Homework 2: Clustering

Code and report due: October 09, 2022, 11:59pm

Problem 1 Write the calculation for the first iteration (i.e., cluster assignment and cluster center computation) of K-means clustering on the following 9 data points. Use Manhattan distance as the distance function. Note that you will need to do cluster assignment by calculating the distance between each point and each cluster center.

Table 1: Data

Data	Feature 1	Feature 2
A1	7	5
A2	-1	4
A3	5	7
A4	2	6
A5	-2	6
A6	5	1
A7	0	5
A8	2	-1
A9	3	6

Suppose there are 3 clusters, and the initial centers are C1 = (3,1), C2 = (1,4), C3 = (5,6). If the distances between the data point and the center points are the same, just randomly choose one center.

Problem 2 Suppose we have 5 data points and the pairwise distance matrix is shown in Table 2. Conduct hierarchical clustering using MIN, MAX, and AVERAGE as the inter-cluster distance respectively, and draw the dendrograms.

Table 2: Distance Matrix

Distance	A	B	C	D	E
A	0	5.5	3.5	5.0	1.0
B	5.5	0	7.5	7.0	6.5
C	3.5	7.5	0	4.0	6.0
D	5.0	7.0	4.0	0	4.5
E	1.0	6.5	6.0	4.5	0

Problem 3 The eigenvectors of covariance matrix of a 3-D dataset are a_1 = [0.24, 0.96, 0.11]^T, a_2 = [0.97, -0.25, 0.06]^T, and a_3 = [-0.09, -0.10, 0.99]^T. The corresponding eigenvalues are λ_1= 585.28, λ_2= 56.58, and λ_3 = 4.45. Given the following adjusted data points (data points after subtracting the mean), transform them to two-dimension by using PCA.

Table 4: Data

x_1	x_2	x_3	y_1	y_2
2	0	1	$y_{11} = ?$	$y_{12} = ?$
4	2	0	$y_{21} = ?$	$y_{22} = ?$

Problem 4: Given a squashing function g(*z*) = 1/ (1 + exp(−*z*)). Show that (i.e., proof) it satisfies the property g(−*z*) = 1−g(*z*). Also show that its inverse is given by g−1(*y*) = ln (*y*/ (1 − *y*)).

Problem 5 (Programming Assignment): In this task, you are asked to the implement five clustering algorithms learned in class. Everyone should submit their codes and a report via ICON.

Recommended Programming Language:
Python

Dataset Description:
Two gene datasets (cho.txt and iyer.txt) can be found on ICON.
Dataset format: Each row represents a gene:
1) the first column is gene_id.
2) the second column is the ground truth clusters. You can compare it with your results. "-1" means outliers.
3) the rest columns represent gene's expression values (attributes).
Please check the README file first for a short description of the two datasets.

Required Tasks:

1. Implement all five clustering algorithms: 1) K-means, 2) Hierarchical Agglomerative clustering with two different inter-cluster distances, 3) density-based, 4) mixture model, and 5) spectral to find clusters of genes that exhibit similar expression profiles for both datasets.
   NOTE: *You **should not** directly call any existing function or package that implements K-means. K-means algorithm should be implemented by yourself. For the rest, you are free to use any existing packages*.

2. For each of the clustering algorithms, you are required to validate your clustering results using the following methods:
   - Using external indexes such as Rand Index and Jaccard Coefficient compare the clustering results of different clustering algorithms. State the pros and cons of each algorithm and findings you get from the experiments. (Note: The ground truth clusters are provided in the datasets)
   - Visualize the datasets and clustering results by Principal Component Analysis (PCA).

3. Prepare your submission. Make a zipped folder named “[HawkID]-*Clustering*.zip”, where "[HawkID]" refers to your HawkID. In the folder, you should include:
    a. Report: (i) Solutions to the problem from 1-4. (ii) Describe the 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. The filename for your report should be “clustering_report.pdf”
    b. A folder named *Code*, which contains all codes used in this part. Inside the folder, please have a file *README* which describes how to run your code. The folder name for your code should be “code.zip”.
    c. An ICON assignment page has been created for homework2. Please submit your zipped folder there.

4. Your report should include “analysis” along the following lines:
    a. How was the initialization for K-means performed? Did you find any empty clusters? If so, how did you handle them?
    b. Do the choice of distance metric (i.e., Euclidean or Manhattan) have any impact on the performance of K-means? What is the best value for *K* on two datasets? What values of K did you experiment with on both datasets and why?
    c. How does the plot of “Sum of Squared Error VS K” look like on two datasets for K-means?
    d. Did you find any outliers in the datasets? If so, how did the K-means handle them?
    e. How comparable is the PCA visualization for your implementation of K-means and the PCA visualization of ground truth clusters? Do the plots look similar?
    f. Did the choice of inter-cluster distance (i.e., single link and complete link) have any impact on the results of hierarchical clustering? Is one better than the other?
    g. What values of K did you experiment with for hierarchical clustering? What trend do you observe for Jaccard Index vs K in the hierarchical clustering and why?
    h. What range of values (for eps and minpts) did you experiment with for DBSCAN algorithm? Any optimal values?
    i. Please perform similar analysis for GMM and Spectral clustering. Finally, can you provide a summarizing table showing the best Jaccard indexes from each algorithm, as well as the parameters that produced them? What is the best performing algorithm for each dataset that produced a reasonable number of clusters?

Note that copying code/results/report from another group or source is not allowed and may result in an F grade.