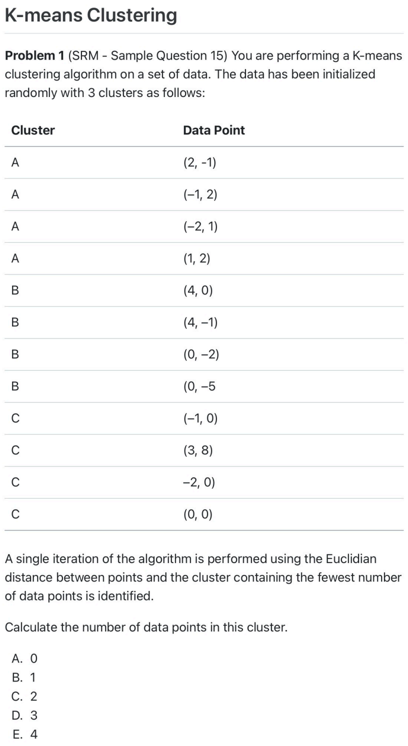
Sean Cottrell
AYU Week 8



K-means Clustering

Problem 1 (SRM - Sample Question 15) You are performing a K-means clustering algorithm on a set of data. The data has been initialized randomly with 3 clusters as follows:

Cluster	Data Point
A	(2, -1)
A	(–1, 2)
A	(–2, 1)
A	(1, 2)
B	(4, 0)
B	(4, –1)
B	(0, –2)
B	(0, –5)
C	(–1, 0)
C	(3, 8)
C	–2, 0)
C	(0, 0)

A single iteration of the algorithm is performed using the Euclidian distance between points and the cluster containing the fewest number of data points is identified.

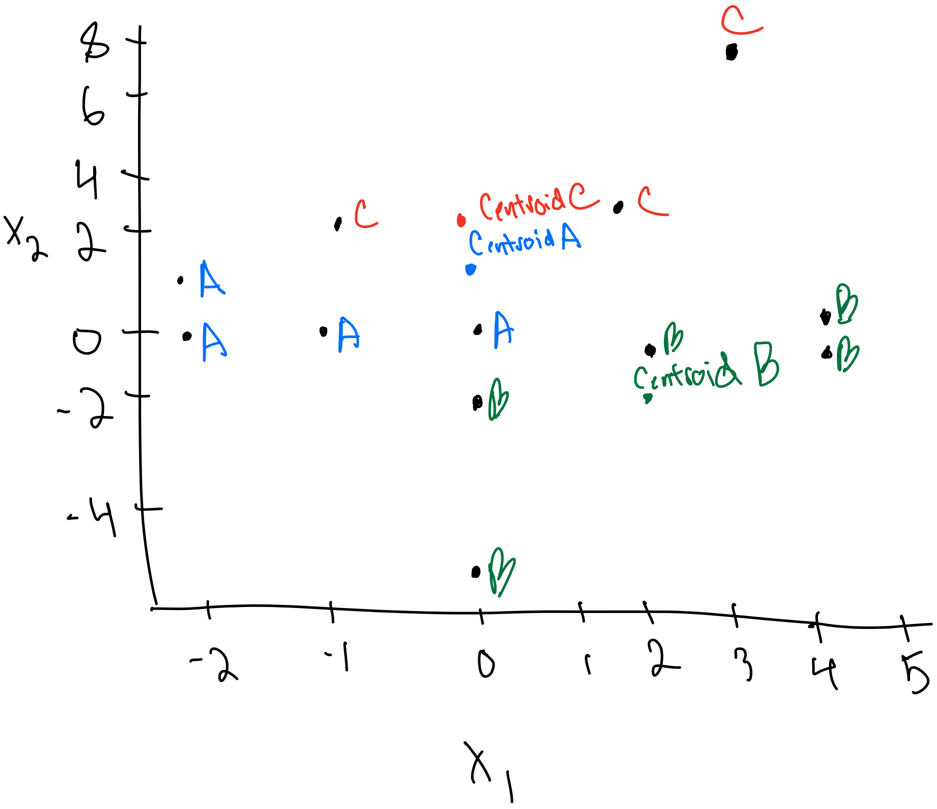
Calculate the number of data points in this cluster.

A. 0
B. 1
C. 2
D. 3
E. 4

1)

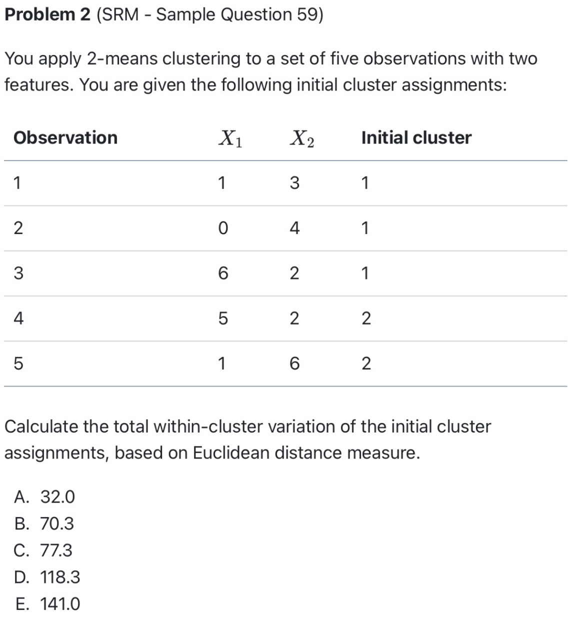
Find the 3 centroids for the clusters

A= (0,1)
B= (2,-2)
C= (0,2)



After graphing, we can tell cluster C has 3 points, which is the fewest points

$\boxed{\text{Answer D}}$



Problem 2 (SRM - Sample Question 59)

You apply 2-means clustering to a set of five observations with two features. You are given the following initial cluster assignments:

Observation	X_1	X_2	Initial cluster
1	1	3	1
2	0	4	1
3	6	2	1
4	5	2	2
5	1	6	2

Calculate the total within-cluster variation of the initial cluster assignments, based on Euclidean distance measure.

A. 32.0
B. 70.3
C. 77.3
D. 118.3
E. 141.0

2)

Centroid of Cluster 1

$$C_1 = \left(\frac{1+0+6}{3}, \frac{3+4+2}{3}\right) = \left(\frac{7}{3}, 3\right)$$

Observation	Distance
1	$(1-7/3)^2+(3-3)^2 = 16/9$
2	$(0-7/3)^2+(4-3)^2 = 58/9$
3	$(6-7/3)^2+(2-3)^2 = 130/9$

$$W(C_1) = 2\left(\frac{16}{9}+\frac{58}{9}+\frac{130}{9}\right) = \frac{136}{3}$$

Centroid for Cluster 2

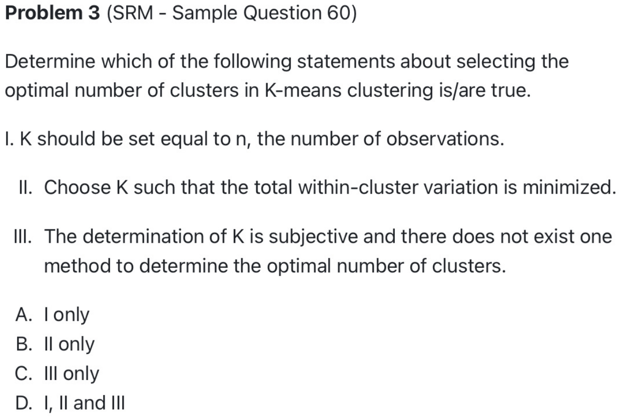
$$C_2 = \left(\frac{5+1}{2}, \frac{2+6}{2}\right) = (3,4)$$

Observation	Distance
4	8
5	8

$$W(C_2) = 2(8+8) = 32$$

Total within cluster variation is 77.33

$\boxed{\text{Answer C}}$



Problem 3 (SRM - Sample Question 60)

Determine which of the following statements about selecting the optimal number of clusters in K-means clustering is/are true.

I. K should be set equal to n, the number of observations.

II. Choose K such that the total within-cluster variation is minimized.

III. The determination of K is subjective and there does not exist one method to determine the optimal number of clusters.

A. I only
B. II only
C. III only
D. I, II and III

3)

Only III is true, therefore $\boxed{\text{Answer C}}$

Problem 4 (SRM - Sample Question 1)

You are given the following four pairs of observations:

$x_1 = (-1, 0),\ x_2 = (1, 1),\ x_3 = (2, -1),\ x_4 = (5, 10)$

A hierarchical clustering algorithm is used with complete linkage and Euclidean distance. Calculate the intercluster dissimilarity between x_1, x_2 and x_4.

A. 2.2

B. 3.2

C. 9.9

D. 10.8

E. 11.7

4) Calculating Euclidean distance between both pairs

$X_1X_4 = \sqrt{(-1-5)^2+(0-10)^2} = \sqrt{136} = 11.7$

$X_2X_4 = \sqrt{(1-5)^2+(1-10)^2} = \sqrt{97} = 9.85$

Taking the max, we find 11.7

$\boxed{\text{Answer E}}$

Problem 5 (SRM - Sample Question 36)

Determine which of the following statements about hierarchical clustering is/are true.

I. The method may not assign extreme outliers to any cluster.

II. The resulting dendrogram can be used to obtain different numbers of clusters.

III. The method is not robust to small changes in the data.

A. None

B. I and II only

C. I and III only

D. II and III only

E. The correct answer is not given by (A), (B), (C), or (D).

5) Both II and III are true therefore

$\boxed{\text{Answer D}}$

Problem 6 (SRM - Sample Question 2)

Determine which of the following statements is/are true.

I. The number of clusters must be pre-specified for both K-means and hierarchical clustering.

II. The K-means clustering algorithm is less sensitive to the presence of outliers than the hierarchical clustering algorithm.

III. The K-means clustering algorithm requires random assignments while the hierarchical clustering algorithm does not.

A. I only

B. II only

C. III only

D. I, II and II

E. The correct answer is not given by (A), (B), (C), or (D)

6) Only III is true, therefore

$\boxed{\text{Answer C}}$

Problem 7 (SRM - Sample Question 16)

Determine which of the following statements is applicable to K-means clustering and is not applicable to hierarchical clustering.

A. If two different people are given the same data and perform one iteration of the algorithm, their results at that point will be the same.

B. At each iteration of the algorithm, the number of clusters will be greater than the number of clusters in the previous iteration of the algorithm.

C. The algorithm needs to be run only once, regardless of how many clusters are ultimately decided to use.

D. The algorithm must be initialized with an assignment of the data points to a cluster.

E. None of (A), (B), (C), or (D) meet the meet the stated criterion.

7) Answer D is true for K means clustering, but false for hierarchial clustering.

Problem 8 (SRM - Sample Question 32)

You are given a set of n observations, each with p features. Determine which of the following statements is/are true with respect to clustering methods.

I. The n observations can be clustered on the basis of the p features to identify subgroups among the observations.

II. The p features can be clustered on the basis of the n observations to identify subgroups among the features.

III. Clustering is an unsupervised learning method and is often performed as part of an exploratory data analysis.

A. None
B. I and II only
C. I and III only
D. II and III only
E. The correct answer is not given by (A), (B), (C), or (D).

8) All I, II, and III are true

Answer E

Proble 9 (SRM - Sample Question 34)

Determine which of the following statements is/are true about clustering methods: I. If K is held constant, K-means clustering will always produce the same cluster assignments.

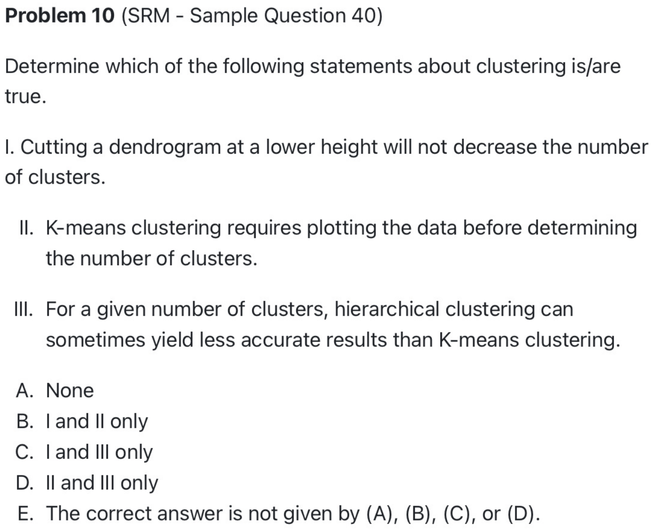
II. Given a linkage and a dissimilarity measure, hierarchical clustering will always produce the same cluster assignments for a specific number of clusters.

III. Given identical data sets, cutting a dendrogram to obtain five clusters produces the same cluster assignments as K-means clustering with K = 5.

A. I only
B. II only
C. III only
D. I, II and III
E. The correct answer is not given by (A), (B), (C), or (D).

9) Only II is true

Answer B



Problem 10 (SRM - Sample Question 40)

Determine which of the following statements about clustering is/are true.

I. Cutting a dendrogram at a lower height will not decrease the number of clusters.

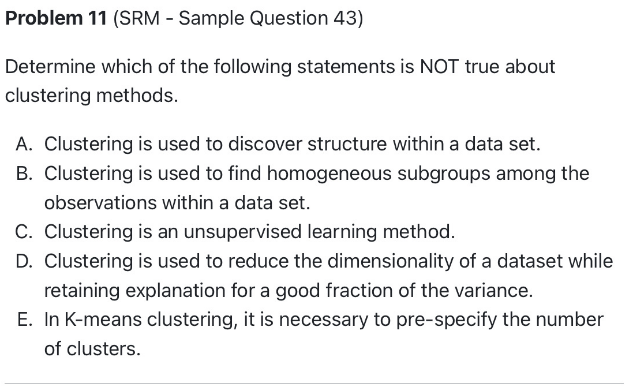
II. K-means clustering requires plotting the data before determining the number of clusters.

III. For a given number of clusters, hierarchical clustering can sometimes yield less accurate results than K-means clustering.

A. None
B. I and II only
C. I and III only
D. II and III only
E. The correct answer is not given by (A), (B), (C), or (D).

10) Both I and III are true

$\boxed{\text{Answer C}}$



Problem 11 (SRM - Sample Question 43)

Determine which of the following statements is NOT true about clustering methods.

A. Clustering is used to discover structure within a data set.
B. Clustering is used to find homogeneous subgroups among the observations within a data set.
C. Clustering is an unsupervised learning method.
D. Clustering is used to reduce the dimensionality of a dataset while retaining explanation for a good fraction of the variance.
E. In K-means clustering, it is necessary to pre-specify the number of clusters.

11) $\boxed{\text{Answer D}}$ is false