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

Laboratory work 4

Instructions

- 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; you are also recommended to show parts of the codes in the flowing text of the report.**
- A typical lab report should 2-4 pages of text plus some amount of figures plus appendix with codes.
- Create a report to the lab solutions in RMarkdown. Make sure that it is can be compiled to HTML and that all paths in RMD file are relative to the current directory where the RMD file is located. **Reports that can not be compiled are returned without revision.**
- Put the RMD file and all supporting files into one ZIP archive when you submit it to LISAM.
- The lab report should be submitted via LISAM before the deadline.

Assignment 1. High-dimensional visualization of economic data

File **prices-and-earnings.txt** shows a UBS's (one of the largest banks in the world) report comparing prices, wages, and other economic conditions in cities around the world. Some of the variables measured in 73 cities are Cost of Living, Food Costs, Average Hourly Wage, average number of Working Hours per Year, average number of Vacation Days, hours of work (at the average wage) needed to buy an iPhone, minutes of work needed to buy a Big Mac, and Women's Clothing Cost.

1. For further analysis, import data to R and keep only the columns with the following numbers: 1,2,5,6,7,9,10,16,17,18,19. Use the first column as labels in further analysis.
2. Plot a heatmap of the data without doing any reordering. Is it possible to see clusters, outliers?
3. Compute distance matrices by a) using Euclidian distance and b) as one minus correlation. For both cases, compute orders that optimize Hamiltonian Path Length and use Hierarchical Clustering (HC) as the optimization algorithm. Plot two respective heatmaps and state which plot seems to be easier to analyse and why. Make a detailed analysis of the plot based on Euclidian distance. Use Euclidian Distance matrix in all coming steps.
4. Compute a permutation that optimizes Hamiltonian Path Length but uses Traveling Salesman Problem (TSP) as solver. Compare the heatmap given by this reordering with the heatmap produced by the HC solver in the previous step – which one seems to be better? Compare also objective function values such as Hamiltonian Path length and Gradient measure achieved by row permutations of TSP and HC solvers (Hint: use criterion() function)
5. Use Ploty to create parallel coordinate plots from unsorted data and try to permute the variables in the plot manually to achieve a better clustering picture. After you are ready with

this, brush clusters by different colors and comment about the properties of the clusters: which variables are important to define these clusters and what values of these variables are specific to each cluster. Can these clusters be interpreted? Find the most prominent outlier and interpret it.

6. Use the data obtained by using the HC solver and create a radar chart diagram with juxtaposed radars. Identify two smaller clusters in your data (choose yourself which ones) and the most distinct outlier.
7. Which of the tools you have used in this assignment (heatmaps, parallel coordinates or radar charts) was best in analyzing these data? From which perspective? (e.g. efficiency, simplicity, etc.)

Assignment 2. Trellis plots for population analysis

File **adult.csv** shown data collected in a population census in 1994. The following metrics are available:

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, etc.
3. fnlwgt: a population index.
4. education: Bachelors, Some-college, etc.
5. education-num: ordered Education variable.
6. marital-status: Married-civ-spouse, Divorced, etc.
7. occupation: Tech-support, Craft-repair, etc.
8. relationship: Wife, Own-child, etc.
9. race: White, Asian-Pac-Islander etc.
10. sex: Female, Male.
11. capital-gain: continuous.
12. capital-loss: continuous.
13. hours-per-week: continuous.
14. native-country: United-States, Cambodia etc.
15. Income level

1. Use ggplot2 to make a scatter plot of Hours per Week versus age where observations are colored by Income level. Why it is problematic to analyze this plot? Make a trellis plot of the same kind where you condition on Income Level. What new conclusions can you make here?
2. Use ggplot2 to create a density plot of age grouped by the Income level. Create a trellis plot of the same kind where you condition on Marital Status. Analyze these two plots and make conclusions.
3. Filter out all observations having Capital loss equal to zero. For the remaining data, use Plotly to create a 3D-scatter plot of Education-num vs Age vs Captial Loss. Why is it difficult to analyze this plot? Create a trellis plot with 6 panels in ggplot2 in which each panel shows a raster-type 2d-density plot of Capital Loss versus Education-num conditioned on values of Age (use cut_number()) . Analyze this plot.

4. Make a trellis plot containing 4 panels where each panel should show a scatter plot of Capital Loss versus Education-num conditioned on the values of Age by a) using cut_number() b) using Shingles with 10% overlap. Which advantages and disadvantages you see in using Shingles?

Submission procedure

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

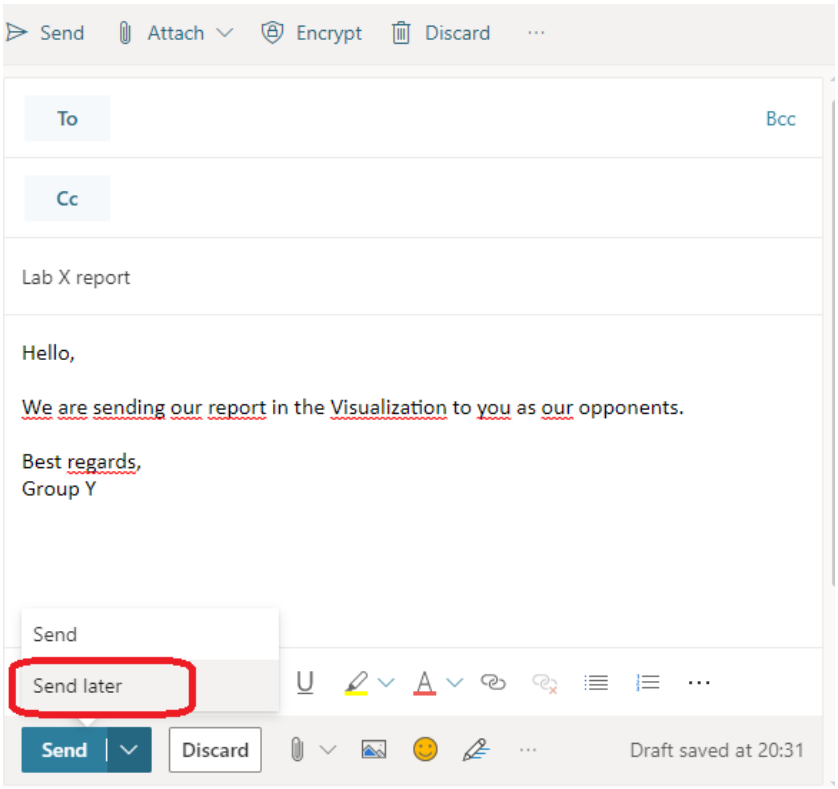
If you are neither speaker nor opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.

If you are a speaker for this lab,

- Make sure that you or your group mate does the following before the deadline:
    1. submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Makes sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
    2. Goes to LISAM→Course Documents→Deadlines.PDF, finds the deadline (date and time) for the current lab.
    3. Goes to LISAM→Course Documents→Seminars.PDF and find the group number of your opponent group
    4. Goes to LISAM→Course Documents→Groups.PDF and finds email addresses of the students in the opponent group
    5. Go to LISAM→Outlook app and in the Outlook web client creates a new message where you
        - Specify Lab X report as a title (X is lab number)
        - Specify email addresses of the opponents in the "To:" field
        - Attach your RMD report and accompanying data files (Note: NOT HTML!)
        - **Important: Click on arrow next to "Send" button, choose "Send Later" and specify the lab deadline as the message delivery time stamp (see figure below)**

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



If you are opponent for this lab,

- Make sure that you or your group mate submits the group report using *Lab X* item in the *Submissions* folder before the deadline. Make sure that the report contains the Statement Of Contribution describing how each group member has contributed into the group report.
- After the deadline for the lab has passed you should be able to receive the RMD report of the speakers per email. Compile it, read it carefully and prepare (in cooperation with your group comrade) **at least three questions/comments/improvement suggestions per lab assignment** in order to put them at the seminar.