DS&AI

Machine Learning

Unsupervised Learning

DPP : 01

Q1 Which of the following is not a Clustering method?

(A) K-Mean method

(B) Self Organizing feature map method

(C) K-nearest neighbor method

(D) Agglomerative method

Q2 For the dataset {(1,1), (2, 2), (3, 3), (8, 8), (9,9)), use single linkage clustering to find the distance at which the first two clusters are merged.

Q3 Which of the following clustering methods is most likely to result in elongated clusters ?

(A) K-means clustering

(B) Complete linkage clustering

(C) Single linkage clustering

(D) Average linkage clustering

Q4 In a k-medoids algorithm with k = 3 and a dataset of 300 points, if the medoids are chosen at random, what is the total number of possible sets of medoids if the dataset is large enough?

(A) $\begin{pmatrix} 300 \\ 3 \end{pmatrix}$

(B) 300^3

(C) $\dfrac{300!}{3!(300-3)!}$

(D) 300 × 299 × 298

Q5 In Principal Component Analysis (PCA), if the eigenvalues for the first three principal components are 5, 3, and 2, respectively, and the total variance in the dataset is 20, what is the percentage of variance explained by the first two principal components ?

(A) 40% (B) 60%

(C) 80% (D) 90%

Q6 Consider a dataset with 10,000 data points and 20 features, and PCA is used to reduce the dimensionality to 8 principal components. If each principal component is computed using an eigenvalue decomposition of the covariance matrix, how many multiplications are needed to compute the eigenvalues and eigenvectors if the covariance matrix is 20 × 20 ? (Find approximate answer)

(A) 800 (B) 1600

(C) 8000 (D) 80000

Q7 If a dataset of 800 samples is clustered into 4 clusters using K-means, and the within-cluster sum of squares (WCSS) for each cluster is 120, 100, 80, and 70 respectively, what is the total WCSS for the dataset ?

(A) 350 (B) 370

(C) 400 (D) 450

Q8 In a dataset with 500 data points and 15 features, K-means clustering is applied with k = 7 If the average distance between each data point and its assigned centroid after convergence is 3.5, what is the total within-cluster sum of squares (WCSS) for the dataset ?

(A) 6125 (B) 87500

(C) 122500 (D) 175000

Q9 Given the distance matrix below, which pair of points will be merged first in single linkage clustering ?

	x_1	x_2	x_3	x_4
x_1	0	4	3	6
x_2	4	0	5	7

Android App | iOS App | PW Website

x_3	3	5	0	8
x_4	6	7	8	0

(A) (x_1, x_2) (B) (x_2, x_3)

(C) (x_1, x_3) (D) None

Q10 In PCA, the principal components are:

(A) Orthogonal vectors.

(B) Parallel vectors.

(C) The eigenvectors corresponding to the largest eigenvalues of the covariance matrix

(D) The eigenvectors corresponding to the smallest eigenvalues of the covariance matrix.

Q11 You are given the following data point in one - dimensional space: $x_1 = 1$ $x_2 = 3$ $x_3 = 8$ and $x = 10$ If single linkage clustering is applied, what is the distance between the clusters $\{x_1, x_2\}$ and $\{x_3, x_4)$?

Q12 Which of the following are advantages of hierarchical clustering over K-means clustering ?

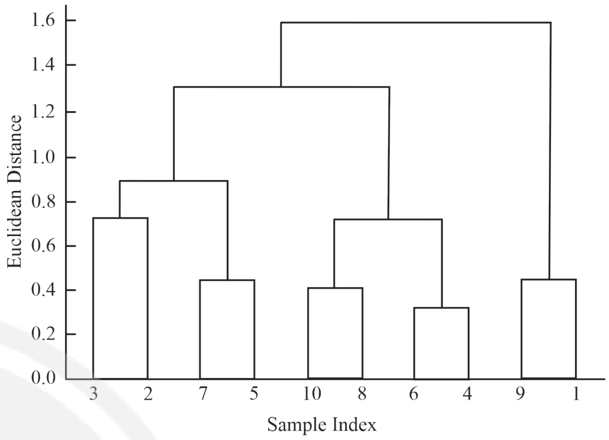
(A) No need to specify the number of clusters in advance.

(B) Hierarchical clustering can find non-convex clusters.

(C) It is less computationally intensive than K-means.

(D) Dendrograms can provide insight into the data structure.

Q13 Consider the following dendogram :



Which is most similar and dissimilar pair.

(A) (1, 9) and (1, 10)

(B) (6, 4) and (3, 5)

(C) (8, 10) and (1, 10)

(D) (6, 4) and (1, 3)

Answer Key

Q1	(C)
Q2	$\Rightarrow \sqrt{2}$
Q3	(C)
Q4	(A, C)
Q5	(A)
Q6	(C)
Q7	(B)
Q8	(A)
Q9	(C)
Q10	(A, C)
Q11	5
Q12	(A, B, D)
Q13	(D)

Android App | iOS App | PW Website

Hints & Solutions

Q1 Text Solution:

(C) K-nearest neighbor method

Here's a brief explanation of each method:

- **K-Mean method**: This is a clustering method where data points are grouped into kk clusters based on their similarity.
- **Self Organizing Feature Map (SOFM) method**: This is a type of neural network used for clustering and dimensionality reduction, often referred to as unsupervised learning.
- **K-nearest neighbor method**: This is a classification method, not a clustering method. It classifies data points based on the labels of their nearest neighbors.
- **Agglomerative method**: This is a type of hierarchical clustering method that builds clusters by successively merging smaller clusters.

So, K-nearest neighbor (C) is not a clustering method; it's used for classification.

Q2 Text Solution:

[Use Eucledian distance]

(1, 1)

(2, 2)

(3, 3)

(8, 8)

(9, 9)

So first of all the (1, 1), (2, 2) or (2, 2) (3, 3) distance between these

$$\Rightarrow \sqrt{1^2+1^2} \Rightarrow \sqrt{2}$$

Q3 Text Solution:

The clustering method most likely to result in elongated clusters is:

(C) Single linkage clustering

Explanation:

- **(A) K-means clustering**: K-means tends to form spherical clusters because it minimizes the variance within each cluster. It generally works best when the clusters are roughly spherical in shape and of similar size.
- **(B) Complete linkage clustering**: This method tends to produce more compact clusters. It merges clusters based on the maximum distance between points in the clusters, which can result in more rounded clusters.
- **(C) Single linkage clustering**: This method, also known as nearest neighbor clustering, tends to form elongated clusters. It merges clusters based on the minimum distance between any pair of points in the clusters. This can lead to "chaining" effects, where clusters can become long and thin if points are linked through a chain of close distances.
- **(D) Average linkage clustering**: This method merges clusters based on the average distance between all pairs of points in the clusters. It generally produces more balanced clusters but is less likely to create highly elongated clusters compared to single linkage.

Q4 Text Solution:

In K medoid we select 3 points in 300 data points

$$\Rightarrow \text{No of ways} = 300C_3 \Rightarrow \binom{300}{3} = \frac{300!}{3!(300-3)!}$$

Q5 Text Solution:

Total Variance : 20

3 eigen values : 5, 3, 2

So the variance explained by any Component = it 's λ.

So first two PC = $\lambda_1 + \lambda_2 = 8$

So % of Variance explained = 8/20 × 100 → 40%.

Q6 Text Solution:

any matrix of order n × n will need calculation of order (n^3) for finding eigen value and eigen vectors.

→ So here we need $20^3 \approx 8000$ calculations.

Q7 Text Solution:

Total wcss of data set = wcss of each cluster

= 120 + 100 + 80 + 70

= 370.

Q8 Text Solution:

So WCSS ⇒Σ (distance of point from its Centroid)^2

All data points

⇒ 500 × $(3.5)^2$

⇒ 6125

Q9 Text Solution:

Since distance is of $(x_3 \rightarrow x_1)$ So (x_3, x_1) are merged.

Q10 Text Solution:

(A) Orthogonal vectors.

and

(C) The eigenvectors corresponding to the largest eigenvalues of the covariance matrix.

Here's why:

- **Orthogonal Vectors**: The principal components in PCA are orthogonal to each other. This orthogonality ensures that the principal components are uncorrelated and represent different directions of maximum variance in the data.
- **Eigenvectors Corresponding to the Largest Eigenvalues**: The principal components are the eigenvectors of the covariance matrix of the data. To capture the maximum variance in the data, PCA selects the eigenvectors corresponding to the largest eigenvalues. These eigenvectors represent the directions in which the data varies the most.

So, while option (A) describes the orthogonality of the principal components, option (C) correctly

Q11 Text Solution:

$\{x_1, x_2\} = (1, 3)$

$\{x_3, x_4) = (8, 10)$

Distance ⇒ (1, 3) ↔ (8,10)

Minimum distance is between 3 – 8 = 5.

Q12 Text Solution:

(C) It is less computationally intensive than K-means: This is generally not true. Hierarchical clustering can be more computationally intensive, especially for large datasets, compared to K-means clustering, which is typically faster and more scalable. K-means involves iterating over data points to refine cluster centroids, which can be less computationally demanding than the pairwise distance calculations required in hierarchical clustering.

Q13 Text Solution:

we can see that (6, 4) merge at least distance they are most similar

So (1, 9) will be most dissimilar with (3,2,7,5,10,8,6,4)

So most dissimilar ⟹ (1, 3).

Android App | iOS App | PW Website