02450 Project 3:
Project description for report 3 and presentation in week 13

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 "Unsupervised learning: Clustering and density estimation" in order to cluster your data, mine for associations as well as detect if there may be outliers in your data.

Resources: You can use the 02450Toolbox on Campusnet to see how the various methods learned in the course are used in Matlab. In particular, you should review exercise 9 to 11 in order to see how the various tasks can be carried out.

Preparation: Exercise 1-11.

Project report 3 should include what you have learned in the third part of the course on unsupervised learning. In particular, you should perform clustering, association mining and outlier detection on your data. The report should therefore contain the following three parts:

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.

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 using the script `binarize`. For categoric variables you can use the function `categoric2numeric` to get the value in a 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.

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 leave-one-out), KNN density, KNN average relative density and distance to Kth nearest neighbor for some suitable K. (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 four scoring methods.

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. A printed copy of the report is to be handed in at the beginning of the lecture Tuesday May 7th and a .pdf file of the report handed in on Campusnet.

Presentation in class week 13: Consider yourself a consultant that was hired by the owner of your data to solve one (or more) of the following machine learning problems; regression, classification, clustering, association mining and anomaly detection. Prepare a maximally 5 minute presentation that will give a brief executive summary of your findings. In the presentation you should introduce the data, the primary problem(s) you considered and your findings. Your presentation will be given in class the 7th of May. You need to upload a .pdf or PowerPoint file of your presentation on Campusnet in the student folder before class the 7th of May.

Below find a suggestion for the structure of your presentation:

1. Presentation of your data, where did you get it, what variables does it contain, what is the primary problem(s) to be solved. (Report 1)

2. Visualize the data by one or two methods that the best convey the structure of the data and include summary statistics.(Report 1)

3. Your results obtained when solving the primary problem(s) considered for your data. (Report 2 or Report 3)

4. Your conclusions. What is the main message you would like to convey to the data owner. Do you think the primary problem(s) considered are feasible. (If available, relate your results to the results others have found when analyzing the same data.)

5. (optional) include any additional interesting findings from your analysis.