Data mining and network analysis ITI8665

Home assignment 1

Distance function, clustering.

Starting date 22.02.2019 deadline 07.03.2019 10:00

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.
- NB! NO E-MAIL SUBMISSIONS!!!
- Submit write-up as **PDF** file by means of ained.ttu.ee environment ained.ttu.ee
- Upload your code and all necessary files to https://gitlab.cs.ttu.ee grant developer rights to sven.nomm@ttu.ee (sven.nomm@taltech.ee)
- During the practice on 07.03 you will have to demonstrate your solution and will be asked few questions. Note it is mandatory to attend practice on 07.03 and demonstrate your solutions.
- If you are unsure about using some third party function contact your teacher.

Exercise 1. Metric function.

Program in MATLAB your own implementations of the following distance functions: Minkowsky for p=1 (Manhattan),2 (Euclidean),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! This exercise is necessary to complete exercises 2 and 3.

Exercise 2. Representative based clustering.

Program in MATLAB your own implementation of: k-means 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. Dataset generation

Generate the dataset, in three dimensional space, such that following conditions are satisfied:

1. Three clusters constitute the data set.
2. Clusters 1 and 2 are distinguishable only using features 1 and 2 but indistinguishable with respect to feature 3.
3. Clusters 2 and 3 are distinguishable only using features 2 and 3 but indistinguishable with respect to feature 1.

Note that simple scatter plot is not enough to illustrate your results. Implement Hopkins statistics and/or entropy function to support your findings.

Good luck!