Project 2: Clustering Algorithms Demo

Time: October 31 2019 10:30AM ~ 2:00PM
Location: 341 Davis Hall

Please bring the following items to Demo:

- A hard copy of your report
- Your laptop, on which you could run and show your implementation
- Your UB card

Four NEW data sets have been posted on Piazza. The data format is the same as the ones we already provided, which is described in *README* file.

You need to run your code to complete the following tasks using provided parameters and settings, and show your results during demo (please do NOT include these results in your report). In addition, questions related to these algorithms will be asked to test your understanding about the tested methods.

For each task you need to visualize the datasets and show your external index (Rand Index and Jaccard Coefficient).
When visualizing the datasets you need to:

1. Color the data points according to the labels obtained by the clustering algorithm.
2. For the datasets with more than two features please use PCA to perform dimension reduction first and then visualize the two-dimensional data.
3. For the datasets with only two features, please just visualize the original data and do not perform PCA on these datasets.

Tasks:

1. Run your K-means implementation on *new_dataset_1.txt* with specific parameter setting. For example, set the number of clusters to be 3, the max iteration number to be 10, and use data points with IDs [3, 5, 9] as initial centers.
2. Run your Hierarchical Agglomerative clustering with Min algorithm on *new_dataset_2.txt* to get N clusters. N will be given during the demo. Except for the visualization and external index, you need to draw a dendrogram which shows the steps of merging on a given piece of paper.
3. Run your DBSCAN algorithm with *DBSCAN.txt*. You will be provided Minpts and Eps (ε) during the demo.
4. Run your Gaussian Mixture Model on *GMM.txt*. The initializations of the following parameters will be given: mean ($\boldsymbol{\mu}$), covariance matrix (Σ) and prior cluster probabilities ($\boldsymbol{\pi}$). In addition, the maximum iterations, convergence threshold, number of clusters and smoothing value will also be given.
   For example, you may be given the following settings in the demo: $\boldsymbol{\mu} = [[\mathbf{0},\mathbf{0}],[\mathbf{1},\mathbf{1}]], \Sigma = [[[1,1],[1,1]],[[2,2],[2,2]]$, $\boldsymbol{\pi} = [\mathbf{0.5},\mathbf{0.5}]$. The number of

clusters is 2, convergence threshold 1e-9, maximum iterations 100, smoothing value 1e-9.
Except for the visualization and external index, you also need to output the estimated parameters ($\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}$) after the iterations.

5. Run your Spectral Clustering algorithm on new_dataset_1.txt. You need to build a fully connected graph using Gaussian kernel similarity, where sigma will be given. After transforming the data into the new embedded space, you are required to use KMeans to get clusters. The initial data IDs and number of clusters will also be given.

For the first three tasks, please use Euclidean distance as the distance metric, while for the Spectral Clustering, you are required to use Gaussian kernel similarity. You can use existing package to calculate these distances and the graph degree. When performing clustering in the new embedded space in Spectral Clustering with KMeans, you can call existing KMeans library. But for the first task you must implement KMeans from scratch by yourself.

Please note:

1. For each part, we will ask questions to test your understanding about the tested methods. Your final score is based on your report, your submitted code, your demo, and your answers to those questions.

2. The demo will be divided into three parts. Each group will be given only 5 minutes to finish each part of the demo. Please make sure your code can run smoothly and your results can be reproduced by your submitted code. We will NOT give extra time.

3. Please arrive five minutes before your scheduled demo time and Do Not be late.