Data mining and network analysis ITI8665

Home assignment 1

Distance function, clustering.

Starting date 21.02.2018 deadline 17.03.2017 14:00

General requirements:

- No plagiarism in any form. Please cite all the sources you used.
- Prepare your solution in such a way, that after extracting files from the archive into a single folder it may be executed on any computer with MATLAB. Data file for evaluating your solutions will follow the same structure as during the practice: single array where rows correspond to the elements and columns correspond to different dimensions.
- Prepare a short write-up with the analysis of achieved results. Maximum 2 pages 12pt (excluding graphs).
- Submit your zip archive with your solution and write-up as separate PDF file by means of Moodle environment ained.ttu.ee
- During the practice on 22.03 you will have to demonstrate your solution and will be asked few questions. Note it is mandatory to attend practice on 22.03 and demonstrate your solutions.
- If you are unsure about using some third party function contact your teacher.
- NB! Please submit two files: .zip with your implementation and all necessary files and the other .pdf containing written report.

Exercise 1. Metric function.

Program in MATLAB your own implementations of the following distance functions: Minkowsky for p=1 (Manhattan),2 (Eucledian),3, and infinity (Chebyshev), Canberra, Mahalanobis, Cosine. Your functions should work with vectors of arbitrary dimensions. You may use standard MATLAB functions for mean, standard deviation, covariance matrix but not for distance functions!

Exercise 2. Representative based clustering.

Program in MATLAB your own implementation of: k-means algorithm OR k-medoids algorithm. It is mandatory to use your own implementation of distance function here.

Exercise 3. Density based clustering.

Program in MATLAB your own implementation of: Generic grid OR DBSCAN algorithm. It is mandatory to use your own implementation of distance function here.

Exercise 4. Evaluation

Evaluate performance of your implementation for the different values of hyper parameters. Compare performance of your implementations with corresponding MATLAB functions.

Bonus Exercises:

Exercise 5.

Implement a function to determine the number of clusters using silhouette coefficient. Assumption: the shape of clusters is closed to elliptic.

Exercise 6.

Program in MATLAB your own implementation of Agglomerative clustering algorithm.