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

Laboratory work 5

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.

Assignment 1. Data mining methods in visualization

The data to be investigated is the study of the Italian olive oils you have considered in a previous lab.

1. Import ***olive.csv*** *to* R. Transform Region column to factor by using function as.factor().
2. Classify oils using decision trees by treating Region as response and all acids as explanatory variables. Use R output to present the resulting decision tree in a graphic form. Comment which variables were selected for making decision and how deep the resulting tree is. How many items were misclassified?
3. Decide using the decision tree which two acids must be used as plot axes to demonstrate the classification results. Use package *classifly* to demonstrate the decision boundaries. Why do you think it can be a problem to detect the three regions from this plot by means of clustering?
4. Scale your data and perform a complete-link hierarchical clustering using all acids as explanatory variables. Plot the resulting dendrogram and comment on the number of natural clusters in the data. Create three clusters and use RGGobi to create a 2D-tour showing the acid contents in the oils where different clusters are marked by different colors. Does the result suggest that a proper clustering was done? Provide a screen-shot demonstrating your conclusions.

Assignment 2: Animation for time series data

The data file ***Oilcoal.xls*** provides time series about the consumption of oil (million tonnes) and coal (million tonnes oil equivalents) in China, India, Japan, US, Brazil, UK, Germany and France. Marker size shows how large a country is(1 for China and the US, 0.5 for all other countries).

1. Open the file in JMP (or import to Google documents) and create a Motion chart for this data.

2. List several noteworthy features of the investigated time series data, and try to find historical facts that could explain the observed behavior.
3. Were there countries that had similar patterns? Include trace plots to motivate your answer
4. Do you think this plot is more informative than a combination of time series plots created per country? Motivate your answer.

5. Fit a thin plate spline model with response $Oil_p = \frac{Oil}{Oil+Coal} * 100$ and predictors $Year$ and $Country$ (use package fields, function Tps())
6. The data has only 45 time points which is not enough for making a smooth video. Decide how many extra points you need to insert between the consecutive years. Afterwards, write a loop that does the following for each time point:
    a. Predicts the values of Oil_p for each country at the given time point by using predict.Krig() with parameter x specified. (function unique() can be useful here)
    b. Creates a bar plot with stacked bars in which one bar corresponds to one country, and each bar is subdivided into percentages of using oil and coal in the country (i.e., each bar has total length 100%)
7. Use the loop from step 6 and **animation** package to produce a video file and analyze the video. What are the advantages of visualizing data in this way compared to the motion chart? What are the disadvantages?

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.