MASTER OF SCIENCE
IN ENGINEERING

HES-SO

Machine Learning

Practical work 12 - Dimensionality reduction

Teachers: A. Perez-Uribe (Email: andres.perez-uribe@heig-vd.ch) & J. Hennebert
Assistant: H. Satizabal (Email: hector-fabio.satizabal-mejia@heig-vd.ch)

Summary for the organization:

- Submit a report before Thursday 11.01.18 23h55 via Moodle.
- Modality: PDF report (max. 5 pages)
- The file name must contain the number of the practical work, followed by the names of the team members by alphabetical order, for example 12_dupont_muller_smith.pdf.
- Put also the name of the team members in the body of the report.
- Only one submission per team.

0. Notebooks and libraries

The notebooks and databases are already available in the servers. 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.

3. Auto-encoders

Yet to come…

Report

Collect the results of the experiments proposed in points 1 to 3 and comment those results. Answer to the stated questions regarding PCA, t-SNE and Auto-encoders. You do not need to provide the notebooks. This will be the last report regarding the practical work.