Foundations of Data Mining (2IMM20)
Homework Assignment 1D
Subspace Clustering

Wouter Duivesteijn

Release date: October 02, 2019.
Guided Self-Study session: October 09, 2019.
Deadline: November 01, 2019.

Preamble

We are going to explore data mining algorithms in Python. Hence, we recommend that you implement your solutions to the following questions in a Jupyter Python notebook. You are free to select another programming language were you so inclined, but if you do, the burden is upon you to provide detailed instructions on the requirements to get your solutions to work in an accompanying README file.

A general Python implementation of subspace clustering is not publicly available at the moment, neither for arbitrarily oriented nor for axis-parallel subspace clustering. However, for smaller datasets, we can explore axis-parallel subspace clustering by hand, using the standard k-means clustering implementation available in `scikit-learn`. You can find all the information you need about the relevant `sklearn.cluster` module at `http://scikit-learn.org/stable/modules/clustering.html`.

We are going to reconstruct the clustering of the Iris dataset into three categories in various subspaces, and compare the results. You can find the Iris dataset in ARFF form at `www.cas.mcmaster.ca/~cs4tf3/iris.arff`. If you are interested in more details, have a look at `https://en.wikipedia.org/wiki/Iris_flower_data_set`.

N.B.: in order to give the correct answers to the following questions, you may need to solve a simple version of the Assignment Problem [1], for instance using the Hungarian Algorithm [2].

Axis-Parallel Subspace Clustering (20 points)

α. (6 points) Use the k-means algorithm with $k = 3$ to cluster the Iris dataset, over the whole 4-dimensional input space. How well do the clusters match the actual labels?

β. (4 points) Project the dataset axis-parallel onto the dimensions `Sepal.Length` and `Sepal.Width` (hint: probably the easiest way of doing this is by simply throwing away the other columns). Then, run the k-means algorithm with $k = 3$ again on the projected dataset. Do the results improve? Do they deteriorate?

γ. (10 points) Repeat the previous exercise with each possible two-dimensional axis-parallel subspace (i.e.: five more times). In which subspace does the clustering mimic the true labels of the dataset most closely? Which type of Iris most often ends up in a wrong cluster? Which records are particularly difficult to cluster?

References

[1] James Munkres, “Algorithms for the Assignment and Transportation Problems”. Journal of the Society for Industrial and Applied Mathematics 5(1):32-–38, 1957.

[2] Harold W. Kuhn, “The Hungarian Method for the assignment problem”. Naval Research Logistics Quarterly 2:83-–97, 1955.