

**Question 1 (Types of Clustering) [12 points]**

Each of the following parts describes a collection of groups, or groupings. Describe each of these groupings in terms of the characteristics we discussed for sets of clusters. More specifically, label every grouping as to whether they are:

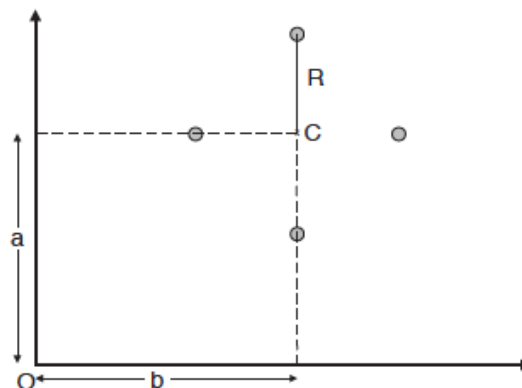
- hierarchical or partitional
- overlapping or non-overlapping
- complete or partial

Each grouping should be labeled with three characteristics, e.g., partitional, overlapping, and incomplete. If you feel there may be some ambiguity about what characteristics a grouping has, provide a short justification of your answer.

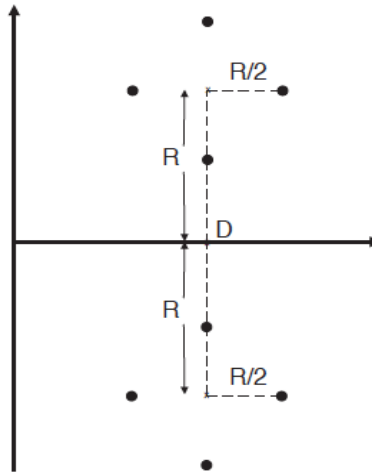
- a) Grouping of movie actors based on the genres of the movies (comedy, drama, sci-fi etc.) they have acted in.
- b) Grouping of students in a university based on the organization (department, college, institute etc.) they belong to. A student may belong to multiple organizations. Also, some students don't have declared majors and hence may not belong to any organization.
- c) Grouping of all the students in the Computer Science department based on the letter grade they get in Machine Learning Class.
- d) You want to group all locations on Earth as to whether they belong to a tropical rainforest, a deciduous forest, or an evergreen forest. Here, each location corresponds to a region of surface area 1km X 1km, and a location can have more than one variety of forests.

**Question 2 (Computing SSE) [2 + 2 + 4 = 8 points]**

Consider the four data points shown in Figure below. The distance between each point to the center C is R.



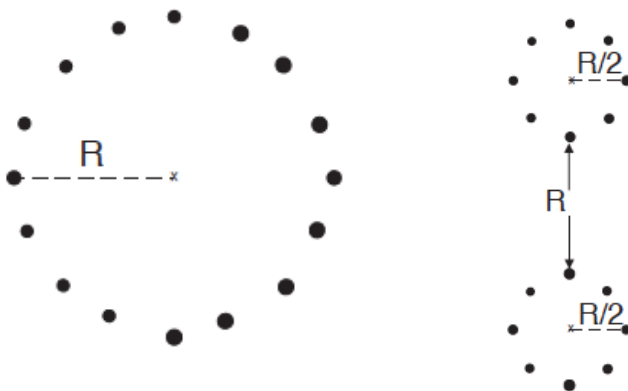
- i. Compute the total SSE of the data points from the centroid, C.
- ii. Compute the total SSE of the data points to the origin, O.
- iii. Use your approach in (ii) to compute the SSE for the 8 data points shown below with respect to the centroid, D.



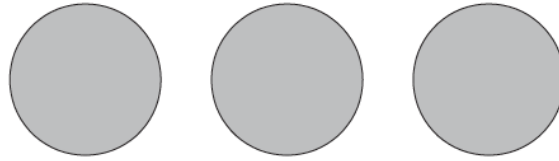
**Question 3 (k-means and Bisecting k-means) [6 points]**

For each of the following data sets, state which method will perform better, k-means or bisecting k-means (where  $k = 3$ ).

a)

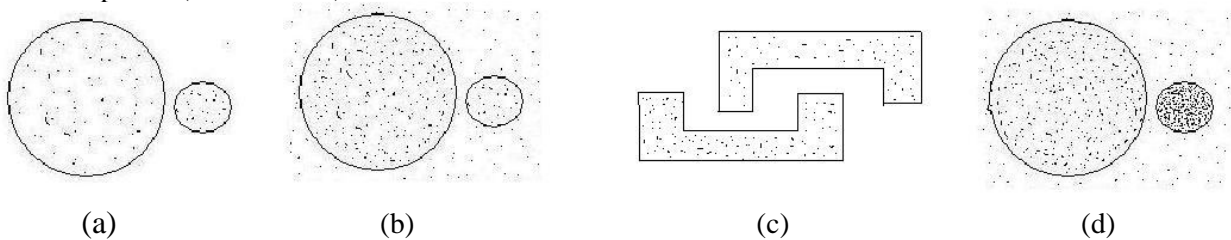


b)



**Question 4 (Hierarchical Clustering and DBSCAN) [12 points]**

How will single-link, complete-link, and DBSCAN clustering algorithms perform in the following cases? The points are evenly distributed for the first three cases (a-c), while the last case (d) has one dense cluster with 50000 points and one relatively sparse cluster with 50 points with noise data points in between. Assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points (case b and d).



**Question 5 (Hierarchical Clustering) [5 + 4 + 4 = 13 points]**

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:

- Compute the matrix of pair-wise distances between the 5 points, where the  $(i, j)$ <sup>th</sup> entry in the matrix corresponds to the distance between point  $i$  and point  $j$ .
- Use the single link (MIN) hierarchical clustering technique for clustering these 5 points, and show the dendrogram of the clustering. The Y-axis of the dendrogram (height at which two clusters are merged) can be chosen to be the pair-wise distance between the two clusters.

- c) Use the complete link (MAX) hierarchical clustering technique for clustering these 5 points, and show the dendrogram of the clustering. The Y-axis of the dendrogram (height at which two clusters are merged) can be chosen to be the pair-wise distance between the two clusters.

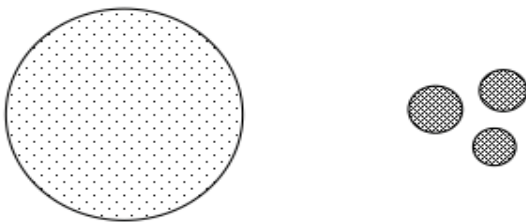
### Question 6 (Hierarchical Clustering) [4 points]

Given the four clusters shown in the figure, answer the following questions. Assume that:

- Data points within the big cluster (on the left) are uniformly distributed
- Between cluster distances of small clusters (on the right) are larger than the distance between points in the large cluster
- The diameter of the larger cluster is larger than the maximum distance of points in the smaller clusters.

a) Will MAX agglomerative clustering identify the four clusters? Explain briefly.

a) Will MIN agglomerative clustering identify the four clusters? Explain briefly.



### Practice Questions

#### Question 7 (Types of Clustering)

Each of the following parts describes a collection of groups, or groupings. Describe each of these groupings in terms of the characteristics we discussed for sets of clusters. More specifically, label every grouping as to whether they are:

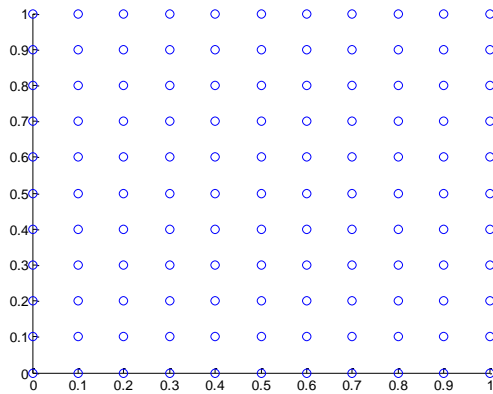
- hierarchical or partitional
- overlapping or non-overlapping
- complete or partial

a) The groups are all possible sets of 10 items drawn from a set of 100 items.

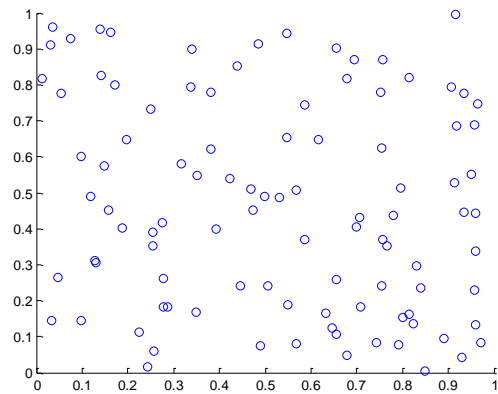
b) Each group is the set of all people in the military who have the same rank.

#### Question 8 (k-means and DBSCAN)

Suppose you are given two sets of 100 points that fall within the unit square. One set of points (a) is arranged so that the points are **uniformly spaced**. The other set of points (b) is **randomly generated** from a uniform distribution over the unit square.



**a.**

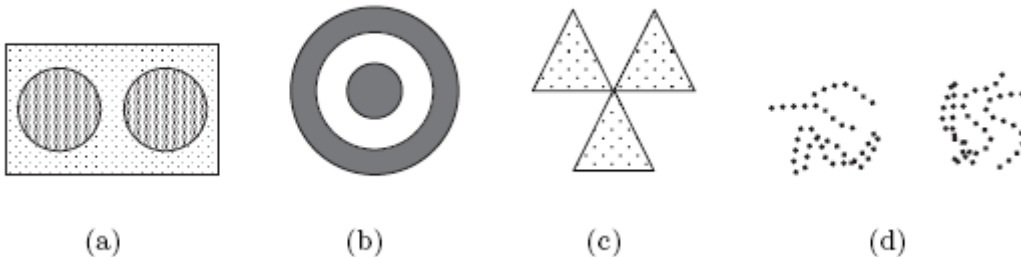


**b.**

- (i) How does the behavior of DBSCAN differ on (a) and (b)?
- (ii) How does the behavior of DBSCAN and K-means differ on (b)?

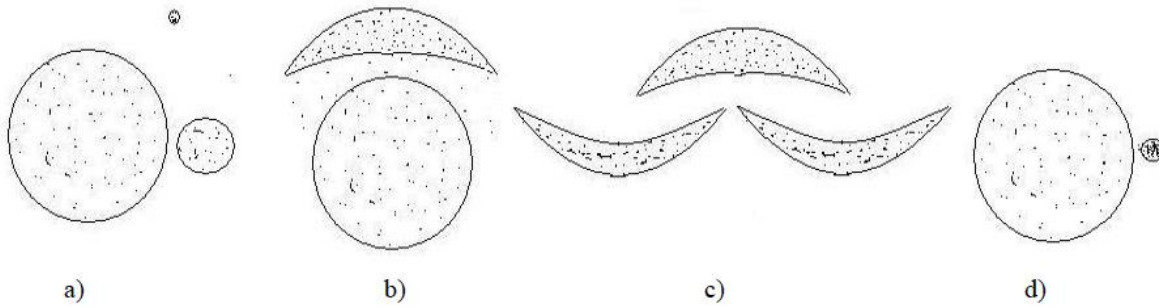
### Question 9 (Comparing clustering methods)

Identify the clusters in the following figure using the center-, contiguity-, and density-based definitions. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN. Also indicate the number of clusters for each case and give a brief explanation of your reasoning. Note that darkness or the number of dots indicates density.




### Question 10 (Hierarchical Clustering)

How will single-link and complete-link will perform for following cases? The points are evenly distributed for first three cases(a-c), where the last case(d) has one dense cluster with 50000 points and one relatively sparse cluster with 50 points only without any noisy data points in between. Assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points (case b).



### Question 11 (k-means)

To answer the following true / false questions on how k-means operates, refer to figures (a), (b), and (c), below. Note that we are referring to the very basic k-means algorithm presented in class and not to any of its more sophisticated variants, such as bisecting k-means or k-means++.

Note that for all three figures, the initial centroids are given by the symbol:  Initial point

For figures (a) and (b) assume the shaded areas represent points with the same uniform density. For Figure (c), the data points are given as red dots and their values are indicated under the dots. No explanation for your answer is necessary unless you feel there is some ambiguity in the figure or the question.

- True or False:** For Figure (a) and the given initial centroids, when the k-means algorithm completes, each shaded circle will have one cluster centroid at its center.
- True or False:** For Figure (b) and the given initial centroids, when the k-means algorithm completes, there will be one cluster centroid in the center of each of the two shaded regions and each of the two final clusters will consist only of points from one of the shaded regions. In other words, none of the two final clusters will have points from both shaded regions.
- True or False:** For Figure (c) and the given initial centroids, the final clustering for k-means contains an empty cluster.

