Data mining ITI8730

Home assignment 2

Classification

Submission deadline 07.11.2021 23:59

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. PDF format only, submit by TalTech Moodle, strictly no e-mail submissions.
- Submit your code by means of https://gitlab.cs.ttu.ee, provide the lecturer and teaching assistant (dmgolo@taltech.ee) with the developer access for your project.
- During the lecture on 09.11.2021 you will have to demonstrate your solution and will be asked few questions. Note it is mandatory to attend / participate in the lecture on 09.11.
- If you are unsure about using some third-party function contact your teacher.
- All the Exercises are mandatory.
- Generate your own data sets to demonstrate the results.
- Assignments are accepted up to one week after the deadline with the penalty of 10% for each day except Saturday and Sunday.
- NB! Students should be able to demonstrate that their implementations are able to perform.
- File naming convention: HA_2_ Name_Surname.pdf
- R codes and functions naming convention. Distance, feature selection and silhouette functions: {student_initials}_nameofthefunction.R. Main codes {student_initials}_ex_{number}_nameofthecode.R. Please avoid using capital letters.
- Please indicate (using bold letters in the beginning of your report) if you are willing to present your work in the class or online.
- The initial plan is to have defense in the hybrid mode.

The following conditions should be satisfied:

a. The students are expected to demonstrate suitability and goodness of their solutions without guidance on behalf of the lecturer.
b. Diagrams are mandatory for the exercises 2 and 3.

Exercise 1. Feature selection.

Program in R your own implementation of Fisher's. Function should allow any finite number of dimensions. You are not allowed to use standard R or third-party implementation of distance function. For this exercise: in the report just state the names of functions you have implemented.

Exercise 2. Classification.

a. Program in R your own implementation of the decision tree classifier.
b. Program in R your own implementation of K- nearest neighbors.

Exercise 3. Kernel trick.

Propose and implement in R kernel allowing to separate two half-moon clusters. Initial clusters are in 2D space and are expected to be projected into 3D space.

Exercise 4. Data preprocessing.

Propose a dataset (max dimensionality 10 plus dependent variable) such that the number of significant (independent) variables is between 3 and 5 and application of the PCA algorithm does not lead to reduce the dimensionality of the dataset. For this exercise the students are allowed to use standard or third party functions for PCA but own implementation will give a higher grade.

Good luck!