NOVA IMS Information Management School

Data Mining

Project - Second part

Ivo Gonçalves
igoncalves@novaims.unl.pt

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Objectives

- Implement a clustering algorithm in Python
- Study different algorithm parameters and how they influence the outcome
- Interpret the results in the context of the dataset being studied
- Propose and test different variations for your baseline clustering algorithm

Datasets and groups

- The datasets are the ones provided by prof. André Melo for the first part of the project
- You must select one of the available datasets to work with
- The groups are also the same
- Send me an email, with CC to prof. André Melo, to specify your group

Description

- Implement the k-means algorithm in Python using only the core language features
  - The exception is the usage of the modules needed to plot the results
- The implementation should work for any value of k
- The details of your baseline implementation are up to you:
  - Centroids initalization method
  - Centroids update rule
  - Etc

Description

- This part of the project does not directly address the tasks regarding data manipulation
- These tasks are already covered in the first part of the project
- You are allowed to use any tool (e.g., SAS) that you are already using in the other part of the project to assist you in this part of the project

Description

- Namely, you can transform the original data if that helps this second part of the project (e.g., normalizing data, removing outliers, etc)

- After having a baseline algorithm implemented, you should test different values of k and interpret the results
  - In other words, in a given clustering scenario, what does cluster 1 represent/what are the common characteristics? What about cluster 2? etc

Description

- Different non-trivial variations of the clustering algorithm are valued. For instance:
  - Defining an approach to automatically find a suitable k
  - Finding better update rules for the particular dataset considered
  - Exploring different distance measures and how they might improve the clustering process
  - Etc

Delivery

- By email until 2016-12-29 23:59 with:
  - The source code
  - A report summarizing the main results