Data mining and network analysis ITI8370/IDN 0110

Home assignment 1

Distance function, classification, clustering.

Submission deadline 18.10. 10:00

General requirements:

- No plagiarism in any form. Please cite all the sources you used.
- Prepare your solution in such a way that it may be executed on any computer with R-studio.
- Prepare a short write-up with the analysis of achieved results. Maximum 2 pages 12pt.
- Submit your code by means of https://gitlab.cs.ttu.ee
- During the practice on 18.10 you will have to demonstrate your solution and will be asked few questions. Note it is mandatory to attend practice on 18.10 and demonstrate your solution.
- If you are unsure about using some third party function contact your teacher.
- Exercises 1-3 are mandatory.
- Exercises 4-6 are bonus exercises.

Exercise 1. Distance function.

Program in R your own distance function for Minkowski (of orders 1,2,3,4,5), Canberra distances and Mahalanobis distance. Function should allow any finite number of dimensions. You are not allowed to use standard R or third party implementation of distance functions.

Exercise 2. Clustering.

Program in R your own k-means or k-medoids algorithms. Compare results achieved with these algorithms. Evaluate clustering performance for different distance functions using silhouette coefficient and ratio of intra- to inter- cluster distances.
You are not allowed to use standard R or third party implementation of distance functions and clustering.

Exercises 3. Classification.

Program in R your own function implementing k nearest neighbor's algorithm. Evaluate performance of your implementation using different distance functions. You are not allowed to use standard R or third party implementation of distance functions and k-NN classification.

Exercise 4. Classification Wrapper.

Use wrapper method (Fisher's score, k-nearest neighbor and accuracy) to find the value of hyper parameter k.

Exercise 5. Rotating surfaces.

Reference: Clustering lecture slides. Implement example demonstrating effect of rotation of underlying dataset on Mahalanobis distance.

Exercise 7. Local outlier factor.

Evaluate how values of Local Outlier Factor change with respect to the location of the data point with respect to the center of the elliptic cluster.