Machine Learning

Practical work 13 - Dimensionality reduction

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

After performing a PCA on the wine dataset, what is the percentage of the variance of the data explained by each one of the first 3 principal components ? and the accumulated explained variance ?

Randomly split the dataset into train (80%) and test (20%) datasets. Train a Multi-layer Perceptron to classify the wine types using the complete set of variables and compare the performance of a model that only uses the 3 principal components (provide the confusion matrices and compare accuracies and F1-score values).

Find the smallest set of components capable of explaining at least 80% of the variance of the data. Use these components to train a Multi-layer Perceptron to classify the

wine types using these components and compare the performance with previous results (provide the confusion matrix and compare accuracy and F1-score values).

2. t-SNE

The goal of this exercise is to use t-SNE to reduce the dimensionality of the MNIST image embeddings. These embeddings are obtained by flattening the outputs of the convolutional layers of the CNN used in PW11. The aim is to visualize the dataset in a 2D space, where datapoints that cluster together should represent the same digit.
Run the notebook and observe the resulting 2D visualization of the embeddings. Are the ten classes clearly separated? Provide that visualization.
What is the dimensionality of the embedding data being fed to t-SNE? What is the final dimensionality at the output of t-SNE?
Identify the values of the parameters used to obtain those results: perplexity, learning rate, momentum, and number of iterations.
What is the formula used to compute the error every 10 iterations?
What happens if you feed the original embeddings directly to the t-SNE algorithm without any preprocessing?
Are you satisfied with the 2D visualization of the ten classes of embeddings? If not, modify the parameters to improve the visualization. Provide the resulting visualization and explain the adjustments you made to achieve it.
Check the computational time required to perform PCA followed by t-SNE or just t-SNE on the raw embeddings and compare the results.

3. UMAP

This exercise builds upon the same MNIST example as before. The task is to use UMAP to reduce the dimensionality of the MNIST image embeddings, which are obtained by flattening the outputs of the convolutional layers of the CNN from PW11. The objective is to generate a 2D visualization of the observations.
As part of the exercise:

- Observe the distinct groups formed in the UMAP output.
- Record the computational time required to produce these results.
- Compare the performance and visualization results of UMAP with those obtained using t-SNE.

Report

Answer to the questions in this guideline sheet.