AIN429 Data Mining Laboratory

Assignment 4: Clustering
Date Issued : 11.12.2023
Date Due : 18.12.2023

Aim of the Experiment

In this assignment, we will focus on clustering , which is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. It is a way of grouping the data points into different clusters, consisting of similar data points. You are required to implement the 3 different clustering algorithms. The assignment should be implemented as a single Jupyter Notebook. Your notebook should be clearly documented, using comments and Markdown cells to explain the code and results. At the end of this exercise, you will become familiar with clustering methods using Python libraries.

Clustering

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

The clustering technique can be widely used in various tasks. Some most common techniques:

- Market Segmentation
- Statistical data analysis
- Social network analysis
- Image segmentation
- Anomaly detection, etc.

Types of clustering algorithms

- ***Connectivity models:*** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. Examples of these models are hierarchical clustering algorithms and its variants.

- ***Centroid models:*** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category.
- ***Distribution models:*** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution. A popular example of these models is the Expectation-maximization algorithm which uses multivariate normal distributions.
- ***Density Models:*** These models search the data space for areas of varied density of data points in the data space Popular examples of density models are DBSCAN and OPTICS.

Experiment

1. Download the dataset. The dataset will be shared on the Piazza group.
2. Perform preprocessing steps that may be necessary to clean or filter the data.
3. Analyze the dataset using tables and graphs.
4. Clearly explain analysis results.
5. Apply the 3 different clustering algorithms of your choice.
6. Apply the normalization and feature selection methods and explain its effect on clustering.
7. Compare the performance of clustering algorithms using tables and graphs.
8. Summarize and interpret your results.
9. You should submit your codes and report as a single Jupyter notebook.

Background information

We provide with you some references related to clustering.

- https://scikit-learn.org/stable/modules/clustering.html
- https://www.javatpoint.com/clustering-in-machine-learning
- https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
- https://github.com/krasserm/machine-learning-notebooks
- https://machinelearningmastery.com/clustering-algorithms-with-python/

Grading

You will present your projects during laboratory hours.

- Import dataset and Preprocessing (%15)
- Visualization (%15)
- Implementing methods (%40)
- Report (%30)

REMARKS:

- Submission format:
    - studentID_name_surname_hw4.ipynb

- Your submission should be matched with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use an online submission system to submit your experiments.
- https://submit.cs.hacettepe.edu.tr/ Deadline is 23:59. No other submission method (such as; CD or email) will be accepted.
- Do not submit any file via email related to this assignment.
- The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group.