Project description for report 1

Objective: The objective of this report is to apply the methods you have learned in the first section of the course,*"Data: Feature extraction and visualization"* on your own data set to get a basic understanding of your data prior to the further analysis (project report 2 and 3).

Material: You can use the 02450Toolbox on Campusnet to see how the various methods learned in the course are used in Matlab, R or Python. In particular, you should review exercise 1 to 4 in order to see how the various tasks can be carried out.

Preparation: Exercise 1–4

Understanding the data you are trying to model well is very important. You can apply very sophisticated machine learning methods but if you are not aware of potential issues with the data the further modeling will be difficult. Thus, the aim of this first project is to get a thorough understanding of your data and describe how you expect the data can be used in the later reports.

Report 1 should cover what you have learned in the lectures and exercises of week 1 to 4 covering the section *"Data: Feature extraction and visualization"* . You should consider yourself as a new employee in a company who has just been given a data set. Your job is to make a useful description of the data set for your co-workers and make some basic plots. In particular, the report ***must*** include the following items and the report will be evaluated based on how it addresses each of the questions asked below and an overall assessment of the report quality. For readability and brevity consider not using one subsection for each item.

1. **A description of your data set.**
   Explain

   - What the problem of interest is (i.e. what is your data about),
   - Where you obtained the data,
   - What has previously been done to the data. (i.e. if available go through some of the original source papers and read what they did to the data and summarize what were their results).
   - What the primary machine learning modeling aim is for the data, i.e. which attributes you feel are relevant when carrying out a classification, a regression, a clustering, an association mining, and an anomaly detection in the later reports and what you hope to accomplish using these techniques. For instance, which attribute do you wish to explain in the regression based on which other attributes? Which class label will you

predict based on which other attributes in the classification task? If you need to transform the data to admit these tasks, explain roughly how you might do this (but don't transform the data now!).

2. **A detailed explanation of the attributes of the data.**

   - Describe if the attributes are discrete/continous, Nominal/Ordinal/Interval/Ratio,
   - give an account of whether there are data issues (i.e. missing values or corrupted data) and describe them if so
   - describe the basic summary statistics of the attributes.

   If your data set contains many similar attributes, you may restrict yourself to describing a few representative features (apply common sense).

3. **Data visualization(s) based on suitable visualization techniques including a principal component analysis (PCA).**
   Touch upon the following subjects, use visualizations when it appears sensible. *Keep in mind the ACCENT principles and Tufte's guidelines when you visualize the data.*

   - Are there issues with outliers in the data,
   - do the attributes appear to be normal distributed,
   - are variables correlated,
   - does the primary machine learning modeling aim appear to be feasible based on your visualizations.

   There are three aspects that needs to be described when you carry out the PCA analysis for the report:

   - The amount of variation explained as a function of the number of PCA components included,
   - the principal directions of the considered PCA components (either find a way to plot them or interpret them in terms of the features),
   - the data projected onto the considered principal components.

   If your attributes have very different scales it may be relevant to standardize the data prior to the PCA analysis.

4. **A discussion explaining what you have learned about the data.**
   Summarize here the most important things you have learned about the data and give also your thoughts on whether your primary modeling task(s) appears to be feasible based on your visualization.

Collaboration

The usual DTU rules for collaboration applies for the reports. The main rule is that if you hand in a report, you must have authored or co-authored the content of the report for this assignment, and if your report contains text you did not write, then it must be with attribution. Notice in particular:

- If you are taking the course again, you are allowed to re-use content from a report that you previously authored or co-authored.
- If you are authoring a report together with a person who has previously taken the course, you cannot re-use that report since you did not originally author it. We recommend that you simply choose another dataset and re-write the text such that the new report can be considered original joint work by both authors.
- You are of course allowed to use the scripts, etc. supplied in this course for the reports.

The report should be 5-10 pages long including figures and tables and give a precise and coherent introduction to and overview of the dataset you have chosen. **Each group member will be main responsible for a given part of the report. You therefore have to specify who have been responsible for each part of the report as well as outline how each member contributed to the report in an appendix to the report. All reports must contain this documentation in order to be accepted.** To ensure all group members get credit for the report, make sure also to put your names and study numbers on the front page and ensure you upload the report as a group hand in and put the name of your dataset on the front page. You cannot work in groups with more than 3 students.

Please hand in the report by uploading it as a single, uncompressed .pdf file to CampusNet no later than 27 February at 13:00.