Data Mining Exercise 2: 17.11.2022

Remember to enroll yourself on the examination of the current course no later than 8 days before an examination date!

1. Let us have a look at file inco13par.txt. What are the types of variables and what kind of statistic can be calculated for each variable? How many different diagnoses the data contains? What is the mean age of all patients? You may need function nanmean().

2. Replace all missing variable values (NaN) using respective mean values from each diagnosis class. Diagnose 1 for instance, can be considered to form a cluster in variable space. Find() function is useful in this task.

3. Boxplot and histogram are valuable visual tools in evaluation of variable quality and distribution. Use boxplot() function and visualize what kind of values variables UVA, US, CYM and PTR have. What can you say about age distribution over all cases in the data? (hist()).

4. You can consider the patients as points in given variable space. Calculate Euclidean distances between cases that are in rows 2, 269 and 393. What cases are closest to each other? What kind of problems you may encounter when you use Euclidean distance measure? In addition to Euclidean distance, there are plenty of others. What other distance measures you have heard of?

5. Load power consumption data from https://archive.ics.uci.edu/ml/machine-learning-databases/00616/. Remove the first row (variable names) and the first column (date info). Inspect visually (plot) all variables over the data and consider what kind of yearly information there might be.

6. Power consumption measurements have been taken using a sampling interval of ten minutes. So there are 144 measurements for each day. Extract daily measurements (time series) for each variable and inspect visually (plot) what kind of daily information There might be.

7. Select the last column from the data you loaded in the task 5. Run *k*-means (help kmeans) using the *k* value of 3. How would you interpret your result? How your interpretation changes, if you let the *k* be for instance 5 or more?