Data Science & Artificial Intelligence

Machine Learning

Unsupervised Learning Discussion Notes

DPP - 01

By- Siddharth Sir

[MCQ]

#Q. 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

Solution: (C)

K-nearest neighbor method

[NAT]

#Q. 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.

[Use Euclidian Distance]

Solution:

(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

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

[MCQ]

#Q. 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

Solution: (C)

Single linkage clustering

[MSQ]

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

A $\binom{300}{3}$

B 300^3

C $\frac{300!}{3!\,(300-3)!}$

D $300 \times 299 \times 298$

Solution: (A, C)

In k-medoids we select 3 points in 300 data points.

$$\text{No. of ways} = {}^{300}\mathrm{C}_3 = \binom{300}{3} = \frac{300!}{3!\,(300-3)!}$$

[MCQ]

#Q. 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%

Solution: (A)

Total Variance : 20

3 Eigen values: 5, 3, 2

So, the variance explained by any component = λ

So, the variance explained by first two PC = $\lambda_1 + \lambda_2 = 8$

So, % of variance explained = $\frac{8}{20} \times 100 = 40\%$

[MCQ]

#Q. 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 value)

A 800

B 1600

C 8000

D 80000

Solution: (C)

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.

[MCQ]

#Q. 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

Solution: (B)

Total WCSS of dataset

= WCSS of each cluster

= 120 + 100 + 80 + 70

= 370

[MCQ]

#Q. 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

Solution: (A)

So, WCSS

$= \Sigma(\text{distance of point from it's centroid all data poitns})^2$

$= 500 \times (3.5)^2$

$= 6125$

[MCQ]

#Q. 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
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

Solution: (C)

Since, minimum distance is of $(x_3 \rightarrow x_1)$.

So, (x_3, x_1) are merged.

[MSQ]

#Q. 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.

Solution: (A, C)

Orthogonal vectors.

The eigenvectors corresponding to the largest eigenvalues of the covariance matrix.

[NAT]

#Q. You are given the following data point in one-dimensional space: $x_1 = 1$, $x_2 = 3$, $x_3 = 9$ and $x_4 = 10$. If single linkage clustering is applied, what is the distance between the clusters $\{x_1, x_2\}$ and $\{x_3, x_4\}$?

Solution:

$(x_1, x_2) = (1, 3)$

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

Distance $\Rightarrow (1, 3) \leftrightarrow (8, 10)$

Minimum distance is b/w 3 – 8 = 5

[MSQ]

#Q. 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.

Solution: (A, B, D)

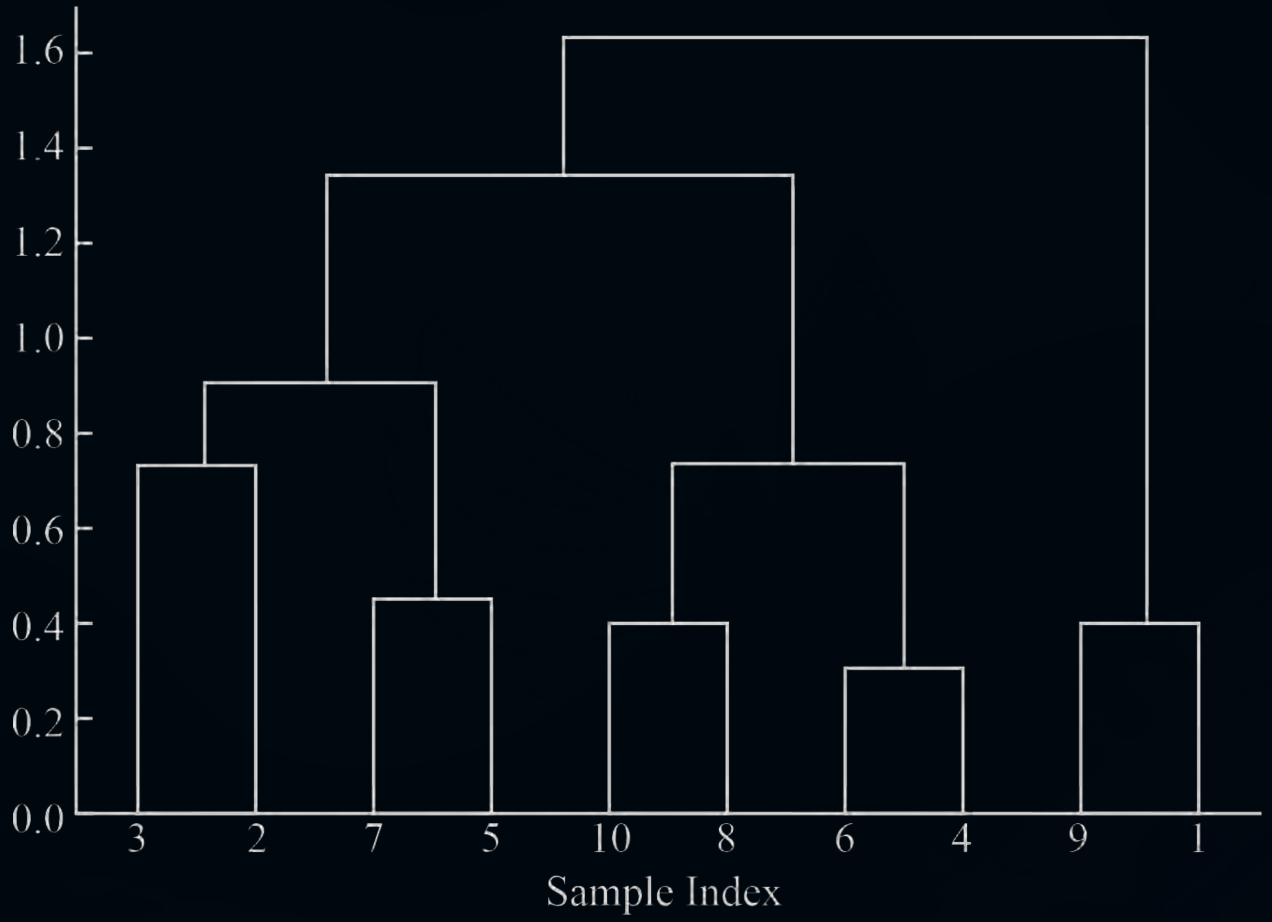
No need to specify the number of clusters in advance.

Hierarchical clustering can find non-convex clusters.

Dendrograms can provide insight into the data structure.

[MCQ]

#Q. Consider the following dendrogram:
Which is the 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)

Solution: (D)

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)

THANK - YOU