732A98 Visualization
Division of Statistics and Machine Learning
Department of Computer and Information Science

Laboratory work 4

Instructions

- Create a report to the lab solutions in PDF.
- Be concise and do not include unnecessary printouts and figures produced by the software and not required in the assignments.
- **Include all your codes as an appendix into your report.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- The lab report should be submitted via LISAM before the deadline.

In this lab, plots are allowed to be "ugly", i.e. copied from GGobi

Assignment 1: Analysis of Italian olive oil

In the storage of GGobi there is a file *olive.csv* that we are going to analyze. Each observation contains information about

- Region (1=North, 2=South, 3= Sardinia island)
- Area (different Italian regions)

Different acids:

- Palmitic
- …
- Eicosenoic

ATTN: All diagrams that support your judgments should be included to the report

1. Open *olive.csv* in GGobi and open Data Viewer. How many observations are present in the data?
2. Create a scatter plot matrix that shows how the contents of different acids are related to each other. Investigate the matrix to find plots where the clusters are present. Close the plot.
3. Create a scatter plot of the eicosenoic against linoleic. Based on section 2, comment why it can be interesting to investigate this pair of variables. You have probably found a group of observations having unusually low values of eicosenoic. Use identification tool to find out the exact values of eicosenoic for these observations.
4. Create a histogram that shows how many observations fall within any given region. Use persistent brushing to identify the regions that correspond unusually low values of eicosenoic. Include the plots into your report and then remove the brushing (one way is to restart GGobi)
5. Create scatter plots eicosenoic against linoleic and arachidic against linolenic. Which outliers in (arachidic, linolenic) are also outliers in (eicosenoic, linoleic)? Are outliers grouped in some way?

6. Use persistent brushing to paint by different colors the observations that fall into different regions. Keep these coloring during steps 7-9.
7. Create a parallel coordinate plot for the available eight acids. Select some proper subset of variables and define their order on the plot. Which variables can be taken for identifying clusters? (suggest at least three variables)
8. Create a 3D-rotation plot by using the variables found in step 7. Can you see clusters? Include proper screenshots motivating your answer.
9. Use all 8 acids and examine a 2D-tour. Try to find a projection with the best separation of the data into clusters. How the clusters detected are related to the regions the oils come from?
10. Based on the analysis above, try to suggest a strategy (or, maybe, several strategies) that would use information about the level of acids to discover which region the oil comes from.

Assignment 2. Multidimensional scaling of a high-dimensional dataset

The data set ***cars.xlsx*** contains information about properties of various cars, such as:

- Country: Nationality of manufacturer (eg. U.S., Japan)
- Car: Car name (Make and model)
- MPG: Miles per gallon, a measure of gas mileage
- Drive_Ratio: Drive ratio of the automobile
- Horsepower: Horsepower
- Displacement: Displacement of the car (in cubic inches)
- Cylinder: Number of cylinders

1. Load the file to R and answer whether it is reasonable to scale these data in order to perform a multidimensional scaling (MDS).
2. Write an R code that performs a metric MDS of the data (numerical columns) into two dimensions. Visualize the resulting observations in GGobi and analyze the plot.
3. Create the Shepard plot for the MDS performed and comment about how successful the MDS was.
4. Repeat steps 2-3 for the nonmetric MDS and Minkowski distance=2. Report the stress value and whether the MDS was converged.
5. Brush the MDS plot by the value of the Country. Draw conclusions.
6. Perform a nonmetric MDS for Minkowski distance p=1. How have this change affected the clustering? Provide a Shepard plot and comment on the quality of the fit.
7. Perform a nonmetric MDS for Minkowski distance p=2, randomly chosen starting points (uniformly from -1 to 1 in each dimension), and the maximum number of iterations equal to 500. How have this change affected the clustering? Have you got a better stress value? Provide a Shepard plot and comment on the quality of the fit.
8. Which of the methods do you think was the best here?

732A98 Visualization
Division of Statistics and Machine Learning
Department of Computer and Information Science

Submission procedure

Assume that X is the current lab number, Y is your group number.

If you are neither speaker nor opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline

If you are a speaker for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members does the following before the deadline:
    - submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
    - Goes to Study room *Group Y→Documents* and opens file *Password X.txt*. Then the student should put your group report into ZIP file *Lab X_Group Y.zip* and protect it with a password you found in Password X.txt
    - Uploads the file to *Collaborative workspace →Lab X* folder

If you are opponent for this lab,

- Submit your report using *Lab X* item in the *Submissions* folder before the deadline.
- Make sure that you or some of your group members submits the group report using *Lab X group report* in the *Submissions* folder before the deadline
- After the deadline for the lab has passed, go to Collaborative workspace→*Lab X* folder and download the appropriate ZIP file. Open the PDF in this ZIP file by using the password available in *Course Documents →Password X.txt,* read it carefully and prepare (in cooperation with other group members) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.