Ministry of Education, Culture, and Research of the Republic of Moldova

Technical University of Moldova

Department of Software Engineering and Automatics

Study Program: Software Engineering

Report

Data analysis and visualisation

Done by: Ion Dodon, IS211-M

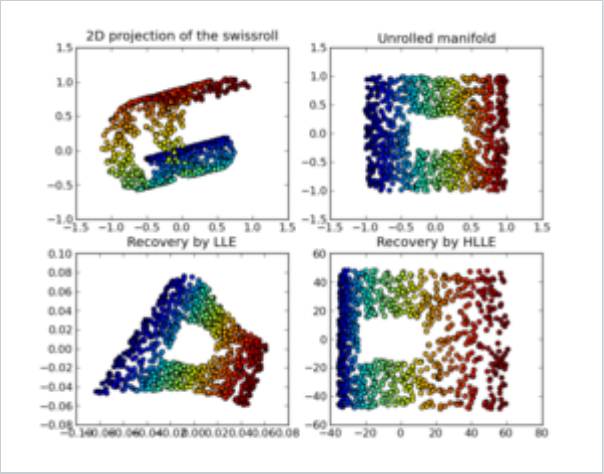
Verified by: Grozavu Nistor

Chisinau, 2021

Laboratory work no. 2

Theoretical material

Nonlinear dimensionality reduction - High-dimensional data, meaning data that requires more than two or three dimensions to represent, can be difficult to interpret. One approach to simplification is to assume that the data of interest lies within lower-dimensional space. If the data of interest is of low enough dimension, the data can be visualised in the low-dimensional space.



Top-left: a 3D dataset of 1000 points in a spiraling band (a.k.a. the Swiss roll) with a rectangular hole in the middle. Top-right: the original 2D manifold used to generate the 3D dataset. Bottom left and right: 2D recoveries of the manifold respectively using the LLE and Hessian LLE algorithms as implemented by the Modular Data Processing toolkit.

Below is a summary of some notable methods for **nonlinear dimensionality reduction**. Many of these non-linear dimensionality reduction methods are related to the linear methods listed below. Non-linear methods can be broadly classified into two groups: those that provide a mapping (either from the high-dimensional space to the low-dimensional embedding or vice versa), and those that just give a visualisation.

Locally-Linear Embedding (LLE) was presented at approximately the same time as Isomap. It has several advantages over Isomap, including faster optimization when implemented to take advantage of sparse matrix algorithms, and better results with many problems. LLE also begins by finding a set of the nearest neighbors of each point. It then computes a set of weights for each point that best describes the point as a linear

combination of its neighbors. Finally, it uses an eigenvector-based optimization technique to find the low-dimensional embedding of points, such that each point is still described with the same linear combination of its neighbors. LLE tends to handle non-uniform sample densities poorly because there is no fixed unit to prevent the weights from drifting as various regions differ in sample densities. LLE has no internal model.

Multidimensional scaling (**MDS**) is a means of visualizing the level of similarity of individual cases of a dataset. MDS is used to translate "information about the pairwise 'distances' among a set of N objects or individuals" into a configuration of N points mapped into an abstract Cartesian space.

More technically, MDS refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. It is a form of non-linear dimensionality reduction.

Given a distance matrix with the distances between each pair of objects in a set, and a chosen number of dimensions, *N*, an MDS algorithm places each object into *N*-dimensional space (a lower-dimensional representation) such that the between-object distances are preserved as well as possible. For $N = 1, 2,$ and 3, the resulting points can be visualized on a scatter plot.

Core theoretical contributions to MDS were made by James O. Ramsay of McGill University, who is also regarded as the founder of functional data analysis.

Conclusions

Working on this laboratory work I've user make_swiss_roll function from sklearn.datasets and plotted the data points user PCA transformation. Then I used LocallyLinearEmbedding and analyzed the resulting error for n_neighbours=[2..15]. After this was used MDS and TSNE and plotted the data to see the difference between them.

In part II I have imported digits dataset with 6 classes and different embedding like Random projection embedding, Truncated SVD embedding, and others were used to transform the dataset. Then the Decision Tree model was used to work with the projections given by each embedding method. Then I compared the error and score after fitting with each projection.

Bibliography

- https://en.wikipedia.org/wiki/Multidimensional_scaling
- https://en.wikipedia.org/wiki/Dimensionality_reduction
- https://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction