Assignment 3

Specific instructions for Assignment 3:

- The deadline is on 10 May 2022 at 23:55, Finnish time.
- The maximum number of points from this assignment is 14.
- This assignment has three exercises that must all be completed to obtain full points.

General instructions:

- The general grading criteria are available on the MyCourses web page.
- Students should complete the assignment individually, but discussions with others are encouraged. but discussions with others are encouraged. However, your final solution must be your own. Please read the *Aalto University Code of Academic Integrity and Handling Violations Thereof* for further details.
- The language of the reports should be English.
- Policy for late submissions: The score is reduced by three points per day after the deadline. (However, you cannot get negative points for an assignment.)
- If you have a pressing reason that causes you to miss the deadline, you can send an email to the lecturers (cs-e4840@aalto.fi) to request an extension without the late submission penalty. The reason must be such that it would entitle you to be absent from work (e.g., illness) and verifiable (e.g., doctor's certificate). The extension must be requested before the deadline.
- The submitted report should be in a single Portable Document Format (PDF) file. If you are using software such as Word, then export the final document as PDF. If you have several PDFs, then please merge them into one before submitting the assignment.
- Do not attach any source code.
- State your name and your student ID clearly in the report.
- Number your answers by the number of the questions, and keep the order.

Exercise 1 (7 points)

Download the dataset mysticdata.csv, which contains $n = 1000$ data points. The first real-valued column is a vector and the next (2nd–4th) real-valued columns are the data in $m = 3$ dimensions—the data matrix X. The data presents a curved point cloud in three dimensions. This exercise aims to study the shape of this three-dimensional data set by embedding it into one or two dimensions. The items (b)–(e) below should contain a brief explanation of what you have done.

(a) Make a trellis (small multiples, similar to Exercise 4 in Assignment 1) of 2-dimensional scatterplots of the point set in X such that you map the value of vector t to a suitable continuous colour scale (e.g., spectral scale).

(b) Use PCA, project the data to the first principal component, and make a plot of the data into one dimension using the same colour scale as in item (a) above. Also, make a histogram of the one-dimensional embedding. With PCA, it is a good idea to center the data first. Why? What would happen if the data would be uncentered when looking for a maximum variance projection?

(c) Then make two-dimensional plots of the data projected to the rst and second PCA components, and to the second and third components. Based on these, what can you tell about the data set's shape?

(d) Use nonmetric MDS or Sammon mapping to embed the data into one or two dimensions and plot the data the same way you did in item (b) above.

(e) Use ISOMAP, discussed in the lectures, to embed the data into one or two dimensions and plot the data the same way you did in item (b) above.

Hints: You can use any software you want to do this and the next exercise. In the lecture slides, you can find links to some examples made with R that you may find helpful. As a measure of the neighborhood, a good definition is that items i and j are neighbours if j is one of the k ($k < 10$) closest points to i or if i is one of the k closest points of j, but you can also use some other denition of neighbourhood.

Exercise 2 (4 points)

Download from MyCourses the dataset population_data.csv which contains statistics of population age structure in Finnish municipalities. The data is organized in dierent age groups with the following columns:

1. Name of the municipality
2. Total population in that area
3. People younger than 4 years
4. People of age 5–9 years

...

20. People of age 85–89 years
21. People 90 years or older

Compute and plot Metric MDS (MMDS) and Sammon mapping. Annotate some selected (or all) places, for example, main cities, provinces, places where you have been/born, etc. Compute Shepard plot (a scatter plot of output distances as a function of input distances) and compare the plots of MMDS and Sammon mapping. Which method predicts which distances better? The easiest way to do this exercise is to use Matlab. The following commands are useful:

- mdscale: computes various projections. Setting the parameter Criterion to "metricstress" produces MMDS. Setting the parameter Criterion to "sammon" produces Sammon mapping.
- pdist: computes Euclidean distance between points.

Exercise 3 (3 points)

From Mycourses, download the dataset network data.tgf, which contains a network defined as an adjacency list (explained below). Visualize it using the principles introduced in the last lecture (and the general Tufte's principles taught earlier). Explain why your visualization is appropriate for this network and how you produced it. Also, visually indicate each node's given attribute labels and try to make different network substructures visible.

You may use any software (e.g., yEd, PowerPoint, Illustrator, etc.) or even hand drawing to develop and present your solution. The TGF (Trivial Graph Format) representation used in network data.tgf is a text file containing first a list of nodes (on each line a node identifier followed by possible attributes), and then a list of edges (pairs of node identifiers), see https://en.wikipedia.org/wiki/Trivial_Graph_Format