Homework Problem Set 3: Basic Clustering

Due Thursday, Jan. 26 at 11:59 PM
Upload a pdf to Canvas

Each question is worth the same number of points.

Question 1:

1. For the following questions, give an answer and a short (1 or 2 sentences) explanation. For the rest of this question, agglomerative hierarchical clustering refers to procedures such as single link, complete link, and group average, while k-means clustering refers to k-means with random initialization of centroids and Euclidean distance. a) Agglomerative hierarchical clustering procedures are better able to handle outliers than k-means. t

b) For any given data set, different runs of k-means can produce different clusterings, but agglomerative hierarchical clustering procedures will always produce the same clustering. t

c) K-means take less time and memory than agglomerative hierarchical clustering and is the most efficient clustering algorithm possible. f

d) During a post-processing step for K-means, a cluster is split by picking one of the points of the cluster as a new centroid and then reassigning the points in the cluster either to the original centroid or the new centroid. What happens to the SSE of the clustering? decrease sse

e) When clustering a dataset using K-means, whenever SSE decreases, cohesion increases. t

f) When clustering a dataset using K-means, whenever SSB (the between sum of squares) increases, separation increases. t

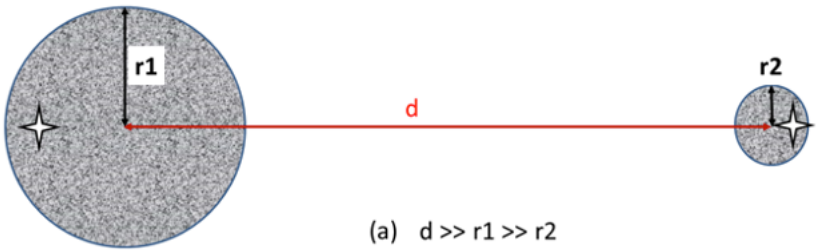
g) Cohesion and separation are independent for K-Means, i.e., improving cohesion (smaller SSE) doesnt necessarily improve separation (larger between sum of squares (SSB)). t

h) When clustering a dataset using K-means, SSE + BSS is a constant. f

i) When clustering a dataset using K-means, whenever cohesion increases, separation increases. f

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 it 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.



(a) d >> r1 >> r2

(b)

(c)

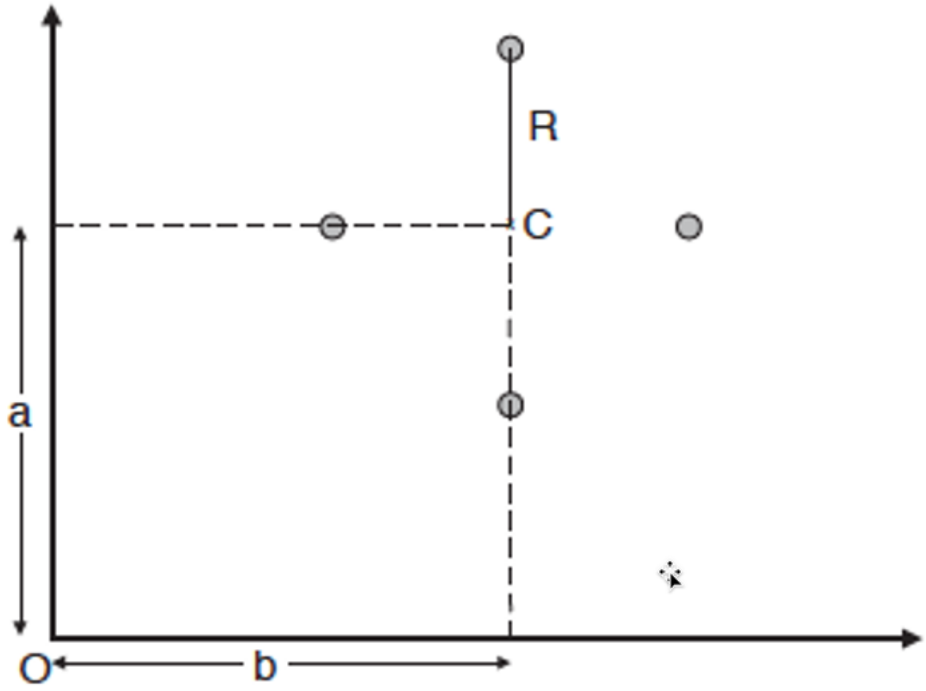
a) True or False: For Figure (a) and the given initial centroid: When the k-means algorithm completes, each shaded circle will have one cluster centroid at its center.

b) True or False: For Figure (b) and the given initial centroids: When the k-means algorithms 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.

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

Question 3:

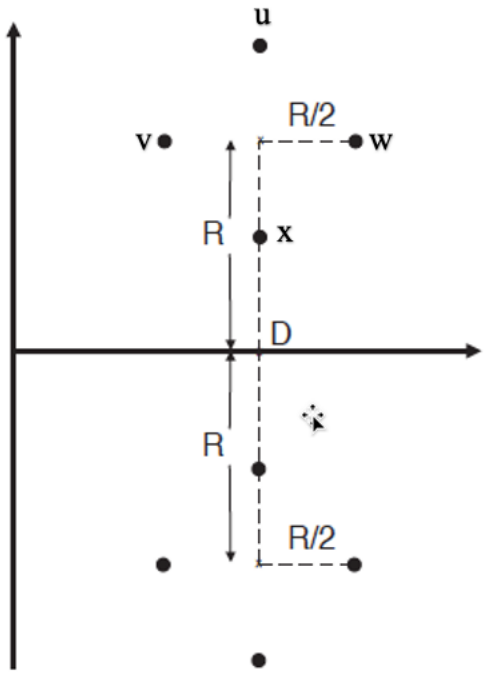
Consider the four data points shown in the following Figure. The distance between each data point to the center C is R.



a) Compute the total SSE of the data points to the centroid, C. 4*R^2

b) Compute the total SSE of the data points to the origin, O. 4*(a^2+b^2+R^2)

c) Using parts (a) and (b), compute the SSE for the 8 data points shown below with respect to the centroid, D. Note that points u, v, w, and x lie on a circle of radius $R/2$. Also, the figure is symmetric with respect to the horizontal line running through D. 10*R^2

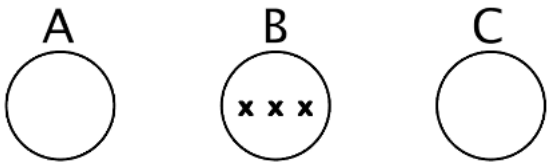


Example of 8 data points in 2-dimensional space.

Question 4:

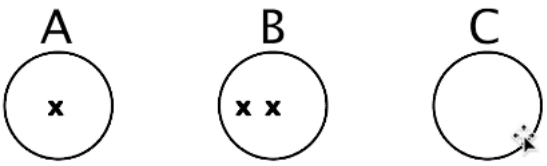
In each of the three sets of figures below, assume that circles A and B contain 100 points each, and circle C contains 100,000 points. The Xs are the centroid initializations for each run of K-means clustering. Assume a uniform distribution of points within each circle. Each circle is the same size,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



The final distribution after convergence should be 0 centroids in circle A, 1 centroid in circle B, and 2 centroids in circle C.
doubt if 111.

Figure 3 (a)



the final expected distribution after K-means clustering converges should be 1 centroid in circle A, 1 centroid in circle B, and 1 centroid in circle C.

Figure 3 (b)

2 in C and 1 between A and B

Figure 3 (c)

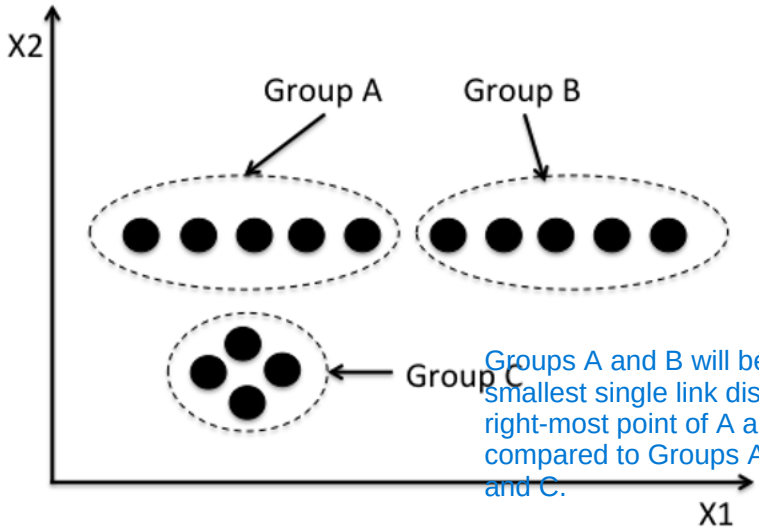
a) The distance between circles A and B is the same as the distance between B and C. (Figure 3 (a))

b) The distance between circles A and B is the same as the distance between B and C. (Figure 3 (b))

c) Circles A and B are much closer than B and C. (Figure 3 (c))

Question 5:

At an intermediate stage of some agglomerative clustering algorithm, you are given three groups of points, as shown in the figure below, which need to be considered for merging. Note that every circle in the figure represents a two-dimensional point, and the Euclidean distance in two dimensions is being used as the distance measure.



Groups A and B will be merged since they have the smallest single link distance (between right-most point of A and left-most point of B), as compared to Groups A and C, and Groups B and C.

Figure 4

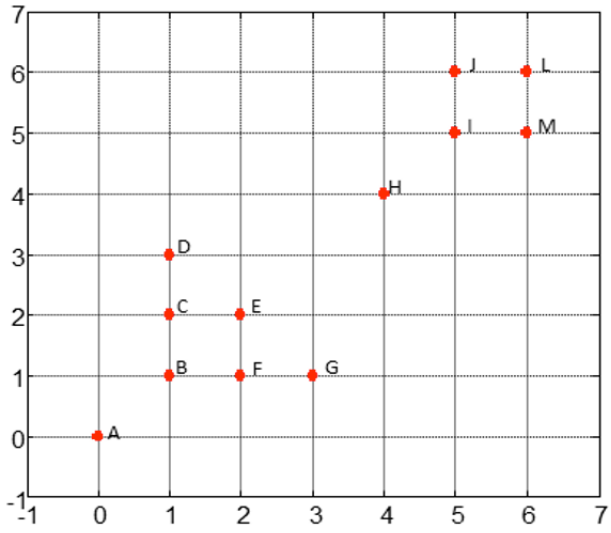
a) Using the single link (MIN) hierarchical clustering technique, which pair of groups would you consider for merging? Provide a one-sentence justification.

: Groups A and C will be merged since they have the smallest complete link distance (between a boundary point of A and the farthest point in C), as compared to the complete link distance of Groups A and B (between the left-most point in A and right-most point in B), and Groups B and C (between right-most-point in B and the farthest point in C).

b) Using the complete link (MAX) hierarchical clustering technique, which pair of groups would you consider for merging? Provide a one-sentence justification.

Question 6:

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



B) There will be two clusters: B, C, D, E, F, G will form a cluster and I, J, L, M will form the other cluster

C) Core points: B, C, E, F, I, J, L, M, D, G, A, H and all points are clustered as a single cluster.

Figure 5

A point is a core point if its density (number of points within ϵ) is $\geq$ MinPts. Given that MinPts = 3 and EPS = 1, answer the following questions.

A) Core points: B, C, E, F, I, J, L, M
Border points: D, G.
Noise: A, H

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

b) What is the clustering result?

c) Repeat the above two questions when $\epsilon = \sqrt{2}$.

Question 7:

The following table (confusion matrix) shows the k-means clustering results for a land cover classification dataset that consists of many pieces of land. The number provided in the table is the number of objects (pieces of land) that are clustered into each cluster that belongs to each category. For example, the number in the forest column and cluster 1 row means that 10 forest items are clustered into cluster 1. Answer the following questions based on the given table. No calculations are necessary. Briefly explain your answer.

	Forest	Farm	Shrubland	Urban	Water
Cluster 1	10	100	20	10	30000
Cluster 2	3000	10	1000	10	0
Cluster 3	10	3000	500	150	200
Cluster 4	2000	2500	1500	3000	1400

0.0486847908152747916
0.8574416759395337
1.090176370248973
2.2615165224362351

Table: k-means clustering results for land cover classification dataset

a) Which cluster has the largest clustering entropy? cluster 4

b) Which cluster has the smallest clustering entropy? cluster 1

Question 8:
The figures below are sorted according to cluster labels, and corresponds to the sets of points (Dataset X, Dataset Y, and Dataset Z). Differences in color distinguish between clusters, and each set of points contains 100 points and four clusters each of equal size. In the distance matrix, blue indicates the lowest distances and red indicates the highest distances.

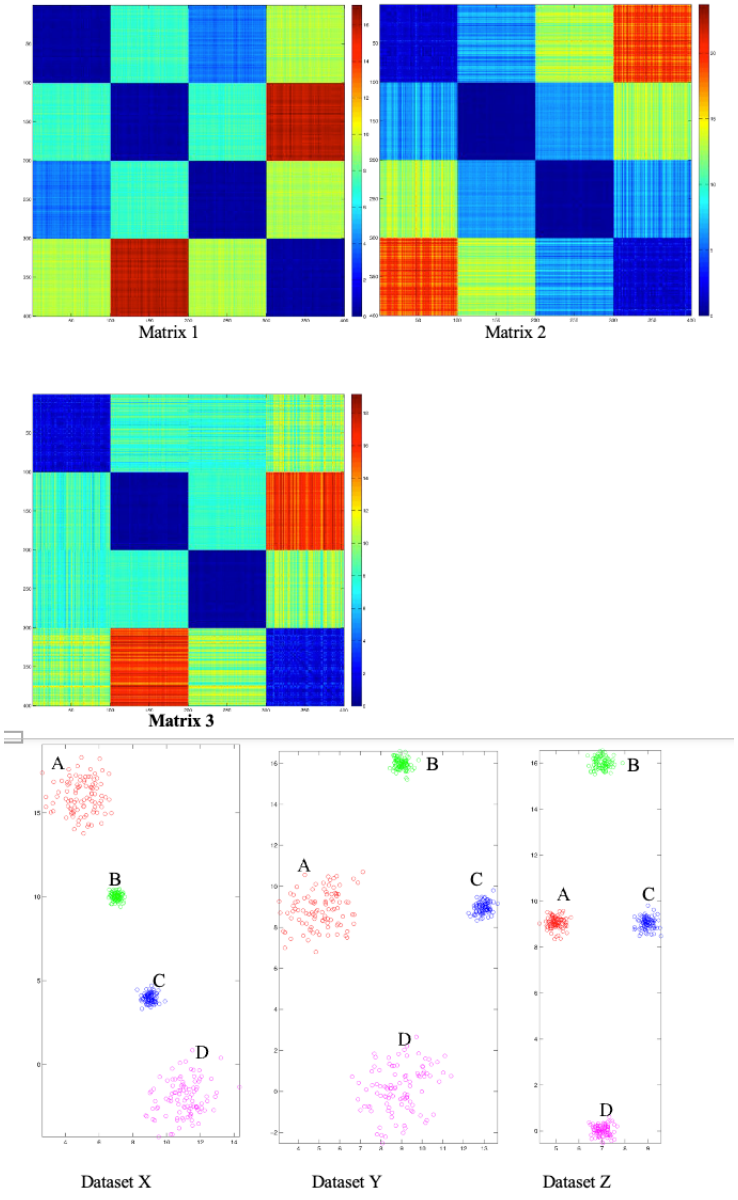


Figure 5

(a) Match the distance matrices (Matrix 1, Matrix 2, Matrix 3) with the sets of points (Dataset X, Dataset Y, and Dataset Z). Provide a brief explanation of the diagaonal entries and the non-diagonal entries.

(b) For the symmetric matrix given in Matrix 2 match the four rows to the corresponding clusters (characterized the nearest alphabet in each cluster (e.g: A, B, C, D)) in the dataset that you match with it in the previous question. Provide a brief explanation

---

Question 9:
For each of the described data sets, decide what type of clustering should be used (hierarchical or partitional, exclusive or overlapping or 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.

An example: Clustering library books based on their literary genre. The genre/topic can have several subtopics, as well. Answer: hierarchical, overlapping, complete

a) Proteins perform different biological functions that 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 be missing functional annotation.

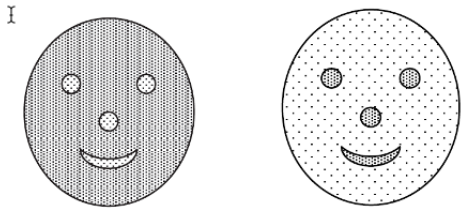
b) A nutritionist asks you several questions (e.g., your calorie intake, types of food you eat, your physical activity labels, and so on) to assess your risks for diabetes in three different groups: low, medium, and high.

c) An international grad student can work on campus only at most for 20 hours. You want to assign each student to different job categories (e.g., TA, RA, another on-campus job, jobless). Hint: the sum of these categories should sum up to 20 hours.

d) Grouping of students in a university-based on the organization (department, college, institute, etc.) to which they belong. A student may belong to multiple organizations. Also, some students dont have declared majors and hence may not belong to any organization.

e) Grouping of all the students in the Computer Science department based on the letter grade they get in the data mining (CSci 5523) class.

Question 10:
Consider the following two sets of points (faces) shown in figures (a) and (b). The darker shading indicates a denser point distribution.



(a) (b)

a) For each figure, could you use DBSCAN to find clusters corresponding to the patterns represented by the nose, eyes, and mouth? Explain.

b) For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain.

a) For (a), could you figure out a clustering method, which can find the patterns represented by the nose, eyes, and mouth?