Assignment 5

Write jupyter notebook scripts for all of the questions.

Check validity of clustering results using within-cluster Sum of Squares (SSE) and between-cluster Sum of Squares (SSB)

1. Cell line data (50 points)

Cluster validity could be measured using **cluster cohesion** and **cluster separation**. Cluster cohesion measures how closely related data objects are in one cluster and cluster separation measures how distinct or sell-separated a cluster is from other clusters. SSE is defined as

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

where C_i denotes the ith cluster and k is the total number of clusters. m_i is the centroid of the ith cluster and x denotes a data object (a sample or a variable).

SSB is defined as

$$SSB = \sum_{i=1}^{k} |C_i|(m - m_i)^2$$

where m is the grand mean (i.e. the centroid) of all of the data objects and $|C_i|$ is the size of cluster i, that is the total number of data objects in cluster i.

- Let $k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, respectively, for kmeans, compute SSB and SSE for each k and make a plot that is similar to Figure 1.

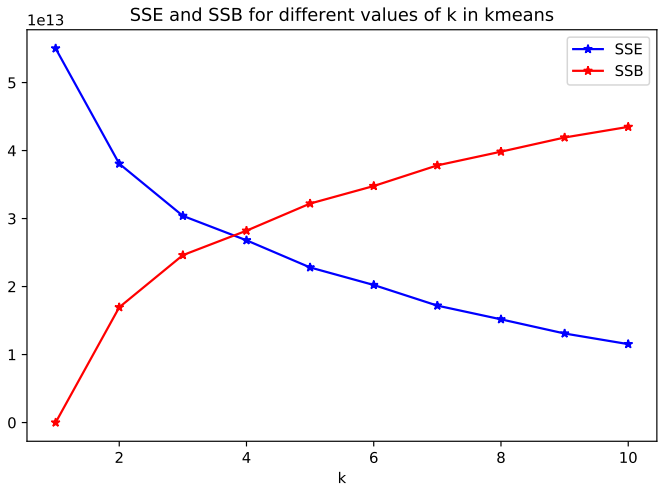

Figure 1: SSE and SSB for different values of k in kmeans.

- For $k = 2$, compute the SSE and SSB for k-means, hierarchical, and DBSCAN. For hierarchical clustering, consider single, complete, average, centroid, AND ward linkages. For DBSCAN, can you find values for parameters `eps` and `min_samples` that will allow you to get meaningful clustering results? After SSE and SSB are computed, make a figure as shown in Figure 2.

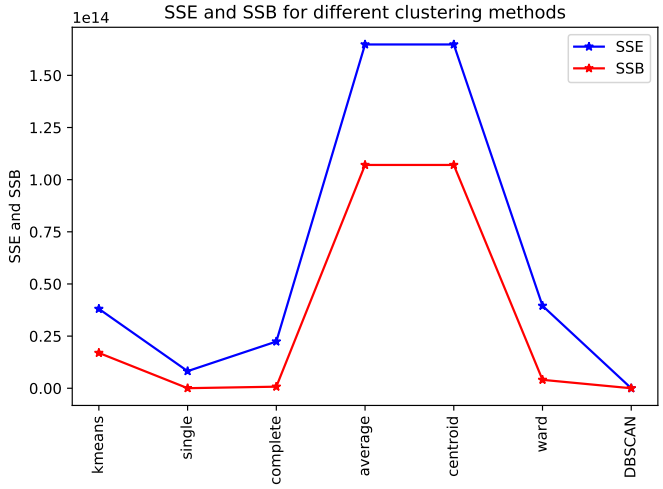

Figure 2: SSE and SSB for different clustering methods.

- Save the cluster results in a csv file as shown in Figure 3. Note that this example table only shows part of the results and you need to save the complete results. A convenient way to save the clustering results with the row names and column names is to use pandas DataFrame to store your SSE and SSB.

	kmeans	single	complete	average	centroid	ward	DBSCAN
NSCLC_A549_1	1	2	1	1	1	1	-1
NSCLC_H1703_2	1	2	1	1	1	1	-1
NSCLC_H1703_1	1	2	1	1	1	1	-1
NSCLC_A549_2	1	2	1	1	1	1	-1
NSCLC_H1437_1	1	2	2	1	1	2	-1
NSCLC_H2228_1	1	2	2	1	1	1	-1
NSCLC_H2228_2	1	2	2	1	1	1	-1
NSCLC_H1437_2	1	2	2	1	1	2	-1
NSCLC_H3122_1	1	2	2	1	1	2	-1
NSCLC_H322_2	1	2	2	1	1	1	-1
NSCLC_H322_1	1	2	2	1	1	1	-1
NSCLC_H358_2	1	1	1	1	1	1	-1
NSCLC_H3122_2	1	2	2	2	2	2	-1
NSCLC_H522_1	1	2	2	1	1	1	-1
NSCLC_H522_2	1	1	2	1	1	1	-1
NSCLC_HCC4006_1	1	2	2	1	1	1	-1
NSCLC_H358_1	1	1	1	1	1	1	-1
NSCLC_PC9_1	1	2	1	1	1	1	-1
NSCLC_PC9_2	1	2	1	1	1	1	-1
NSCLC_HCC4006_2	1	2	1	1	1	1	-1
SCLC_86M1_2	0	2	2	1	1	2	-1
SCLC_86M1_1	0	2	2	1	1	2	-1
SCLC_16HV_1	0	2	2	1	1	2	-1
SCLC_16HV_2	0	2	2	1	1	2	-1
SCLC_DMS79_1	0	2	2	1	1	2	-1
SCLC_DMS79_2	0	2	2	1	1	2	-1
SCLC_H187_2	0	2	2	1	1	2	-1
SCLC_H187_1	0	2	2	1	1	2	-1
SCLC_H209_1	0	2	2	1	1	2	-1
SCLC_H524_1	0	2	2	1	1	2	-1
SCLC_H209_2	0	2	2	1	1	2	-1
SCLC_H524_2	0	2	2	1	1	2	-1

Figure 3: Clustering results using kmeans, hiarchical, and DBSCAN.

- What conclusion can you draw from the clustering results you obtained?

2. Simulated data

Repeat problem 1 with simulated data stored in `simulated-data.csv`. The simulated data has two variables named $X1$ and $X2$ as shown in Figure 4.

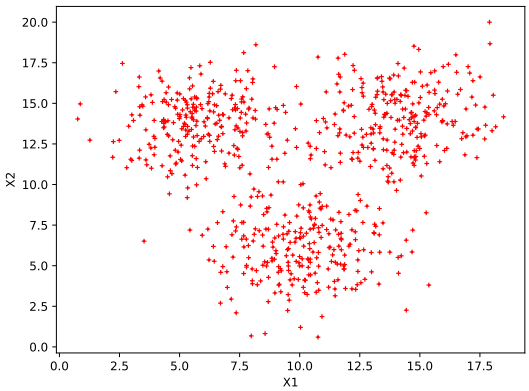

Figure 4: Simulated data.

- Let $k = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, respectively, for kmeans, compute SSB and SSE for each k and make a plot that is similar to Figure 1.
- For $k = 2$, compute the SSE and SSB for k-means, hierarchical, and DBSCAN.
  - For hierarchical clustering, consider single, complete, average, centroid, AND ward linkages. After SSE and SSB are computed, make a figure as shown in Figure 2.
  - In addition, for EACH linkage in the hierarchical clustering, make two plots as shown in Figure 5.

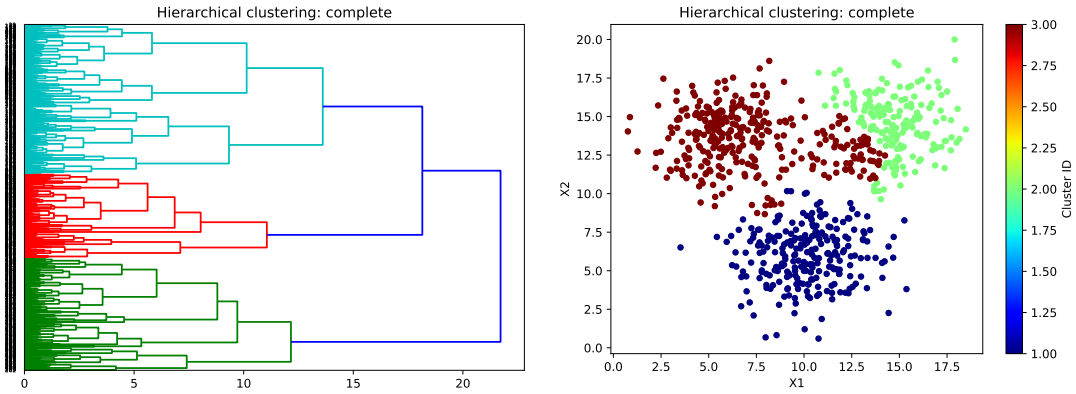

Figure 5: Results from a particular linkage in hierarchical clustering.

  - For kmeans and DBSCAN each, make a figure that is similar to the scatter plot in Figure 5
- What conclusion can you draw from the clustering results you have obtained?