Project 2: Clustering Algorithms

Demo time: Start from 10:30am, Nov. 5
Code&report submission due: 11:59pm Nov. 4

Two gene datasets (*cho* and *iyer*) can be found on Piazza. Please check the README file first for a short description of the two datasets.

Complete the following tasks:

1. Implement five clustering algorithms to find clusters of genes that exhibit similar expression profiles: K-means, Hierarchical Agglomerative clustering with Min approach, density-based, mixture model, and spectral clustering. Compare these five methods and discuss their pros and cons.

For each of the above tasks, you are required to validate your clustering results using the following methods:

- Using external index (Rand Index and Jaccard Coefficient) and compare the clustering results from different clustering algorithms. The ground truth clusters are provided in the datasets.
- Visualize data sets and clustering results by Principal Component Analysis (PCA). You can use the PCA you implemented in Project 1 or use any existing implementation or package.

Your final submission should include the following:

- Codes: A folder named *Code*, that contains five clustering algorithms and a *README* that shows how to run your code.
- Report: A pdf file named *Cluster_report.pdf*. Describe your implementation details about all the algorithms. Compare the performance of these approaches using visualization and external index on the two given data sets. State the pros and cons of each algorithm and any findings you get from the experiments.

Project Submission:

1. Your final submission should be a zip file named as *project2.zip*. In the zip file, you need to include aforementioned folder *Code* and folder *Report*.
2. Log in any CSE department server and submit your zip file as follows:

   >> submit_cse601 project2.zip

The details about Demo will be released two days before the demo date **through Piazza**. Note:

- New data sets will be given to check your implemented clustering algorithms and validation measures. The data format is the same as that of the project data sets we have provided.
- During the demo, the specific parameter setting will be changed, so you need to keep your code flexible enough to allow for the change in the parameters.