CSE 347/447 Data Mining: Project 1 – Clustering Analysis

Due on 11:59 PM, March 8, 2024

Standard and General Requirements

- Note that copying code/report from another student or source is not allowed and may result in an F in the grade. Partial credit will be given for partial solutions.
- Discussion is allowed at the level of technical conversation only. Students are expected to abide by Lehigh Academic Integrity Policy.
- Partial credit will be given for partial analysis/solutions, but not for long off-topic discussion that leads nowhere. Overall, think before you write, and try to give concise and crisp answers.
- You can use any programming language you like to implement this project, but Python is preferred.
- **Late policy:** You can be at most 3 days late; for every late date you lose 10% of your grade, unless some other arrangement is agreed to before the due date.
- **Submission Instructions:** Please submit your Code and Report as .zip file to CourseSite.

Datasets for Clustering:
You will use the below two types of datasets for clustering analysis and practice. You can download them at: `https://drive.google.com/drive/folders/1hDrjq3aktFfvQ2lOV1Yr1HERnAL5vAxF?usp=drive_link`

- **Simulation Datasets**: Square and Elliptical are two simulated datasets, as illustrated in Fig. 2.
- **Real-Word Datasets**: Cho and Iyer are two gene sequences datasets. A short description of these datasets can be found at `https://drive.google.com/file/d/1EqKZngBeOdLoYsypTQRAa9czNO711lv1/view?usp=drive_link`.

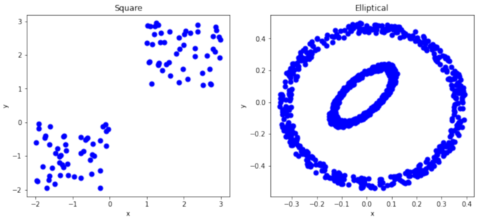

Figure 1: Visualization of Square and Elliptical datasets.

Complete the Following Tasks on Simulation Datasets:

- Implement k-means and spectral clustering algorithms to find 2 clusters on Square and Elliptical datasets and visualize your results. Compare the two methods and discuss their pros and cons. Fig. 2 shows the ideal clustering results for your reference.
- Discuss the effects of centroid initialization on k-means clustering results.
- Present the performance analysis of the spectral clustering algorithm using different similarity measures like cosine similarity and Gaussian kernel similarity (set an appropriate bandwidth parameter for Gaussian kernel).
- Present the performance analysis of the spectral clustering algorithm using different Laplacian matrices like unnormalized Laplacian and normalized symmetric Laplacian.
- **Note**: You are allowed to call k-means and spectral clustering functions from any packages or libraries if needed, but be sure you know how to properly use them.

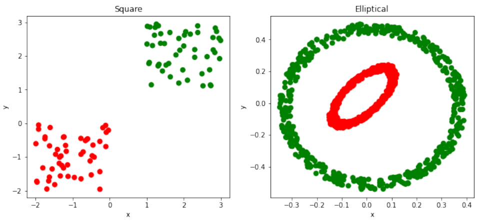

Figure 2: Ideal clustering results on Square and Elliptical datasets.

Complete the Following Tasks on Real-World Datasets:

- Use k-means and spectral clustering algorithms to find clusters of genes on Cho and Iyer datasets which exhibit similar expression profiles. Note that there are some noise samples in Iyer dataset where a label of "-1" means outliers, you should consider preprocessing your data.
- Validate your clustering results using the following methods:
  - External Index: Use the Accuracy measure to compare the clustering results between k-means and spectral clustering algorithms on Cho and Iyer datasets (the ground truth clusters are provided in the data sets).
  - Internal Index: Use the Sum of Squared Error (SSE) measure to compare the clustering results between k-means and spectral clustering algorithms Cho and Iyer datasets.
- Discuss the impact of data normalization on k-means and spectral clustering results based on Cho and Iyer datasets in terms of clustering accuracy (i.e., Accuracy).
- Discuss the impact of noise on k-means and spectral clustering results based on Iyer dataset in terms of clustering accuracy. Namely, compare the results with and without noise data.

Your final submission of .zip file should include the following:

- Code: Two clustering algorithms implemented with two functions, respectively.
- Report (PDF file): Please be sure to answer each question sequentially and adhere to the specific guidelines for each section. You can either use tables or figures to discuss your results or any findings you get from the experiments.