Data Mining Exercise 4: 01.12.2022

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

1. Let us consider the power consumption data as an ordinary data set and omit the idea of time series. Select the fifth case from the data and search for its nearest neighbor using Euclidean distance, Manhattan distance and Cosine distance.

2. Normalize the variables so that the means become zero and the variances become one. Repeat the first task using normalized data.

3. Scale the variables into the interval [0 1] and repeat the first task. Scaling forces the minimum values to zero and maximum values to one.

4. Let's get back to the time series idea. Find out what kind of clusters (kmeans) you can find from temperature measures using the daily measurements and *k* values of 3 and 5. (Your initial data in this task is a matrix that has 364 rows and 144 columns)

5. Use *k* means algorithm (*k*=3) for the data in the file Iris.txt. Select the 40 first cases from each category for creating the clusters. Classify the remaining ten cases from each category using the clusters you created (nearest neighbor). First column in the file is a running number for the cases. The last column gives the classes of the cases. 50 first cases belong to class 1 and so on. You can use original data and Euclidean distance in this task.

6. Reduce the dimension of Iris data by replacing the highly correlated variables using their mean (new feature is the mean of the two variables that have the highest correlation). Repeat the fifth task using the reduced data.

7. Propose another way to reduce the dimension of Iris data than that given above. Repeat again the fifth task using your reduced data.