CS-E4840 Information Visualization

Assignment 3

Specific instructions for Assignment 3:

- The first deadline is on 6 April 2018 at 23:55, local time, and the second deadline is on 29 April 2018 at 23:55, local time.
- If the student misses the 1st deadline there is no late submission penalty and the 2nd deadline automatically applies. For this reason no extension is granted for the 1st deadline. If the student misses the second deadline there is a late submission penalty (see below). The only practical difference between the deadlines is that the assignments submitted by the 1st deadline will be graded and dated earlier than the assignments submitted after it. The gradings will be completed within 4 weeks of the respective deadlines.
- You can only submit one answer to Assignment 3. If you submit an answer by the 1st deadline and then resubmit an answer after the 1st deadline your new answer replaces your old one and the 2nd deadline applies. We will not publish model answers before the 2nd deadline.
- Maximum number of points from this assignment is 14.
- This assignment has 3 exercises that must all be completed to obtain full points.

General instructions:

- The general grading criteria are available at `https://mycourses.aalto.fi/course/view.php?id=16959&section=5`
- The assignment should be completed by one person, 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 assignments is English.
- The deadline has a late submission policy: each day being late automatically reduces 3 points of the assignment. 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. Otherwise, the extension will be refused.
- 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 clearly your name and your student id in the report.
- Number your answers to correspond the questions in each assignment, and do it in order corresponding to the questions.

Exercise 1 (7 points)

Download the dataset toydata.csv, which contains $n = 1000$ data points. The first real-value column is a vector $s \in \mathbb{R}^n$ and the second and third real-valued columns are the data in $m = 2$ dimensions, denoted here by a data matrix $X \in \mathbb{R}^{n \times m}$. The data presents a spiral in two dimensions. The purpose of this exercise is to study the embedding of this two-dimensional data set into one dimension. The items (b)–(d) below should contain a brief explanation of what you have done.

(a) Make a scatterplot of the spiral in X such that you map the value of vector s 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 dimensions 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) Use nonmetric MDS *or* Sammon mapping to embed the data into one dimension and plot the data into one dimension in the same way you did in item (b) above.

(d) Use ISOMAP *or* LLE *or* Laplacian eigenmap *or* CCA, discussed in the lectures, to embed the data into one dimension and plot the data into one dimension in the same way you did in item (b) above.

(e) Define stress, precision, and recall, and compute them for the three embeddings obtained in items (b)–(d) 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 useful. As a measure of neighbourhood, a good definition is that items i and j are neighbours if j is one of the 10 closest points to i or if i is one of the 10 closest points of j, but you can also use some other definition of neighbourhood.

Exercise 2 (4 points)

Download from Mycourses the dataset M.csv which contains the results of Helsingin Sanomat questionnaire to $n = 795$ candidates in the 2017 municipal elections in Helsinki. The first columns of the CSV file give the first name, surname, and party of the candidate. The remaining $m = 49$ columns give the positions of the candidates to various claims on a scale from 1 being disagree to 5 being agree (see the lectures for an equivalent dataset from Espoo). Using the three methods you used in items (b)–(d) of Exercise 1 produce the respective two-dimensional embeddings of the data. Compare the embeddings to the ones from Espoo and presented in the lectures and point out any differences; can you say if the differences are "real" (i.e., the distribution of the candidates' answers in the two cities are really different) or just random artefacts? Briefly explain how the different principles of the three different dimensionality reduction methods used show up in the three visualisations.

Exercise 3 (3 points)

The fisheye view by Furnas[1], discussed in the lectures, tries to solve the so called focus+context problem by displaying detailed information of a chosen point, while giving only a coarse overview of more distant points. Assume that we are interested in one candidate in the 2017 municipal election data such as used in Exercise 2. Describe how you could visualise this candidate and his or her surroundings by using the ideas of Furnas' fisheye view. You do not have to do the implementation, but you should include any necessary drawings etc. that are necessary to make your idea clear.

[1] https://doi.org/10.1145/258549.258800