MACHINE_LEARNING_ASSIGNMENT – 1

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram?

a) 2

b) 4

c) 6

d) 8

SOLUTION: B

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

SOLUTION: D

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) Accessing the validity of clustering

d) Formulating the clustering problem

SOLUTION: 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

SOLUTION: A

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

SOLUTION: 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

SOLUTION: 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

SOLUTION: A

8. Clustering is a-

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

SOLUTION: 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

SOLUTION: 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

SOLUTION: 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

SOLUTION: D

12. For clustering, we do not require-

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

SOLUTION: A

13. How is cluster analysis calculated?

ANSWER:

Cluster Analysis is calculated by 3 methods as follows:

1. K-means cluster.

2. Hierarchical cluster.

3. Two-step cluster.

K-means cluster :

It is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster :

It is the most common method. It generates a series of models with cluster solutions from *1* (all cases in one cluster) to *n* (each case is an individual cluster). Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.

The hierarchical cluster analysis follows three basic steps:
1) calculate the distances,
2) link the clusters, and
3) choose a solution by selecting the right number of clusters.

Two-step cluster :

This analysis identifies groupings by running pre-clustering first
and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal

14. How is cluster quality measured?

ANSWER:

The Quality of a Cluster is measured by few methods.
In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts. If ground truth is available, it can be used by extrinsic methods, which compare the clustering against the group truth and measure. If the ground truth is Unavailable, we can use intrinsic methods, which evaluate the goodness of a Clustering by considering how well the clusters are separated. Ground truth can be Considered as supervision in the form of "cluster labels." Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

Extrinsic Methods

When the ground truth is available, we can compare it with a clustering to assess the clustering. Thus, the core task in extrinsic methods is to assign a score, Q(C,Cg), to a clustering, C, given the ground truth, Cg. Whether an extrinsic method is effective largely depends on the measure, Q, it uses.

Intrinsic Methods

When the ground truth of a data set is not available, we have to use an intrinsic method to assess the clustering quality. In general, intrinsic methods evaluate a clustering by examining how well the clusters are separated and how compact the clusters are. Many intrinsic methods have the advantage of a similarity metric between objects in the data set.

15. What is cluster analysis and its types?

ANSWER:

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.

TYPES OF CLUSTER ANALYSIS

There are four basic types of cluster analysis used in data science.

i. Centroid Clustering,
ii. Density Clustering
iii. Distribution Clustering
iv. Connectivity Clustering.