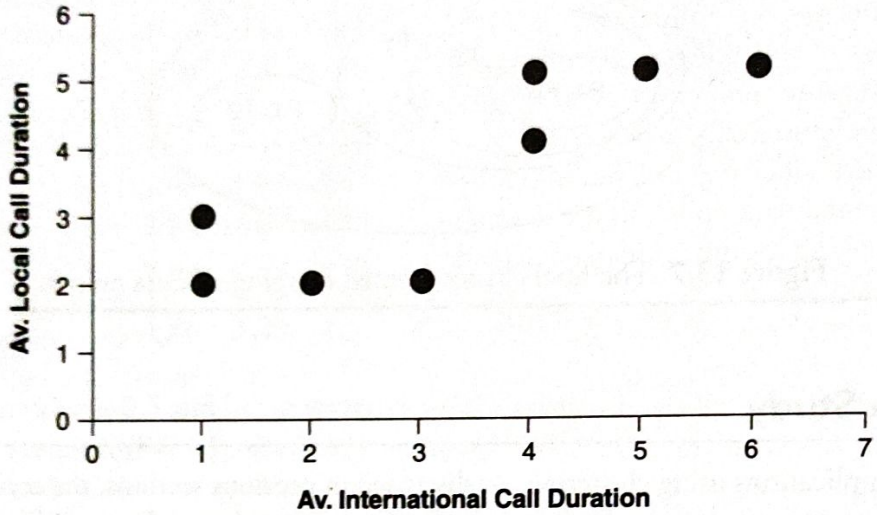
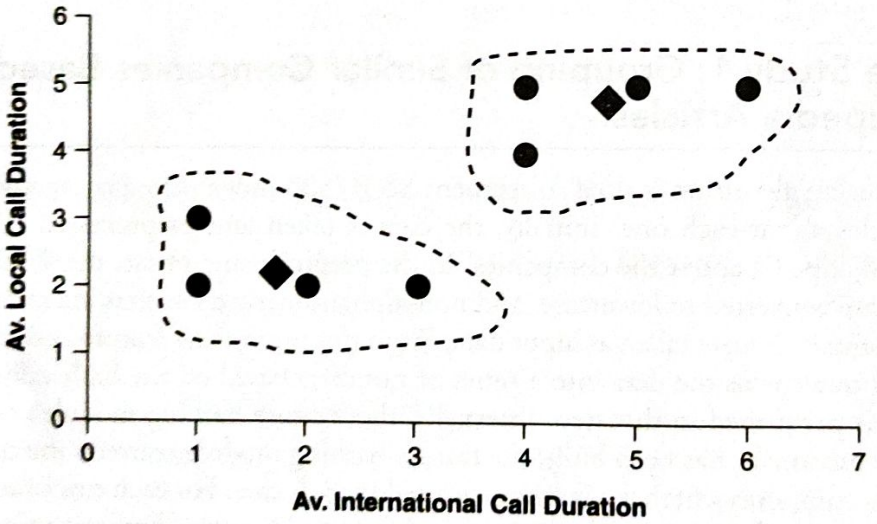
Let us take a small sample size of eight customers. Based on their duration of national and international calls, a scatter plot is drawn as shown below.



Using Euclidean distance metric to compute the centroids, the final clusters forms are shown in the figure below.



Based on the proximity between cluster centres and centriod of the formed clusters, the customer plans can be decided so that the company as well as the customers benefit from a chosen plan.

Summary

- Clustering is the process of grouping together data objects into multiple sets or clusters, so that objects within a cluster have high similarity when compared to objects outside of it.
- Similarity is measured by distance metrics and the most common among them is the Euclidean distance metrics.
- Clustering is also called data segmentation because clustering partitions large datasets into groups according to their similarity.
- Clustering is known as unsupervised learning because the class label information is not present.

- The applications of clustering are varied and include business intelligence, pattern recognition, image processing, biometrics, web technology, search engine, and text mining.
- The requirements of clustering is dependent on its scalability, number and types of attributes which have to be clustered, shape of the cluster to be identified, efficiency in handling noisy data and incremental data points to the existing clusters, handling high-dimensional data, and data with constraints.
- The basic types of clustering are hard clustering and soft clustering depending on whether the data points belong to only one cluster or whether they can be shared among clusters.
- Clustering algorithms are classified based on partitioning, hierarchical, density-based and grid-based clustering.
- Partitioning-based clustering algorithms are distance based. *k*-means and *k*-medoids are popular partition-based clustering algorithms. The number of clusters to be formed is initially specified.
- The result of hierarchical clustering is a tree-based representation of the objects, which is also known as dendrogram.
- Density-based clustering algorithm finds non-linear shaped clusters based on density. Density-based spatial clustering of applications with noise (DBSCAN) is the most widely used density-based algorithm. It uses the concept of density reachability and density connectivity.
- The grid-based clustering approach differs from conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points.

Multiple-Choice Questions

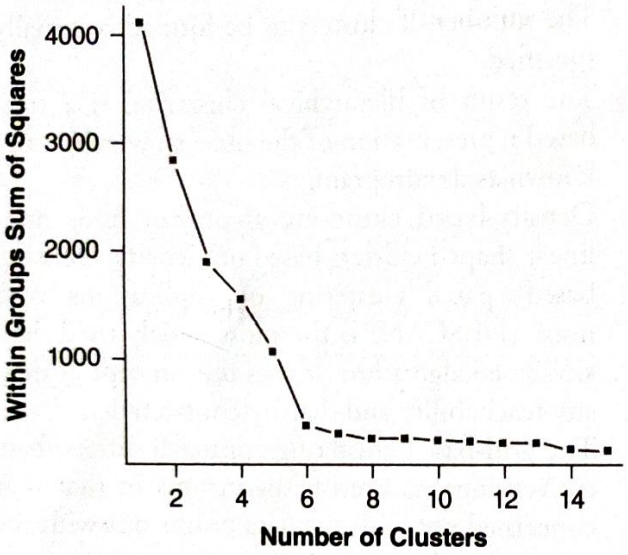
1. Which of the following statements is true?
   (a) Assignment of observations to clusters does not change between successive iterations in *k*-means.
   (b) Assignment of observations to clusters changes between successive iterations in *k*-means.
   (c) Assignment of observations to clusters always decrease between successive iterations in *k*-means.
   (d) Assignment of observations to clusters always increase between successive iterations in *k*-means.

2. Which of the following can act as possible termination conditions in *k*-means?
   i. For a fixed number of iterations.
   ii. Assignment of observations to clusters does not change between iterations, except for cases with a bad local minimum.
   iii. Centroids do not change between successive iterations.
   iv. Terminate when RSS falls below a threshold.
   (a) i, iii, and iv
   (b) i, ii, and iii
   (c) i, ii, and iv
   (d) All of the above

3. Which of the following algorithm is most sensitive to outliers?
   (a) *k*-means clustering algorithm
   (b) *k*-medians clustering algorithm
   (c) *k*-modes clustering algorithm
   (d) *k*-medoids clustering algorithm

4. Which of the following is finally produced by hierarchical clustering?
   (a) Final estimate of cluster centroids
   (b) Tree showing how close things are to each other
   (c) Assignment of each point to clusters
   (d) All of the above

5. What is the best choice for the number of clusters based on the following graph?



(a) 5
(b) 6
(c) 14
(d) None of the above

Very Short Answer Questions

1. Give an example of an application for *k*-means clustering algorithm. Explain in brief.
2. Explain the different distance measures used for clustering.
3. Using *k*-means clustering, cluster the following data into two clusters. Show each step.

{2, 4, 10, 12, 3, 20, 30, 11, 25}

4. Compare between single link, complete link, and average link based on distance formula.
5. Draw the flowchart of *k*-means algorithm.

Short Answer Questions

1. Compute the distance matrix for the *x*–*y* coordinates given in the following table.

Point	x coordinate	y coordinate
p1	0.4005	0.5306
p2	0.2148	0.3854
p3	0.3457	0.3156
p4	0.2652	0.1875
p5	0.0789	0.4139
p6	0.4548	0.3022

2. How will you define the number of clusters in *k*-means clustering algorithm?
3. Use *k*-means algorithm to cluster the following dataset consisting of the scores of two variables on each of the seven individuals.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

4. Use the k-means algorithm and Euclidean distance to cluster the following eight examples into three clusters: A1 = (2, 10), A2 = (2, 5), A3 = (8, 4), A4 = (5, 8), A5 = (7, 5), A6 = (6, 4), A7 = (1, 2), A8 = (4, 9). The distance matrix based on the Euclidean distance is given in the following table.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose the initial seeds (centers of each cluster) are A1, A4, and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:

(a) The new clusters (that is, the examples belonging to each cluster).

(b) The centers of the new clusters.

5. Use single and complete link agglomerative clustering to group the data given in the following distance matrix. Show the dendrograms.

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Review Questions

1. Use k-means algorithm to create three clusters for a given set of values: {2, 3, 6, 8, 9, 12, 15, 1s8, 22}.
2. Apply agglomerative clustering algorithm on the given data and draw the dendogram. Show three clusters with its allocated points by using the single link method.

	a	b	c	d	e	f
a	0	$\sqrt{2}$	$\sqrt{10}$	$\sqrt{17}$	$\sqrt{5}$	$\sqrt{20}$
b	$\sqrt{2}$	0	$\sqrt{8}$	3	1	$\sqrt{18}$
c	$\sqrt{10}$	$\sqrt{8}$	0	$\sqrt{5}$	$\sqrt{5}$	2
d	$\sqrt{17}$	3	$\sqrt{5}$	0	2	3
e	$\sqrt{5}$	1	$\sqrt{5}$	2	0	$\sqrt{13}$
f	$\sqrt{20}$	$\sqrt{18}$	$\sqrt{2}$	3	$\sqrt{13}$	0

3. Apply complete link agglomerative clustering techniques on the given data to find the prominent clusters.

	P1	P2	P3	P4	P5	P6
P1	0	0.23	0.22	0.37	0.34	0.24
P2	0.23	0	0.14	0.19	0.14	0.24
P3	0.22	0.14	0	0.13	0.28	0.10
P4	0.37	0.19	0.13	0	0.23	0.22
P5	0.34	0.14	0.28	0.23	0	0.39
P6	0.24	0.24	0.10	0.22	0.39	0

4. Explain expectation–maximization algorithm.
5. What are the requirements for clustering?
6. What are the applications of clustering?

Answers

Multiple-Choice Answers

1. (a) 2. (d) 3. (a) 4. (b) 5. (b)