- For this practical work you will need Pandas. Pandas is a library that allows the processing of data structures like « data frames ». A « data frame » is used to store data tables. The top line of the table, called the header, contains the column names. Each horizontal line afterward denotes a data row, which begins with the name of the row, and then followed by the actual data. Each data member of a row is called a cell.

1. PCA

The objective of this exercise is to use PCA to reduce the dimensionality of a wine base, that has served as a benchmark for many Machine Learning studies. This database contains 13 features of 3 types of wines from Toscana.

First, visualize the capability of each pair of variables for explaining the three classes of wine, by means of a scatter matrix. Observe for which pairs of variables the three classes of wine appear more or less separated.

- Provide the scatter matrix and select the pair of variables (by visual inspection of the scatter matrix) that appears to allow the recognition of the three classes of wine. Explain.
- Find the smallest set of components capable of explaining at least 50% of the variance of the data.
- What is the percentage of the variance of the data explained by each one of the first 3 principal components ?
- Find the smallest set of components capable of explaining at least 60% of the variance of the data. How do you print the resulting eigenvectors ? print them and if only 3 components are required, provide the 3D projection of the original dataset.

- Find the smallest set of components capable of explaining at least 80% of the variance of the data.

2. t-SNE

The objective of this exercise is to use t-SNE to reduce the dimensionality of the MNIST database such that we can visualize the dataset in a 2D space. Datapoint appearing clustered together in the 3D space should correspond to the same digit.

- Run the notebook and observe the resulting 2D visualization of the dataset. Are there ten classes clearly separated ? provide that visualization.
- What is the dimensionality of the input data ? what is the final dimensionality at the output of t-SNE ? What is the dimensionality of the input data being fed to t-SNE ?
- Identify the values of the parameters having been used to obtain those results: perplexity, learning rate, momentum, number of iterations.
- What is the formula being used to compute the error given every 10 iterations ?
- What happens if you feed the original datapoint to the t-SNE algorithm directly ?
- Are you happy with the 2D visualization of the ten classes of datapoint ? if not, modify the parameters to get a better one. Provide the resulting visualization and then explain how did you get to it.
- Check the computational time required to perform PCA and then t-SNE or t-SNE on the raw data and compare the results.

3. UMAP

We have prepared a notebook that uses the same MNIST previous example, but this time we propose you to use UMAP as a means for reducing the dimensionality of the 28x28 inputs and to generate a 2D visualization of the observations.

- Notice the appropriate groups that are found in the UMAP output and the computational time it takes to generate those results.
- Compare them with t-SNE.

Report

Answer to the questions in this guideline sheet.