The following problems are for you better understanding of the material. Solving them is not required, but recommended. We will not grade your solutions, but we will release a solution. So, please do **NOT** submit your answers via Submit Admin.

Practice problems of Basic Cluster Analysis:4, 9, 11, 12, 16, 32 of chapter 8 in our textbook.

Homework #2

Due date: November 22

The following homework is a practical exercise to learn about cluster analysis. This part of the homework will be graded. Please do submit your solutions via Submit Admin. You may use R, Matlab, etc., or code the necessary algorithms by yourself.

Deliverables: please submit the answers to the questions below as a single pdf document. Please also include in this document a visualization of **all** computed clusterings. In addition, please provide all analysis scripts, plus a specification of the platform and the corresponding data analysis software used in an **additional program file**.

k-means clustering

1.(15 points) Use k-means on the **non-noise data** by setting k=2,3,4,5,6,7,8,9,10. For **each value of k,** compute the **SSE (sum of squared error)** of clustering result, and the **average silhouette** of clustering result. Plot the **SSE curve and silhouette curve** w.r.t the various k values. (You can use any kind of distance measure.)

2. (15 points) Use k-means on the **noisy data** by setting k=2,3,4,5,6,7,8,9,10. For **each value of k**, compute the **SSE** of clustering result, and the **average silhouette** of clustering result. Plot the **SSE curve and silhouette curve** w.r.t the various k values.

3. (5 points) Can you tell the “correct” number of clusters from the SSE curve and silhouette curve? What is the correct number?

Hierarchical clustering method

4. (15 points) Use agglomerative hierarchical clustering method on the **non-noise data**. Try 4 different setting of linkage,

* single (min, shortest)
* complete (max, furthest)
* average (average distance)
* ward (inner squared distance, minimum variance algorithm)

Compute the “Cophenetic correlation coefficient” for each clustering result with different setting of linkage. Make a table to show the 4 Cophenetic correlation coefficients and tell which one is the best.

5. (15 points) Use agglomerative hierarchical clustering method on the **noisy data**. Try 4 different setting of linkage,

* single (min, shortest)
* complete (max, furthest)
* average (average distance)
* ward (inner squared distance, minimum variance algorithm)

Compute the “Cophenetic correlation coefficient” for each clustering result with different setting of linkage. Make a table to show the 6 Cophenetic correlation coefficients and tell which one is the best.

6. (5 points) Are the best settings of linkage in Question 4 and Question 5 the same? If they are not the same, why ?

DBScan Clustering Method

7. (15 points) DbScan on the **non-noise data**. Try different settings of parameter Minpts and Eps. Plot the best clustering result you think (using different colors to show different clusters) and answer:

a) How the clustering result is changing when you increase Minpts ?

b) How the clustering result is changing when you increase Eps ?

8. (15 points) Use DbScan on the **noisy data**. Try different settings of parameter Minpts and Eps. Plot the best clustering result you think, and answer:

c) How the clustering result is changing when you increase Minpts ?

d) How the clustering result is changing when you increase Eps ?