

Homework 5

Kelsey Neis - neis@umn.edu - 3942198

11/15/2021

Question 1

For each of the described data sets, decide what type of clustering should be used (hierarchical or partitional, exclusive or non-exclusive (overlapping/fuzzy), complete or partial (incomplete)). Briefly explain your reasoning if you feel there may be several possible answers. Note: we are using partitional and hierarchical in the more relaxed use of the terms to mean un-nested or nested, respectively.

a) Grouping of cities based on the various locations, namely county (local region), state or province, and country.

hierarchical, exclusive, complete

b) A supermarket manager wants to group all goods into several categories, each of which has multiple subcategories.

hierarchical, non-exclusive, complete

c) 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 grid cell of surface area 1km by 1km, and a location can have more than one variety of forest or may be unforested.

partitional, non-exclusive (fuzzy), partial. Partial because there will be many grid cells that do not have any forest and will not be interesting to my clustering experiment.

Question 2

To answer the following true/false questions about 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:

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.

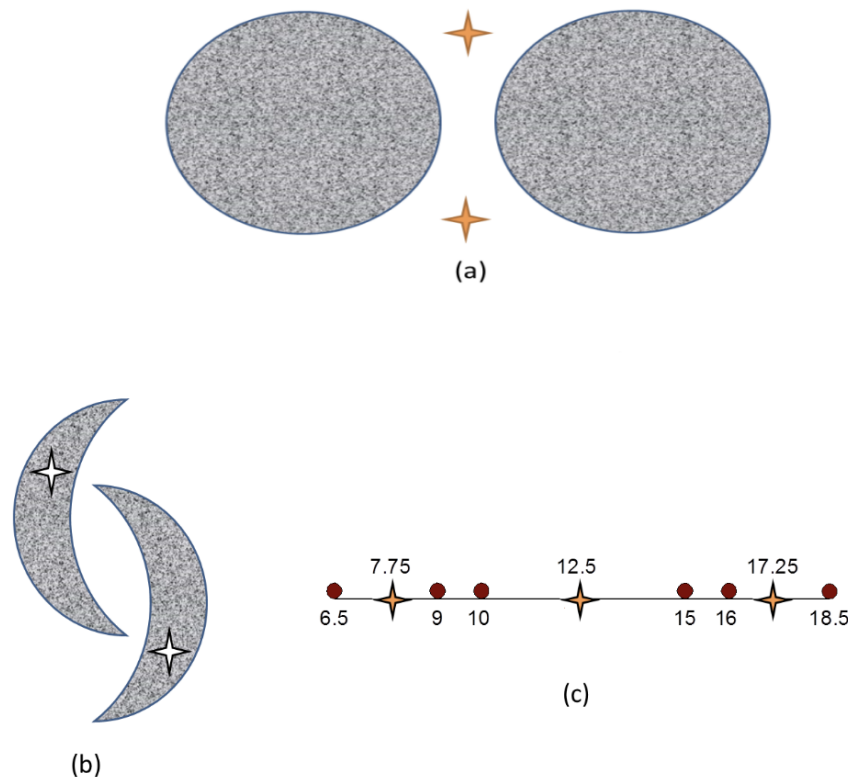


Figure 1: q2

a) True or False: For Figure (a) and the given initial centroid: When the k-means algorithm completes, each shaded circle will have one cluster centroid at its center.

False

b) 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.

False

c) True or False: For Figure (c) and the given initial centroids, the final clustering for k-means contains an empty cluster.

True

Question 3

In each of the three sets of figures below, assume that circles A and C contain 100,000 points each, and circle B contains 100 points. The X's are the centroid initializations for each run of K-means clustering. Assume a uniform distribution of points within each circle. Each circle is the same size, and the distances between the circles are to scale. 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)

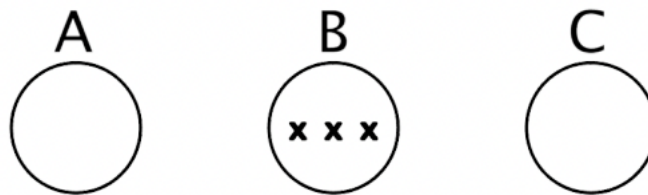


Figure 2: a

Number of Centroids in Circle (a): 1

Number of Centroids in Circle (b): 1

Number of Centroids in Circle (c): 1

Brief explanation: The centroids on the left and right will move to A and C respectively after the points from those areas are assigned to them, and the x in the center will remain there, since there are equal number of points on either side.

b)

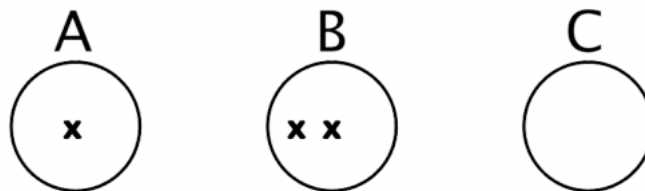


Figure 3: b

Number of Centroids in Circle (a): 1

Number of Centroids in Circle (b): 1

Number of Centroids in Circle (c): 1

Brief explanation: The centroid in A will stay where it is, the right centroid will take the points from C and move over there, and the remaining centroid will end up in the middle since it will not get any points from A or C.

c)

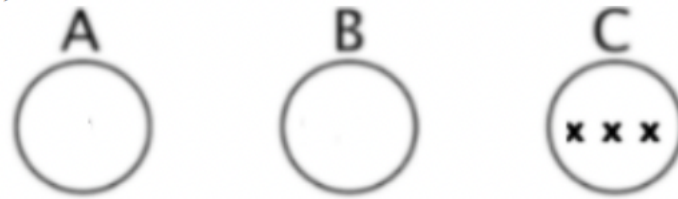


Figure 4: c

Number of Centroids in Circle (a): 1

Number of Centroids in Circle (b): 0

Number of Centroids in Circle (c): 2

Brief explanation: The left centroid will get the points from A and B, and then move until it's in A. The other centroids will stay put, taking each a portion of the points in C.

Question 4

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

Figure 5: q4-1

Assuming Euclidean distance as the distance measure. The pairwise distance matrix between the 5 points has been computed in the below figure for you, where the ij th entry in the matrix corresponds to the distance between point i and point j .

Pairwise Distances	Point 1	Point 2	Point 3	Point 4	Point 5
Point 1	0	3	5	2.23	9.2
Point 2	3	0	4	4.47	7.6
Point 3	5	4	0	4.47	4.2
Point 4	2.23	4.47	4.47	0	8.6
Point 5	9.2	7.6	4.2	8.6	0

For this question you are allowed to paste an image of your handwritten/plotted dendrogram as part of the solution.

a) Use the single link (MIN) hierarchical clustering technique for clustering these 5 points and show the dendrogram of the clustering. The dendrogram should clearly show the order in which the points are merged.

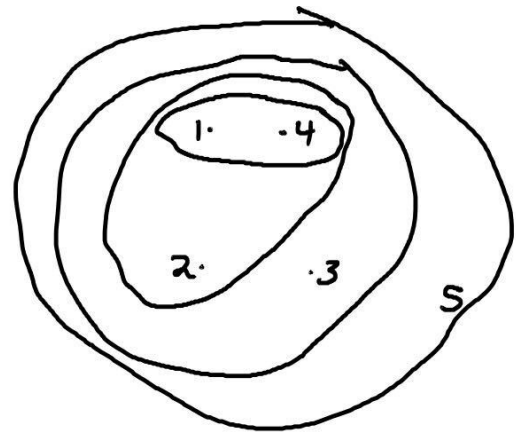
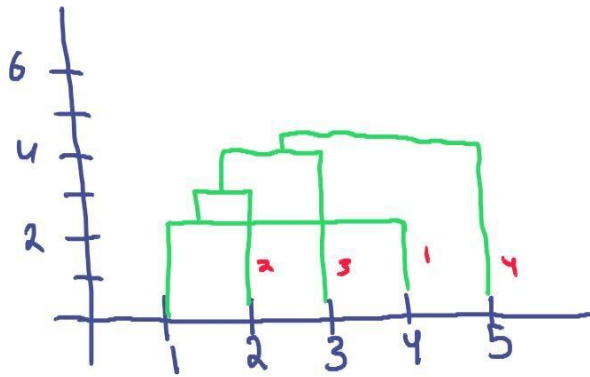


Figure 6: MIN clustering

b) Use the complete link (MAX) hierarchical clustering technique for clustering these 5 points and show the dendrogram of the clustering. The dendrogram should clearly show the order in which the points are merged.

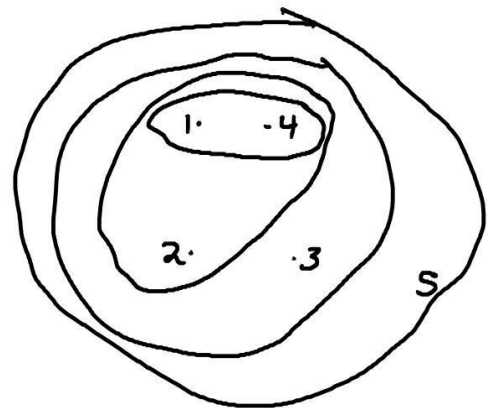
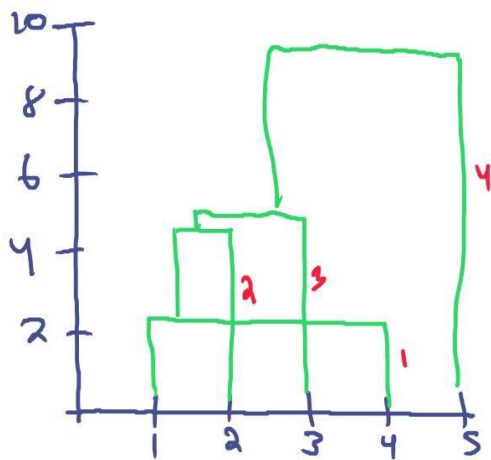


Figure 7: MAX clustering

Question 5

Suppose we apply DBSCAN to cluster the following dataset using Euclidean distance.

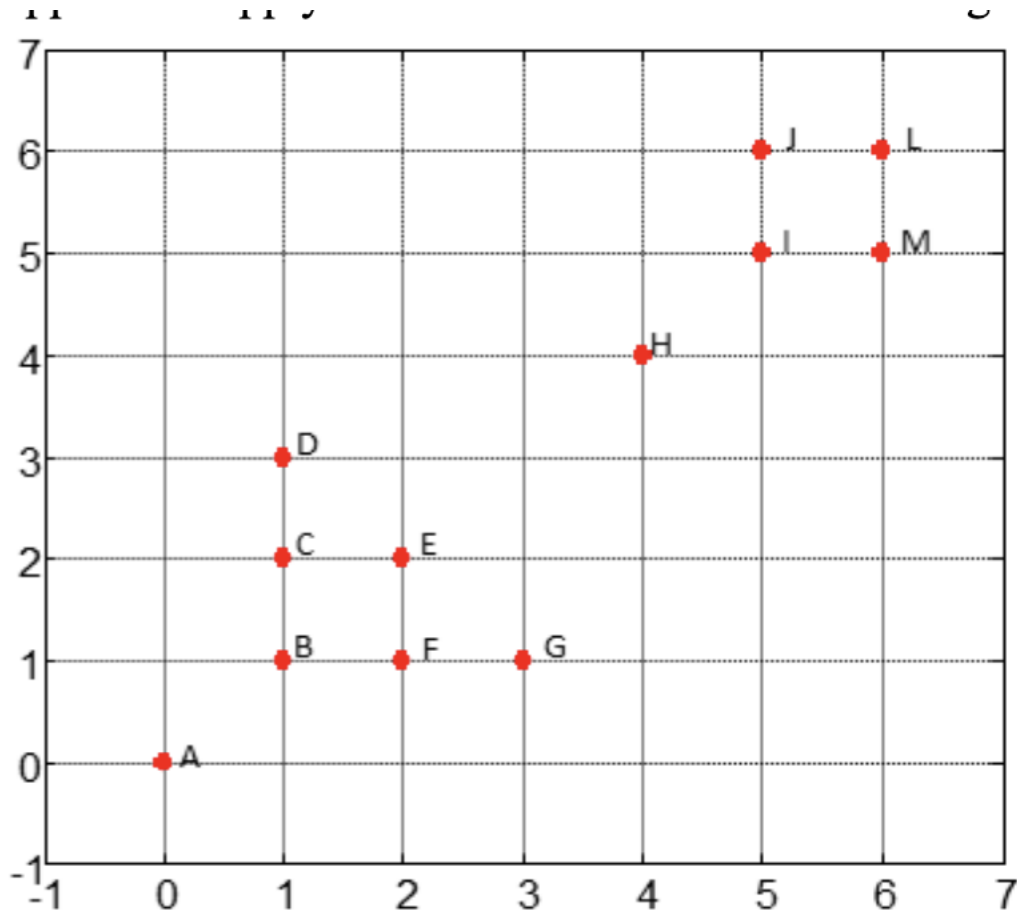


Figure 8: q5

A point is a core point if its density (num point within EPS) is $\geq \text{MinPts}$. Given that $\text{MinPts} = 3$ and $\text{EPS} = 1$, answer the following questions.

a) Label all points as “core points,” “boundary points,” and “noise.”

A: noise, B: core, C: core, D: boundary, E: core, F: core, G: boundary, H: noise, I: core, J: core, L: core, M: core

b) What is the clustering result?

One cluster with B, C, D, E, F, G (B, C, E, F are connected points) and another cluster with I, J, L, M and A and H are eliminated

c) Repeat the above two questions when $\text{EPS} = \sqrt{2}$

A: boundary, B: core, C: core, D: core, E: core, F: core, G: core, H: boundary, I: core, J: core, L: core, M: core

Result: One cluster with A, B, C, D, E, F, G and another cluster with H, I, J, L, M