**Question 1:**

Assume we have a simple dataset with 10 two-dimensional points (x, y).

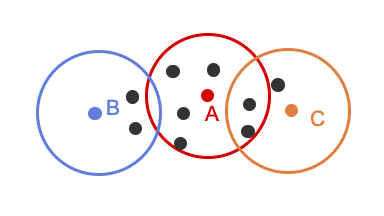
Dataset: (2, 3), (3, 3), (3, 4), (4, 4), (7, 5), (9, 4), (6, 8), (8, 8), (9, 9), (8, 10)

Use KMeans algorithm and group data points into two clusters.

Initial centroids are Centroid 1: (3, 3), Centroid 2: (8, 8)

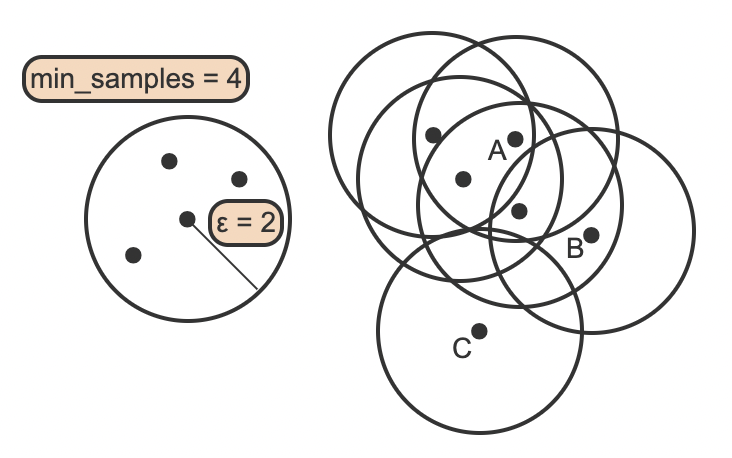
**Question 2:**

A data scientist plans to use DBSCAN with the minimum number of points set to 5. Identify each labeled point in the scatter plot as a border point, core point, or outlier. (Discuss it).

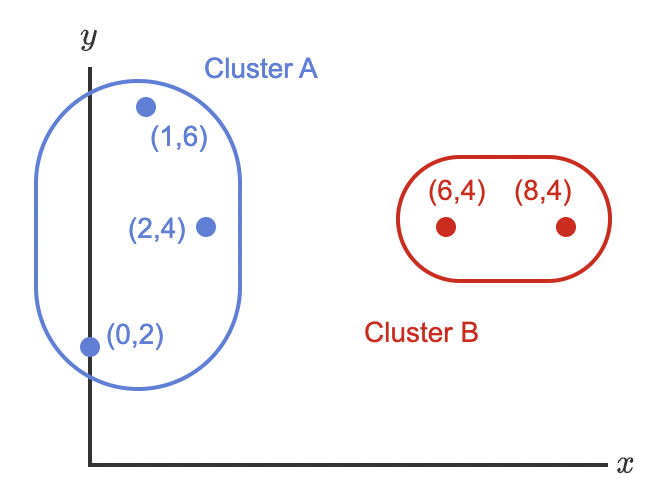


**Question 3:**

Determine whether each labeled point in the figure below is a core point, a boundary point, or an outlier given ε = 2 and the minimum number of points for a core point is 4. (Discuss it).



**Question 4:** What is the distance between the two clusters using centroid linkage?



**Question 5:** Create a dendrogram from the following figure. (Height of dendrogram in your solution is not important.)

A diagram of a diagram

Description automatically generated

**Question 6:**

In real estate, outliers represent a special circumstance that drastically affects the price of a house. One possibility is that a home received multiple offers and one bidder submitted a high offer to guarantee the offer is accepted. Another possibility is that an unexpected event happened and the owner needs to sell quickly.

The figure below uses a dataset that contains 76 single family homes with list price and square feet as features. The values for list price and square feet have been standardized, because the units and ranges of both variables are different.

A graph of a function

Description automatically generated with medium confidence

1. How many clusters of houses are obtained when ε = 1 and min\_samples = 12?
2. Is the 1,440 square foot house listed at $277,000 an outlier? Why?
3. would decreasing ε most likely increase the number of houses identified as outliers? Why?

**Question 7:** Suppose we apply DBSCAN to cluster the following dataset using Euclidean distance.

A graph with red dots and numbers

Description automatically generated

Given that minpoint = 3 and epsilon = 1, answer the following questions.

1. Label all points as “core points”, “boundary points” and “noise”.
2. What is the clustering result?

**Question 8:** After performing anomaly detection, data miner A wants to find clusters of outliers. Data miner B claims that this does not make any sense and suggests that A re-read the definition of an anomaly. Do you think it is meaningful to cluster anomalies? Explain.

**Question 9:** Referring to the figure below, what is the optimal number of clusters for the dataset? Why?

A graph with a blue line

Description automatically generated