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

Due on 11:59 PM, March 17, 2021

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 solutions, but not for long off-topic discussion that leads nowhere. Overall, think before you write, and try to give concise and crisp answers.
- **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. Schedule a meeting with TA (zhk219@lehigh.edu) after submission to demo your code.

Datasets: Cho and Iyer are gene sequences datasets which need to be clustered according to the similar gene patterns. You can download them at: `https://drive.google.com/open?id=1I00QuYllu0ahnEy-hkEN_MTwATm2uXiZ` and `https://drive.google.com/open?id=1KHTaMp_9yfID6Qlwg1YsLnfmwu2g2C7A`.

A short description of the datasets can be found at `https://drive.google.com/open?id=1Nm7lbYSUsIp8cFGq7UQmgJUhCcY2yuqT`.

Complete the following tasks:

- Implement k-means and spectral clustering algorithms (step by step) to find clusters of genes which exhibit similar expression profiles. Compare the two methods and discuss their pros and cons.
- Directly call k-means and spectral clustering functions from packages is not allowed.

You are required to validate your clustering results using the following methods:

- Choose an external index and compare the clustering results from different clustering algorithms with an external index (the ground truth clusters are provided in the data sets).
- Choose an internal index and compare the clustering results.

Your final submission should include the following:

- Code: Two clustering algorithms implemented with two functions, respectively. Your code should allow the users to choose either of the clustering algorithm. Just an example:
  cin 'the method you choose:'
  cout 'clustering result is:'
- Report: Describe the flow of all the implemented algorithms. Compare the performance of these approaches using external and internal index on the two given datasets. State the pros and cons of each algorithm and any findings you get from the experiments.

Note*: If you choose MATLAB (every Lehigh student has free access), you could use the GUI interface with menu bars. This is not required but you will find it useful. Enjoy yourself, and if you have any question, feel free to ask.