Project description for report 3

Objective: The objective of this third and final report is to apply the methods you have learned in the third section of the course on *"Recap"* in order to cluster your data, mine for associations as well as detect if there may be outliers in your data.

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 10 to 11 in order to see how the various tasks can be carried out.

Handin Checklist

- Specify **names *and* study numbers** of each group member on the front page
- According to the DTU regulations, each students contribution to the report must be clearly specified. Therefore, for each section, specify which student was responsible for it (use a list or table). **A report must contain this documentation to be accepted**
- Your handin should consist of a `.pdf` file containing the report, and the code you have used as one (or more) files with the extension `.py`, `.R` or `.m`. The reports are not evaluated based on the quality of the code (comments, etc.), however we ask the code is included to avoid any potential issues of illegal collaboration between groups. Please do not compress or convert these files.
- Reports are evaluated based on how well they address the questions below. Therefore, to get the best evaluation, address all questions
- Use the group handin feature on campusnet. **Do not upload separate reports for each team member as this will lead to duplicate work and unhappy instructors**
- **Deadline for handin is no later than 7 May at 13:00**. Late handins will not be accepted under normal circumstances

Description

Project report 3 should include what you have learned in the third part of the course on unsupervised learning and naturally follow report 1 on *"Supervised learning: Classification and regression"* and report 2 on *"Unsupervised learning: Clustering and density estimation"*. In particular, you should perform clustering, outlier detection, and association mining on your data. The report should therefore include

three sections. A section on clustering, a section on outlier detection, and finally a section on association mining.

The material to be covered in each of these three sections is outlined below 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.

Clustering: In this part of the report you should attempt to cluster your data and evaluate how well your clustering reflects the labeled information. If your data is a regression problem define two or more classes by dividing your output into intervals defining two or more classes as you did in report 2.

1. Cluster your data by the Gaussian Mixture Model (GMM) and use cross-validation to estimate the number of components in the GMM. Try to interpret the extracted cluster centers.

2. Perform a hierarchical clustering of your data using a suitable dissimilarity measure and linkage function. Try to interpret the results of the hierarchical clustering.

3. Evaluate the quality of the clustering in terms of your label information for the GMM as well as for the hierarchical clustering where the cut-off is set at the same number of clusters as estimated by the GMM.

Outlier detection/Anomaly detection: In this part of the exercise you should apply some of the scoring methods for detecting outliers you learned in Exercise 11. In particular, you should

1. Rank all the observations in terms of the Gaussian Kernel density (using leaveone-out), KNN density, KNN average relative density (ARD). (If the scale of each attribute in your data are very different it may turn useful to normalize the data prior to the analysis).

2. Discuss whether it seems there may be outliers in your data according to the three scoring methods.

Association mining: In this part of the report you are to investigate if there are associations among your attributes based on association mining. In order to do so you will need to make your data binary, see also exercise 12. (For categoric variables you can use the one-out-of-K coding format). You will need to save the binarized data into a text file that can be analyzed by the Apriori algorithm.

1. Run the Apriori algorithm on your data and find frequent itemsets as well as association rules with high confidence.

2. Try and interpret the association rules generated.

The report should be 5-10 pages long including figures and tables and give a precise and coherent account of the results of the clustering, association mining and outlier detection methods applied on your data.

Transferring/reusing reports from previous semesters

If you are retaking the course, you are allowed to reuse your previous report. You can either have the report transferred in it's entirety, or re-work sections of the report and have it evaluated anew.

To have a report transferred, *do absolutely nothing*. Reports from previous semesters are automatically transferred. Therefore, please do not upload old reports to campusnet as this will lead to duplicate work. As a safeguard, we will contact all students who are missing reports shortly after the exam.

If you wish to redo parts of a report you have already handed in as part of a group in a previous semester, then to avoid any issues about plagiarism please keep attribution to the original group members for those sections you choose not to redo.