CSE 304 Homework 2: Clustering

Introduction

This assignment aims to deepen your understanding of two key clustering algorithms: k-means++ and DBSCAN, core concepts of the algorithms briefly covered during our lectures. This task will involve studying, analysing, implementing, and conducting experiments with the specified algorithms.

Instructions and Contact

Please carefully read the assignment manual for detailed instructions. Should you have any questions, do not hesitate to reach out to the instructor, Junghoon Kim, at junghoon.kim@unist.ac.kr. For HW #2, there is no feedback session like HW #1

Topics and References

Topic 1: k-means++ Algorithm

Reference: Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." Soda. Vol. 7. 2007.

https://courses.cs.duke.edu/spring07/cps296.2/papers/kMeansPlusPlus.pdf

Topic 2: DB-SCAN algorithm

Reference: Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." KDD. Vol. 96. No. 34. 1996.

https://dl.acm.org/doi/10.5555/3001460.3001507

Minimum requirement

k-means++: Develop a method to determine the appropriate number of clusters (k). Implement your approach and report the accuracy of your results using the provided dataset.

DBSCAN: Design a method to select appropriate parameters for mu and epsilon. Implement your approach and report the accuracy using the provided & your dataset.

Implementation and Data

Sample datasets can be downloaded from Blackboard. The data is in a CSV format, with each line representing one data point. Your implementation should accept the input file name and parameters. The program should operate as follows:

k-means++) The input parameter may include a value for k. If not specified, k must be estimated. Examples:

```
> java A2_G2_t1 ./artd-31.csv 15

Cluster #1 => p301 p302 p303 p304 … p2400

Cluster #2 => p4651 p4652 p4653 p4543 … p5000

…
```

```
> java A2_G2_t1 ./artd-31.csv

estimated k: 15

Cluster #1 => p301 p302 p303 p304 … p2400

Cluster #2 => p4651 p4652 p4653 p4543 … p5000

…
```

DBSCAN) The input parameter may include values for mu and epsilon. If the input is an integer, it represents mu; if a floating number, it represents epsilon. Examples:

```
> java A2_G2_t2 ./artd-31.csv 5 0.5

Number of clusters : 26

Number of noise : 2113

Cluster #1 => p1 p2 p3 … p1086 p1088

Cluster #2 => p16 p76 p1024 … p2500

…
```

```
> java A2_G2_t2 ./artd-31.csv 0.5

Estimated MinPts : 4

Number of clusters : 23

Number of noise : 2080

Cluster #1 => p1 p2 p3 … p2479
Cluster #2 => p16 p76 p1024 … p2500

…
```

```
> java A2_G2_t2 ./artd-31.csv 4

Estimated eps : 0.5

Number of clusters : 23

Number of noise : 2080

Cluster #1 => p1 p2 p3 … p2479

Cluster #2 => p16 p76 p1024 … p2500

…
```

* Note that estimated mu(MinPts) and eps are fake values.

* A# indicates assignment number (HW number), G2 indicates group 2, and t1 indicates topic #1 (here, k-means++).

- As previously discussed, our homework schedule has changed. For HW#2, the maximum page limit is adjusted from 10 to 15 pages.

Submission Guidelines

- Please adhere to the submission format provided in the above description.

- Only one git repository per team is allowed; avoid creating too many nested folders.

- You are not required to submit the code itself; instead, provide the git link in your report.

- Avoid modifying the given format if possible.

Deadline: 2024-05-31 23:59 (Late submission is not accepted)