CS 2756, Spring 2022
Homework 5
Due: April 21, 2022, 11:59 PM, Gradescope

Instructions:

- The main submission of HW 5 must be uploaded on Gradescope. Your submission will be graded via Gradescope platform.
- Your PDF file should be named as “lastname_firstname_hw5”. For example, if your name is Jane Doe, name your file as “doe_jane_hw5”.
- Please write your name (First Name, Last Name) legibly, as it appears on Canvas. Please also include your PITT email.
- Type out your solutions on a separate blank file (using Word, Latex, etc.), and don't forget to convert the file to PDF extension before submitting. You should not type out on the original pdf file.
- Only a subset of the questions will be graded. However, such questions are not determined a priori, therefore please do your best to answer all the questions correctly.

Question 1.

For each of the following situations, decide what type of clustering should be used (hierarchical or partitional; exclusive, overlapping, or fuzzy; and complete or partial) to obtain the desired grouping. If you believe that data could be clustered using several different types of clustering, please state your assumptions.

Example: A nutritionist asks you various questions to assess your risk for diabetes. Based on this data, the nutritionist hopes to group people in three different categories: low, medium and high.
Answer: Partitional, exclusive, complete

(a) You want to group all X-ray images taken in a hospital based on the **likelihood** of the patient having COVID-19. You will group into 4 clusters, based on severity level (e.g., no-covid, mild, moderate, severe).

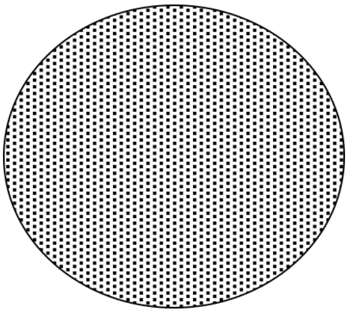
(b) You want to group pixels in an image of a crowd, based on whether the pixel indicates if the person object associated with that pixel is wearing a mask or not.

(c) Proteins perform different biological functions which are organized into a hierarchical taxonomy (GO) defined by biologists. Some proteins can be multi-functional as well. You want to group them based on those functions. Some proteins may also have missing functional annotation.

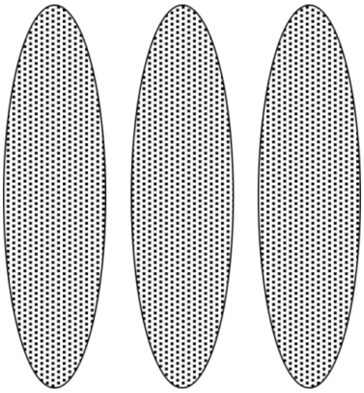
Question 2.

For the following sets of two-dimensional points, (1) draw a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that the minimization of the sum of squares error (SSE) is being considered as the optimization criterion. If you think that there can be multiple solutions for clustering depending on the initial choice of centroids, then please indicate whether each solution is a global or local minima. Assume that the shaded regions in each of the figures have uniform density of points.

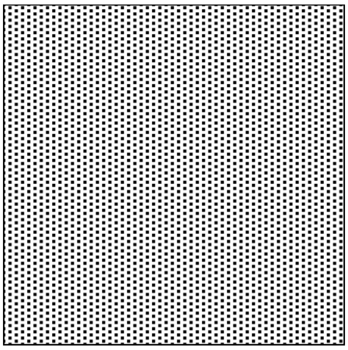
a) Use K = 4 on the dataset described in the figure below. How many possible ways are there (in theory) to partition this dataset into 4 clusters using the K-means algorithm?



b) Use K = 2 on the dataset described in the figure below.



c) Use K = 4 on the dataset described in the figure below.



Question 3.

Consider a set of 5 points in two-dimensional space, shown in the following table:

Point ID	X	Y
1	9	8
2	6	8
3	6	4
4	10	6
5	3	1

Assuming Euclidean distance as the distance measure, answer the following questions:

a) Compute the matrix of pair-wise distances between the 5 points, where the (i, j)th entry in the matrix corresponds to the distance between point i and point j.

b) Use the single link (MIN) hierarchical clustering technique for clustering these 5 points and show the dendrogram of the clustering. If you wish, you can draw this part by hand, but please note that only readable answers can be graded.

c) Use the complete link (MAX) hierarchical clustering technique for clustering these 5 points, and show the dendrogram of the clustering. If you wish, you can draw this part by hand, but please note that only readable answers can be graded.

Question 4.

For the figures below, each dataset contains four clusters, and each cluster has 80 points. In the distance matrices, points are sorted according to the four cluster labels. Different darkness indicates differences in distance: black indicates the lowest distances and white indicates the highest distances.

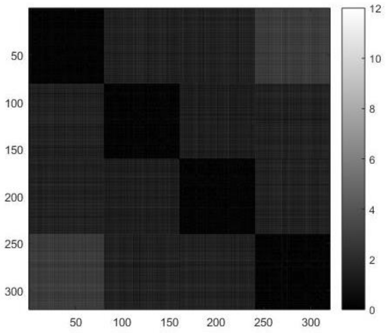


Fig. (a)

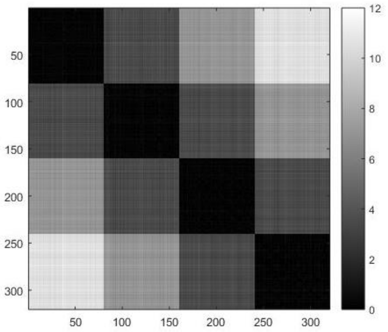


Fig. (b)

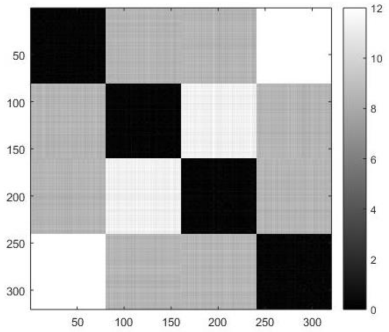


Fig. (c)

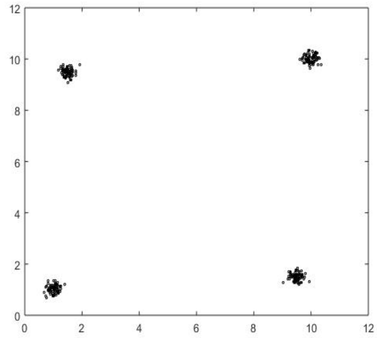


Fig. (d)

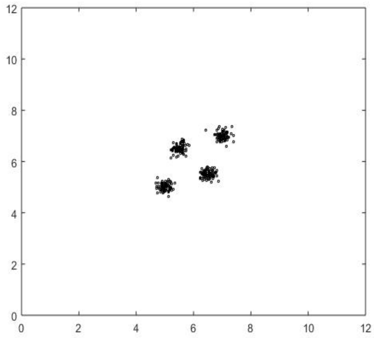


Fig. (e)

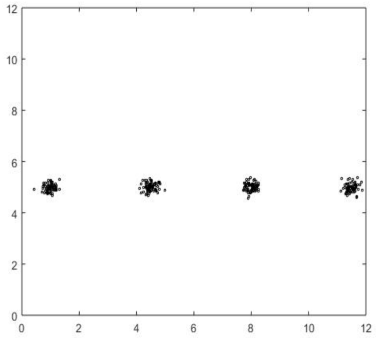


Fig. (f)

Match the distance matrices (Fig. (a), (b) and (c)) with their corresponding datasets (Fig. (d), (e) and (f)). Include a brief explanation.

(i) Fig (a)
Corresponding dataset (((d), (e) or (f)):
Explanation:

(ii) Fig. (b)
Corresponding dataset (((d), (e) or (f)):
Explanation:

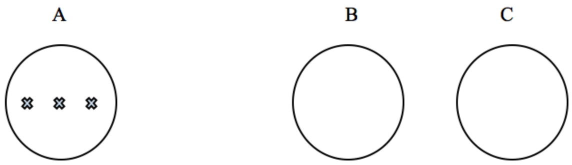
(iii) Fig. (c)
Corresponding dataset (((d), (e) or (f)):
Explanation:

Question 5.

In the three sets of figures below, assume that the leftmost circle (A) has 50,000 points, and the other two circles (B and C) have 50 points each. The X's are the centroid initializations for each run of K-means clustering. Assume a uniform distribution of points within each circle.

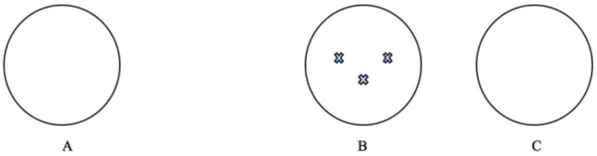
For each figure, you should tell how many centroids should end up in each circle after convergence of K-means clustering. Your answer should be 0, 1, 2, or 3. You should provide a brief justification for each case.

(a)



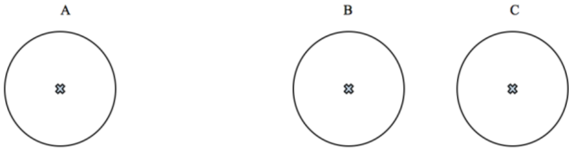
Number of Centroids in Circle (a):
Number of Centroids in Circle (b):
Number of Centroids in Circle (c):
Brief explanation:

(b)



Number of Centroids in Circle (a):
Number of Centroids in Circle (b):
Number of Centroids in Circle (c):
Brief explanation:

(c)

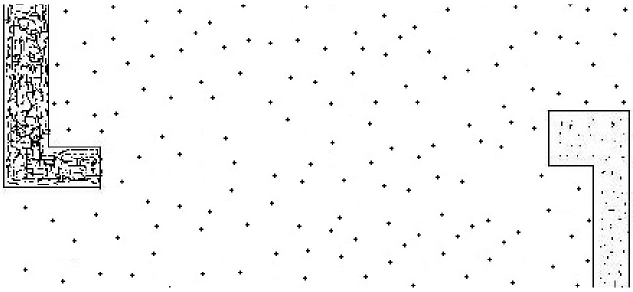


Number of Centroids in Circle (a):
Number of Centroids in Circle (b):
Number of Centroids in Circle (c):
Brief explanation:

Question 6.

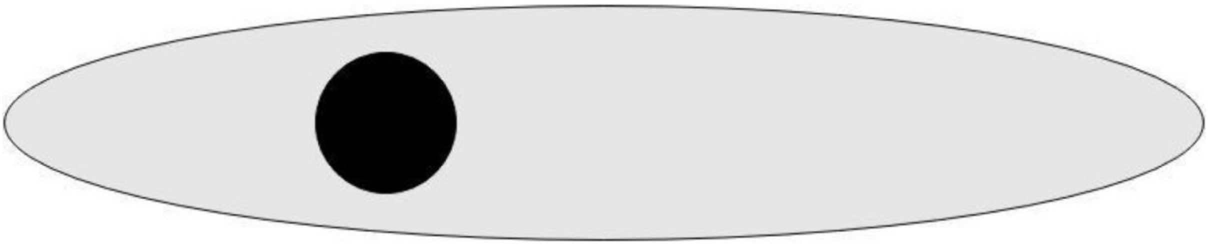
What challenges are faced by complete-link and DBSCAN respectively for the following case? Between the two algorithms, which one will you prefer for this specific data set? And why?

The upper block is denser than the lower block. And assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points.



Question 7.

For the following set of two-dimensional points, sketch (or if you think you can illustrate your answer by words, go for it) how they would be split into two clusters by K-means (when a global minimum of SSE is achieved) and by Gaussian mixture model clustering. Draw your answers on the figures below. You can assume the density of points in the darker area is much higher than the density of points in the lighter area.



b) Name one other clustering method that might be able to accurately capture the two clusters.

Question 8.

a) List one similarity and one difference between SOM and k-means.

b) List one similarity and one difference between bisecting k-means and hierarchical clustering using group average.

c) List one advantage and one disadvantage of SNN similarity over direct similarity.