MACHINE LEARNING

ASSIGNMENT – 1

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

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

Answer :b) 4

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

Answer :d) 1,2 and 4

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

Answer :d) formulating the clustering problem

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

Answer :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

Answer : b)divisive clustering

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

Answer : d)All answers are correct

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

Answer : d)All the above

8. Clustering is a- a) Supervised learning b) Unsupervised learning c) Reinforcement learning d) None

Answer : a)Supervised learning

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

Answer : a)K-means clustering

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

Answer : K-means clustering algorithm

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

Answer : D) All the above

12. For clustering, we do not require- a) Labeled data b) Unlabeled data c) Numerical data d) Categorical data

Answer a) Labeled Data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

. 13. How is cluster analysis calculated?

copy your data into the table, select more than one variable & select the number of cluster you want to calculate..

K means calculator :

initial partition with k clusters is created , randomly placing cluster focal point, assigning clusters nearby to the respective focal points, calculating new focal point from the separated cluster & reassigning according to the new focal points. steps being repeated until there is no change in cluster.

elbow curve to identify in identifying number of clusters (k),cluster taken when there is only small variation after a certain point.

14. How is cluster quality measured?

Cluster quality can be measured by below methods

Dissimilarity/Similarity metric : The similarity between the clusters can be expressed in terms of a distance function d(i,j)

Cluster Completeness : if two objects have same characteristics, they belong to same cluster category and if the objects are high then the cluster is of high quality...

Ragbag : when objects of some categories cannot be merged with other cluster then ragbag method is used... those dissimilar objects can be put in ragbag category\

Small cluster Preservation : small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are unique

15. What is cluster analysis and its types?

Cluster analysis is something to identify objects with different characteristics or attributes and cluster (grouping) them with the similarity of characteristics or attributes with different algorithms..

Types:

connectivity models: data points closer to each other has more similar characteristics data points can be classified into separate clusters and can be aggregated with decrease in distance or whole dataset classified into one cluster and partitioned into separate clusters when distance increases

Distribution Models: all data points in a cluster belonging to same distribution...i.e. Normal distribution or Gaussian distribution

Density Models: it searches data space for varied densities of data points and isolate the different density regions...it then assigns the data points within the same region as clusters..

Centroid Models : similarity between data points is derived based on their closeness to clusters centroid..example is k -means algorithm