1. Which of the following is an application of clustering?

a. Biological network analysis

b. Market trend prediction

c. Topic modeling

d. All of the above

Ans: d All the above.

2. On which data type, we cannot perform cluster analysis?

a. Time series data

b. Text data

c. Multimedia data

d. None

Ans: d None

3. Netflix’s movie recommendation system uses

a. Supervised learning

b. Unsupervised learning

c. Reinforcement learning and Unsupervised learning

d. All of the above

Ans c Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is

a. The number of cluster centroids

b. The tree representing how close the data points are to each other

c. A map defining the similar data points into individual groups

d. All of the above

Ans b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?

a. A distance metric

b. Initial number of clusters

c. Initial guess as to cluster centroids

d. None

Ans d None

6. Which is the following is wrong?

a. k-means clustering is a vector quantization method

b. k-means clustering tries to group n observations into k clusters

c. k-nearest neighbour is same as k-means

d. None

Ans c. k-nearest neighbour is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in

hierarchical clustering?

i. Single-link

ii. Complete-link

iii. Average-link

Options:

a.1 and 2

b. 1 and 3

c. 2 and 3

d. 1, 2 and 3

8. Which of the following are true?

i. Clustering analysis is negatively affected by multicollinearity of features

ii. Clustering analysis is negatively affected by heteroscedasticity

Options:

a. 1 only

b. 2 only

c. 1 and 2

d. None of them

Ans a.1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?

Ans a.2

10. For which of the following tasks might clustering be a suitable approach?

a.Given sales data from a large number of products in a supermarket, estimate future sales for each

of these products.

b.Given a database of information about your users, automatically group them into different market

segments.

c. Predicting whether stock price of a company will increase tomorrow.

d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

Ans b.Given a database of information about your users, automatically group them into different market segments.

11.

First cluster is formed between P3 and P4 since it has the lowest value.

Update the distance :

MIN(dist(P3,P1),(P6,P1))

=MIN[(0.2218,0.2347)]

= 0.2218

MIN(dist(P3,P2),(P6,P2))

=MIN[(0.148,0.0.2540)]

= 0.148

MIN(dist(P3,P4),(P6,P4))

=MIN[(0.1513,0.2216)]

= 0.1513

MIN(dist(P3,P5),(P6,P5))

=MIN[(0.2843,0.3921)]

= 0.2843

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | UpdatedValues |  |  |  |
| P1 | P2 | P3,P6 | P4 | P5 |
| 0 |  |  |  |  |
| 0.235 | 0 |  |  |  |
| 0.221 | 0.148 | 0 |  |  |
| 0.368 | 0.204 | 0.151 | 0 |  |
| 0.342 | 0.138 | 0.284 | 0.293 | 0 |
|  |  |  |  |  |

Second cluster is between P5 and P2

MIN(dist(P2,P1),(P5,P1))

=MIN[(0.221,0.341)]

= 0.221

MIN(dist(P2,(P3,P6),(P5,(P3,P6))

=MIN[(0.151,0.284)]

= 0.151

MIN(dist(P2,P4),(P5,P4)

=MIN[(0.204,0.293)]

= 0.204

Updated values

|  |  |  |  |
| --- | --- | --- | --- |
| P1 | P2,P5 | P3,P6 | P4 |
| 0 |  |  |  |
| 0.235 | 0 |  |  |
| 0.221 | 0.151 | 0 |  |
| 0.368 | 0.204 | 0.151 | 0 |
|  |  |  |  |

Next cluster will be formed between P2,P5 and P3,P6

MIN(dist((P2,P5),P1)(P3,P6),P1))

=MIN[(0.235,0.221)]

= 0.221

MIN(dist((P2,P5),P4)(P3,P6),P4))

=MIN[(0.204,0.151)]

= 0.151

|  |  |  |
| --- | --- | --- |
| P1 | P2P5P3P6 | P4 |
| 0 |  |  |
| 0.22 | 0 |  |
| 0.368 | 0.151 | 0 |
|  |  |  |

Next cluster is between P2,P5,P3,P6 and P4

MIN [dist(P2,P5,P3,P6),P1,(P4,P1)]

MIN[(0.22,0.368)

0.22

|  |  |  |
| --- | --- | --- |
|  | P1 | P2,P5,P3,P6,P4 |
| P1 | 0 |  |
| P2,P5,P3,P6,P4 | 0.22 | 0 |

So the last cluster is P2,P5,P3,P6,P4 and P1

**Hence the answer is a)**

12. First cluster is formed between P3 and P6

Max(dist[P3,P1),(P6,P1)]

Max(0.2218,0.2347)

0.2347

Max(dist[P3,P2),(P6,P2)]

Max(0.1483,0.2540)

0.2540

Max(dist[P3,P4),(P6,P4)]

Max(0.1513,0.0.2216)

0.2216

Max(dist[P3,P5),(P6,P5)]

Max(0.2843,0.3921)

0.3921

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P1 | P2 | P3,P6 | P4 | P5 |
| 0 |  |  |  |  |
| 0.235 | 0 |  |  |  |
| 0.234 | 0.254 | 0 |  |  |
| 0.368 | 0.204 | 0.2216 | 0 |  |
| 0.342 | 0.138 | 0.3921 | 0.293 | 0 |

Second cluster is formed between P2 and P5

Max(dist[P2,P1),(P5,P1)]

Max(0.234,0.342)

0.342

Max(dist[P2,(P3,P6),(P5(P3,P6)]

Max(0.254,0.3921)

0.3921

Max(dist[P2,P4),(P5,P4)]

Max(0.204,0.293)

0.293

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| P1 | P2,P5 | P3,P6 | P4 |  |
| 0 |  |  |  |  |
| 0.34 |  |  |  |  |
| 0.234 | 0.3921 | 0 |  |  |
| 0.368 | 0.293 | 0.2216 | 0 |  |

Cluster is formed between P3,P6 and P4

Max(dist[(P3,P6)P1,( P4,P1)]

Max(0.234,0.368)

0.368

Max(dist[(P3,P6)(P2,P5),( P4,(P2,P5)]

Max(0.3921,0.368)

0.3921

|  |  |  |  |
| --- | --- | --- | --- |
| P1 | P2,P5 | P3,P6,P4 |  |
| 0 |  |  |  |
| 0.34 | 0 |  |  |
| 0.368 | 0.3921 | 0 |  |

Cluster will be between P2,P5and P1

Max(dist[(P2,P5)(P3,P6,P4),( P1,( P3,P6,P4)]

Max(0.3921,0.368)

0.3921

|  |  |  |
| --- | --- | --- |
| P1,P2P5 | P3,P6,P4 |  |
| 0 |  |  |
| 0.3921 | 0 |  |

Last cluster is between P1,P2,P5 and P3,P6,P4

**Hence the answer is a**

13. What is the importance of clustering?

Clustering is a type of unsupervised machine learning that involves grouping together the dataset into clusters such that each cluster has common attributes. Hence when presented with unstructured data, we can draw inferences from the data by using clustering technique. For example, if we have a huge data set that has stored user experience and usage of online stores from the users, we can identify factors that attract customers and also identify the types of customers and their shopping patterns which will help make business strategies.

14. How can I improve my clustering performance.

If we think about clustering, it is based on patters on the basis of which they can be categorized into each cluster. So it is important that we have enough data so that the dataset can be split into more that two groups. There should also be strong commonality between the each element in the cluster.The data shouldn’t be biased, so it should be collected from different sources.