Linnaeus University

Data Mining, 4DV510

Rafael M. Martins

rafael.martins@lnu.se

Assignment: Exploratory Data Analysis

1 Introduction

In this assignment you will make use of your project framework to test different algorithms with data sets of your choice. You can deliver the results in either (a) a jupyter notebook, combining the code, text, and images in a nice readable sequence; or (b) send the source code and the text separately, with the full report (with text and images) in a PDF file.

For the following two exercises, you will need three different DR techniques and three different clustering techniques integrated into your project framework. You will then visualize and explore their results in a simple manner, using scatterplots. This will be a relatively open-ended task; you will choose three data sets and explore them. The only restrictions are that (a) the data sets must be multidimensional (i.e., more than 4 features), (b) they must have labels, and (c) they must have at least 1000 data points.

These are some examples of interesting places to obtain new data sets:

- `https://archive.ics.uci.edu/`
- `https://www.openml.org/search?type=data`
- `https://www.kaggle.com/datasets`

Think about the size of the data set you choose so the algorithms do not run too slow and the scatterplots do not get very crowded. It is up to you to find interesting datasets that will result in nice insights!

2 Comparison of DR Techniques

Generate a scatterplot matrix comparing the results of three DR techniques: PCA, Sammon mapping, and t-SNE, for each data set. The resulting visualization should be a 3×3 matrix where each cell is a scatterplot of a DR technique applied to a data set. Color the points by their target variables (i.e., class/labels) using a qualitative colormap.

Then answer this shortly (in a couple of paragraphs): In your opinion, which technique performed the best for each data set, regarding the separation of the classes? How are the classes in the data sets separated? Are some classes easier to separate than others?

2.1 Class Preservation

Write a function to compute the *class preservation* of each point x_i in the following way: given a certain k, return the number of neighbors of x_i (in the final two-dimensional layout) that belong to the same class as x_i, divided by k. The output will be in the interval $[0, 1]$, such that 0 means no neighbors were preserved (bad) and 1 means perfect preservation (good). Recreate the previous scatterplot matrix and show these values on each point using a quantitative colormap (such as `parula` or `hot`). If you use `jet` or `hsv` your assignment will automatically receive a zero, you will be expelled from LNU, and I will personally hunt you.[1]

Which techniques performed better regarding neighborhood preservation? How is the neighborhood preservation distributed among the points in the layout, i.e., do all points have similar neighborhood preservation or are the values different in different areas? If so, can you find a pattern for this?

[1] Seriously, though, please do not use them! (`http://idl.cs.washington.edu/papers/quantitative-color/`)

3 Comparison of Clustering Techniques

Choose one of the DR techniques from the previous exercises and generate a similar scatterplot matrix to compare the results of K-Means, DBSCAN, and Hierarchical Clustering for each data set. The resulting visualization should be a 3×3 matrix where each cell is a scatterplot of the chosen DR technique applied to a data set, with the colors of the points showing the clusters using a qualitative colormap (see, e.g., https://matplotlib.org/tutorials/colors/colormaps.html).

Then answer this shortly (in a couple of paragraphs): In your opinion, which clustering technique performed the best for each data set? How are the clusters in the data sets separated? Are some clusters easier to separate than others?

Note: The clustering should be applied to the original, multidimensional dataset, and only visualized in 2D. You should not apply the clustering to the 2D output of the DR method.

3.1 Cluster Preservation

Write a function to compute the *cluster preservation* of each point x_i in the following way: given a certain k, return the number of neighbors of x_i (in the final two-dimensional layout) that belong to the same cluster as x_i, divided by k. The output will be in the interval $[0, 1]$, such that 0 means no cluster neighbors were preserved (bad) and 1 means perfect cluster preservation (good). Recreate the previous scatterplot matrix and show these values on each point using a quantitative colormap (such as `parula` or `hot`). If you use `jet` or `hsv` your assignment... well, you get the idea.

Answer the same questions here: Which techniques performed better regarding cluster preservation? How is the cluster preservation distributed among the points in the layout, i.e., do all points have similar cluster preservation or are the values different in different areas? If so, can you find a pattern for this?