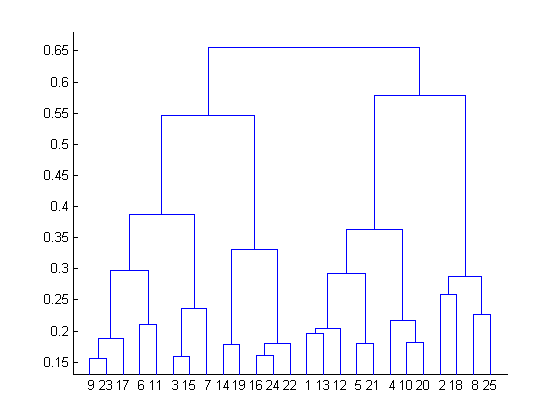
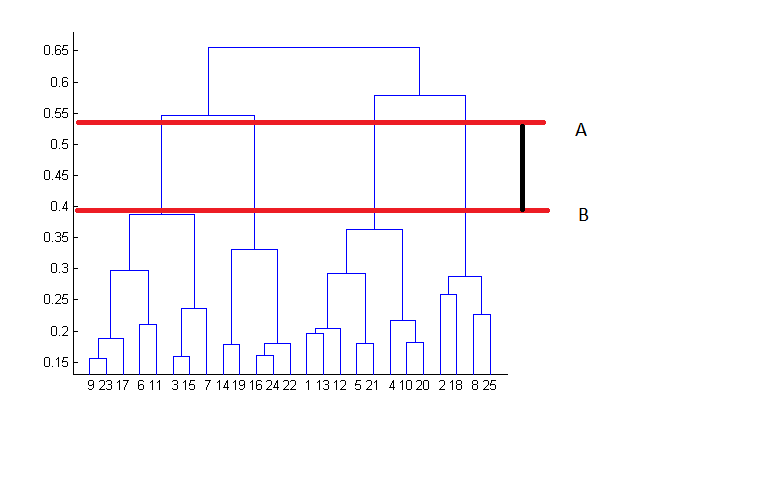
1. What is the most appropriate number of clusters for data points represented by following dendogram?



=

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers

2. Data points with different densities

3. Data points with round shapes

4. Data points with non-convex shapes

Options:

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4

E. 1, 2, 3 and 4

= d)

K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.

3.The most important part of \_\_\_\_\_\_ is selecting the variables on which clustering is based.

a) interpreting and profiling clusters

b) selecting a clustering procedure

c) assessing the validity of clustering

d) formulating the clustering problem

=d)

4. The most commonly used measure of similarity is the \_\_\_\_\_\_\_ or its square.

a) Euclidean distance

b) city-block distance

c) Chebyshev’s distance

d) Manhattan distance

= a) Euclidean distance

5. \_\_\_\_\_\_\_ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

a) Non-hierarchical clustering

b) Divisive clustering

c) Agglomerative clustering

d) K-means clustering

=b)

6. Which of the following is required by K-means clustering?

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

d) All answers are correct

=d)

7.The goal of clustering is to-

a) Divide the data points into groups

b) Classify the data point into different classes

c) Predict the output values of input data points

d) All of the above

=b)

8. Clustering is a-

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

=b)

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a) K- Means clustering

b) Hierarchical clustering

c) Diverse clustering

d) All of the above

=d)

10. Which version of the clustering algorithm is most sensitive to outliers?

a) K-means clustering algorithm

b) K-modes clustering algorithm

c) K-medians clustering algorithm

d) None

=a)

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

a) Data points with outliers

b) Data points with different densities

c) Data points with non-convex shapes

d) All of the above

=d)

12. For clustering, we do not require-

a) Labelled data

b) Unlabeled data

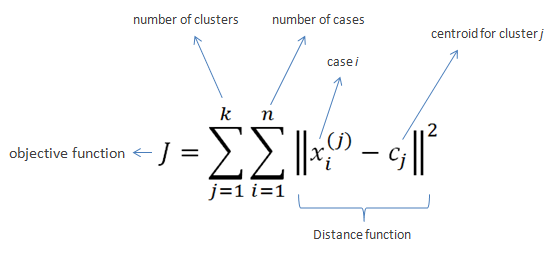
c) Numerical data

d) Categorical data

=a) Labelled data

13. How is cluster analysis calculated?

=let us take kmeans clustering, K-means clustering intends to partition n objects into k clustering in which each objects belongs to cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to greatest separation is not known as priori and must computed from data. The objective of k-means clustering is to minimize total intra-cluster variance or the squared error function.



|  |  |  |
| --- | --- | --- |
| **Algorithm** |  |  |
| 1. Clusters the data into *k* groups where *k*  is predefined. 2. Select *k* points at random as cluster centers. 3. Assign objects to their closest cluster center according to the *Euclidean distance* function. 4. Calculate the centroid or mean of all objects in each cluster. 5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | ***Example***: | | | | | | | |  | | | | | |  | | | | | | | Suppose we want to group the visitors to a website using just their age (one-dimensional space) as follows: | | | | | | | |  | | | | | |  | | | | | | | ***n* = 19** | | | | | | | |  | | | | | |  | | | | | | | 15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65 | | | | | | | |  | | | | | |  | | | | | | |  | | | | | | | |  | | | | | |  | | | | | | | **Initial clusters (random centroid or average):** | | | | | | | |  | | | | | |  | | | | | | | ***k* = 2** | | | | | | | |  | | | | | |  | | | | | | | *c1* = 16 *c2* = 22 | | | | | | | |  | | | | | |  | | | | | | | https://www.saedsayad.com/images/kmeans_distance.png | | | | | | | |  | | | | | |  | | | | | | | **Iteration** **1**: | | | | | | | |  | | | | | |  | | | | | | | *c1* = 15.33 *c2*  = 36.25 | | | | | | | |  | | | | | |  | | | | | | | *xi* | | *c1* | *c2* | Distance 1 | Distance 2 | Nearest Cluster | New Centroid | | 15 | | 16 | 22 | 1 | 7 | 1 | **15.33** | | 15 | | 16 | 22 | 1 | 7 | 1 | | 16 | | 16 | 22 | 0 | 6 | 1 | | 19 | | 16 | 22 | 9 | 3 | 2 | **36.25** | | 19 | | 16 | 22 | 9 | 3 | 2 | | 20 | | 16 | 22 | 16 | 2 | 2 | | 20 | | 16 | 22 | 16 | 2 | 2 | | 21 | | 16 | 22 | 25 | 1 | 2 | | 22 | | 16 | 22 | 36 | 0 | 2 | | 28 | | 16 | 22 | 12 | 6 | 2 | | 35 | | 16 | 22 | 19 | 13 | 2 | | 40 | | 16 | 22 | 24 | 18 | 2 | | 41 | | 16 | 22 | 25 | 19 | 2 | | 42 | | 16 | 22 | 26 | 20 | 2 | | 43 | | 16 | 22 | 27 | 21 | 2 | | 44 | | 16 | 22 | 28 | 22 | 2 | | 60 | | 16 | 22 | 44 | 38 | 2 | | 61 | | 16 | 22 | 45 | 39 | 2 | | 65 | | 16 | 22 | 49 | 43 | 2 | |  | | | | | | | | | | |  | | | | | | | | |  | | **Iteration** **2**: | | | | | | | | | | |  | | | | | | | | |  | | *c1* = 18.56 *c2*  = 45.90 | | | | | | | | | | |  | | | | | | | | |  | | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | *xi* | *c1* | *c2* | Distance 1 | Distance 2 | Nearest Cluster | New Centroid | | 15 | 15.33 | 36.25 | 0.33 | 21.25 | 1 | **18.56** | | 15 | 15.33 | 36.25 | 0.33 | 21.25 | 1 | | 16 | 15.33 | 36.25 | 0.67 | 20.25 | 1 | | 19 | 15.33 | 36.25 | 3.67 | 17.25 | 1 | | 19 | 15.33 | 36.25 | 3.67 | 17.25 | 1 | | 20 | 15.33 | 36.25 | 4.67 | 16.25 | 1 | | 20 | 15.33 | 36.25 | 4.67 | 16.25 | 1 | | 21 | 15.33 | 36.25 | 5.67 | 15.25 | 1 | | 22 | 15.33 | 36.25 | 6.67 | 14.25 | 1 | | 28 | 15.33 | 36.25 | 12.67 | 8.25 | 2 | **45.9** | | 35 | 15.33 | 36.25 | 19.67 | 1.25 | 2 | | 40 | 15.33 | 36.25 | 24.67 | 3.75 | 2 | | 41 | 15.33 | 36.25 | 25.67 | 4.75 | 2 | | 42 | 15.33 | 36.25 | 26.67 | 5.75 | 2 | | 43 | 15.33 | 36.25 | 27.67 | 6.75 | 2 | | 44 | 15.33 | 36.25 | 28.67 | 7.75 | 2 | | 60 | 15.33 | 36.25 | 44.67 | 23.75 | 2 | | 61 | 15.33 | 36.25 | 45.67 | 24.75 | 2 | | 65 | 15.33 | 36.25 | 49.67 | 28.75 | 2 | | | | | | | | | | | | | | | | | | | | | | | | **Iteration** **3**: | | | | | | | | | | | | | |  | | | | |  | | | | | *c1* = 19.50 *c2* = 47.89 | | | | | | | | | | | | | |  | | | | |  | | | | | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | *xi* | *c1* | *c2* | Distance 1 | Distance 2 | Nearest Cluster | New Centroid | | 15 | 18.56 | 45.9 | 3.56 | 30.9 | 1 | **19.50** | | 15 | 18.56 | 45.9 | 3.56 | 30.9 | 1 | | 16 | 18.56 | 45.9 | 2.56 | 29.9 | 1 | | 19 | 18.56 | 45.9 | 0.44 | 26.9 | 1 | | 19 | 18.56 | 45.9 | 0.44 | 26.9 | 1 | | 20 | 18.56 | 45.9 | 1.44 | 25.9 | 1 | | 20 | 18.56 | 45.9 | 1.44 | 25.9 | 1 | | 21 | 18.56 | 45.9 | 2.44 | 24.9 | 1 | | 22 | 18.56 | 45.9 | 3.44 | 23.9 | 1 | | 28 | 18.56 | 45.9 | 9.44 | 17.9 | 1 | | 35 | 18.56 | 45.9 | 16.44 | 10.9 | 2 | **47.89** | | 40 | 18.56 | 45.9 | 21.44 | 5.9 | 2 | | 41 | 18.56 | 45.9 | 22.44 | 4.9 | 2 | | 42 | 18.56 | 45.9 | 23.44 | 3.9 | 2 | | 43 | 18.56 | 45.9 | 24.44 | 2.9 | 2 | | 44 | 18.56 | 45.9 | 25.44 | 1.9 | 2 | | 60 | 18.56 | 45.9 | 41.44 | 14.1 | 2 | | 61 | 18.56 | 45.9 | 42.44 | 15.1 | 2 | | 65 | 18.56 | 45.9 | 46.44 | 19.1 | 2 | | | | | | | | | | | | | | | | |  | | | | |  | | | |  | | | | | | | | | | | | | | | |  | | | | |  | | | | **Iteration** **4**: | | | | | | | | | | | | | | | | | |  | | | | |  | | | *c1* = 19.50 *c2* = 47.89 | | | | | | | | | | | | | | | | | |  | | | | |  | | | |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | *xi* | *c1* | *c2* | Distance 1 | Distance 2 | Nearest Cluster | New Centroid | | 15 | 19.5 | 47.89 | 4.50 | 32.89 | 1 | **19.50** | | 15 | 19.5 | 47.89 | 4.50 | 32.89 | 1 | | 16 | 19.5 | 47.89 | 3.50 | 31.89 | 1 | | 19 | 19.5 | 47.89 | 0.50 | 28.89 | 1 | | 19 | 19.5 | 47.89 | 0.50 | 28.89 | 1 | | 20 | 19.5 | 47.89 | 0.50 | 27.89 | 1 | | 20 | 19.5 | 47.89 | 0.50 | 27.89 | 1 | | 21 | 19.5 | 47.89 | 1.50 | 26.89 | 1 | | 22 | 19.5 | 47.89 | 2.50 | 25.89 | 1 | | 28 | 19.5 | 47.89 | 8.50 | 19.89 | 1 | | 35 | 19.5 | 47.89 | 15.50 | 12.89 | 2 | **47.89** | | 40 | 19.5 | 47.89 | 20.50 | 7.89 | 2 | | 41 | 19.5 | 47.89 | 21.50 | 6.89 | 2 | | 42 | 19.5 | 47.89 | 22.50 | 5.89 | 2 | | 43 | 19.5 | 47.89 | 23.50 | 4.89 | 2 | | 44 | 19.5 | 47.89 | 24.50 | 3.89 | 2 | | 60 | 19.5 | 47.89 | 40.50 | 12.11 | 2 | | 61 | 19.5 | 47.89 | 41.50 | 13.11 | 2 | | 65 | 19.5 | 47.89 | 45.50 | 17.11 | 2 | | | | | | | | | | | | | | | | | | | | |  | | | |  | | |  | | | | | | | | | | | | | | | | | | | |  | | | |  | | | No change between iterations 3 and 4 has been noted. By using clustering, 2 groups have been identified 15-28 and 35-65. The initial choice of centroids can affect the output clusters, so the algorithm is often run multiple times with different starting conditions in order to get a fair view of what the clusters should be. | | | | | | | | | | | | | | | | | | | |  | | | |  | | |  |  |

14. How is cluster quality is measured?

=cluster quality can be measured using precision/Recall, confusion matrix, entropy, F-measure

15.what is cluster analysis and its types?

=Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

four basic types of cluster analysis used in data science. These types are Centroid Clustering, Density Clustering, Distribution Clustering, and Connectivity Clustering.

Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you’re a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

The algorithm will start by randomly selecting centroids (cluster centers) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iteratations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats.

Density Clustering

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

Distribution Clustering

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.

These maps look a lot like targets at an archery range. In the event that a data point hits the bulls eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bulls eye represents lessening percentage or certainty.

Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to acluster.

Connectivity Clustering

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.