1.

We have a dataset whose points have one features and are in two classes: $\{-4, 2, 3\}$ are in the negative class and $\{-3, -2, 1\}$ in the positive class.

(a) Disregard the labels and cluster the data in two groups using K-means. Label all data in each cluster using the label of the center of the cluster. What is the misclassification error rate (%) of the classifier you built?

2.

We have a dataset with 10 points, each having two features:

Point	**X**	**Y**
p0	-4	-8
p1	7	-1
p2	2	9
p3	**-6**	**6**
p4	4	1
p5	-2	11
p6	-5	-4
p7	**0**	**-3**
p8	-9	2
p9	4	3

(a) Assume that we want to perform **k-medoids** clustering with **k=2**. Let us initialize the algorithm with two medoids, points p3 and p7. What are the final two clusters after one iteration of k-medoid clustering? Use Manhattan distance (NOT square Euclidean distance) to calculate the distance (dissimilarity) between medoid and non-medoid points.

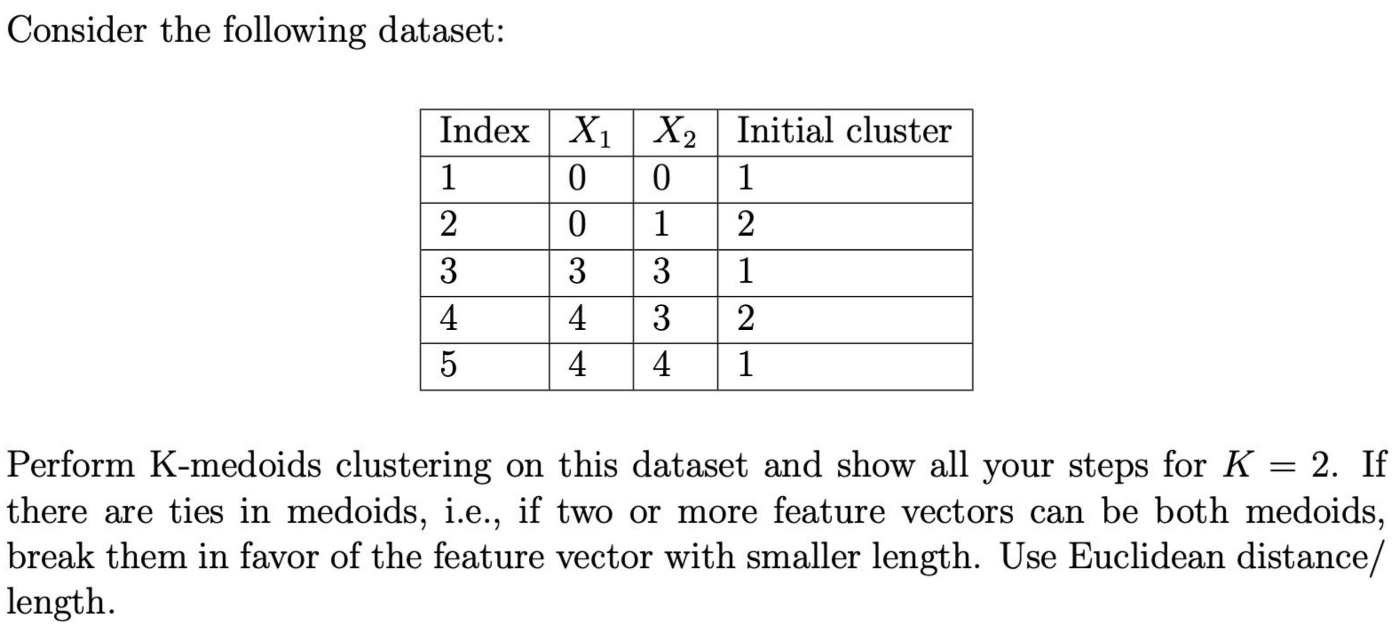
(b) What is the total cost (total within cluster variation) for assigning the points to the two clusters? Use Manhattan distance for within cluster variation (NOT Euclidean distance).

[Hint: Manhattan Distance between two points (x_1, x_2) and (x_2, y_2) is $d = |x_1 - x_2| + |y_1 - y_2|$]

(c) Let us initialize **p3** and **p9** as our initial medoids to perform **k-medoids** clustering with **k=2**. Will this change the final two clusters after one iteration of k-medoid clustering, compared to the two clusters of part a?

(d) Which medoids and clusters among parts a and b are better? Explain why.

3.



Consider the following dataset:

Index	X_1	X_2	Initial cluster
1	0	0	1
2	0	1	2
3	3	3	1
4	4	3	2
5	4	4	1

Perform K-medoids clustering on this dataset and show all your steps for $K = 2$. If there are ties in medoids, i.e., if two or more feature vectors can be both medoids, break them in favor of the feature vector with smaller length. Use Euclidean distance/ length.

4.

We have a dataset whose points have two features: $\{\mathbf{x}_1 = (0,0), \mathbf{x}_2 = (0,1), \mathbf{x}_3 = (2,0), \mathbf{x}_4 = (2,2)\}$.

(a) Using the K-means algorithm, cluster the data into two groups. Initially, assign $\mathbf{x}_1$ and $\mathbf{x}_4$ to cluster one and $\mathbf{x}_2$ and $\mathbf{x}_3$ to cluster 2.

(b) Using Euclidean distance and single linkage, perform hierarchical clustering on the dataset and show the results by a dendrogram.

5.

The table below is a similarity matrix for 6 objects. We define similarity of two objects as $1-$distance of those two objects.

Index	A	B	C	D	E	F
A	1					
B	0.88	1				
C	0.49	0.75	1			
D	0.16	0.84	0.86	1		
E	0.72	0.23	0.30	0.55	1	
F	0.66	0.39	0.07	0.80	0.33	1

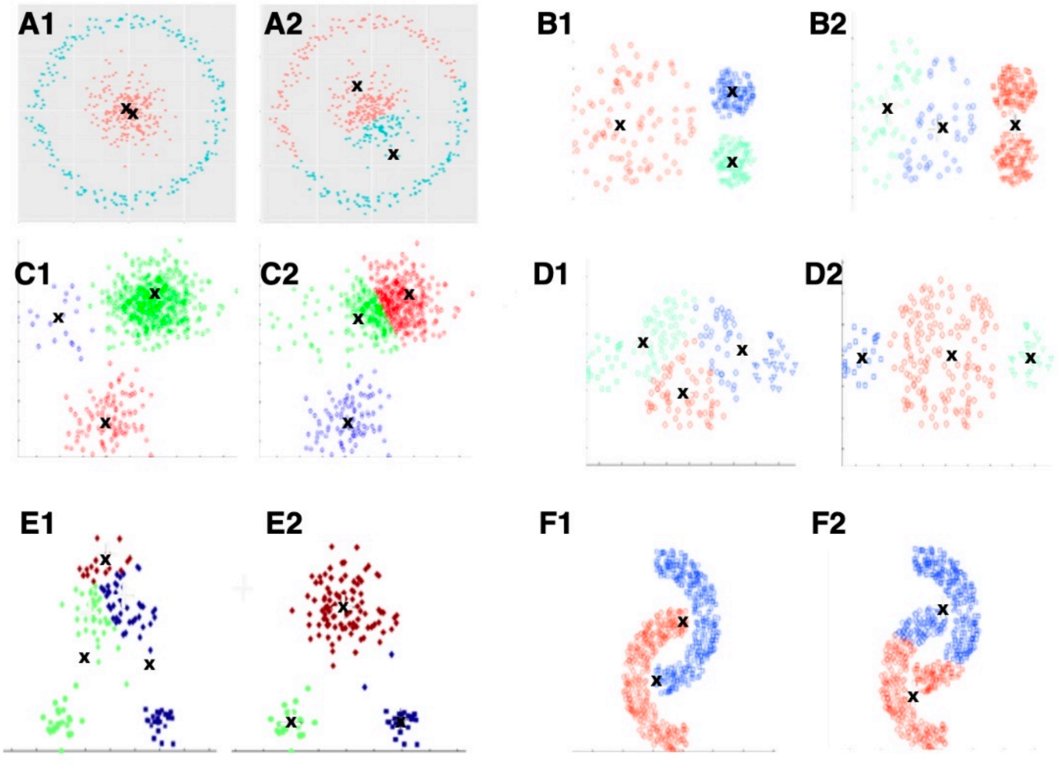
(a) Perform hierarchical clustering with single linkage (minimum of pairwise distances) on the data and draw the corresponding dendrogram. Show your work.

(b) Perform hierarchical clustering with complete linkage (maximum of pairwise distances) on the data and draw the corresponding dendrogram. Show your work.

(c) Change two values from the matrix so that the dendrograms in the last two question would be same.

6.

There are 6 different datasets noted as A,B,C,D,E,F. Each dataset is clustered using two different methods, and one of them is K-means. All results are shown in figure below. You are required to determine which result is more likely to be generated by K-means method. (Hint: check the state when K-means converges; Centers for each cluster have been noted as x; Since x and y axis are scaled proportionally, you can determine the distance to centers geometrically). The distance measure used here is the Euclidean distance.



(a) Dataset A (write A1 or A2)

(b) Dataset B (write B1 or B2)

(c) Dataset C (write C1 or C2)

(d) Dataset D (write D1 or D2)

(e) Dataset E (write E1 or E2)

(f) Dataset F (write F1 or F2)

7.

A research team is using hierarchical clustering with single linkage to study the relationship between different research papers based on their citation patterns. The goal is to group papers that cite similar sets of other papers, suggesting thematic or methodological similarities. [By Soumyaroop Nandi]

The team decides to use a simplified pairwise distance formula to quantify the dissimilarity between any two papers, (A) and (B), based on the number of unique citations they have in common ((C)) and the total number of unique citations between them ((T)):

Pairwise Distance = 1 - $\frac{C}{T}$

Given the following citation data for the first four papers:

Paper	**Citation**
A	1, 2, 3
B	2, 3 , 4
C	3, 4, 5
D	1, 4, 5

(a) Calculate the pairwise distance between Paper A and Paper B using the provided formula.

(b) Explain the significance of using single linkage in the context of hierarchical clustering for this study.

(c) After constructing the dendrogram based on the pairwise distances among all papers, the research team observes that Papers A and B are the first to be merged into a cluster. What does this indicate about their citation patterns, and how might this information be useful for the research team?

8.

Consider the following dataset:

Index	X_1	X_2	Z_1	Z_2
1	0	0		
2	1	1		
3	-1	-1		
4	-2	-2		
5	2	2		

Calculate the first and second principal components Z_1 and Z_2 for each datapoint by calculating the loading vector for each principal direction.

Hint: To solve this problem, you DO NOT need to solve any optimization problem, SVD, or eigenvalue problem. DO NOT forget the constraints we have for loading vectors.

9.

Consider the following unlabeled dataset:

$$\left\{\mathbf{x}_1 = \begin{bmatrix}1\\1\end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix}-1\\-1\end{bmatrix}\right\}, \left\{\mathbf{x}_3 = \begin{bmatrix}1\\-1\end{bmatrix}, \mathbf{x}_4 = \begin{bmatrix}-1\\1\end{bmatrix}\right\}.$$

(a) Determine the (loading) vector φ_1, $\|\varphi_1\|_2 = 1$ that represents the first principal direction of the data.

(b) Determine the (loading) vector φ_2, $\|\varphi_2\|_2 = 1$ that represents the second principal direction of the data.

Hint: you do not need to solve an optimization problem to answer this question.