

# Machine Learning - Worksheet - 1

By:

Yatika Taneja

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

Sol) (b) 4

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

Sol) (d) 1, 2 and 4.

③ The most important part of — is selecting the variables on which clustering is based?

Sol) (d) formulating the clustering problem

④ The most commonly used measure of similarity is the — or its square

Sol) (a) euclidean distance

(5) \_\_\_\_\_ is a clustering procedure where all objects start out as one giant cluster. Clusters are formed by dividing the cluster into smaller & smaller clusters.

Sol > (b) Divisive cluster

(6) Which of the following is required for k-means clustering?

Sol > all answers are correct

(7) The goal of clustering is to -

Sol > divide the data points into groups

(8) Clustering is a -

Sol > (b) unsupervised learning

(9) Which of the following clustering algorithms suffer from the problem of convergence at local optima?

Sol > a) ~~K-means algorithm~~

d) All of the above.



(10) Which version of clustering algorithm is most sensitive to outliers?

Sol > (A) K-means clustering algorithm

(11) What of the following is a bad characteristic of a dataset for clustering analysis?

Sol > (D) All of the above

(12) For clustering, we do not require -

Sol > (A) Labeled Data

(13) How is cluster analysis calculated?

sol) ① Hierarchical cluster generates a series of model with cluster solutions from 1 to  $n$ .

② It also works with variables as opposed to cases; it can cluster variables together in a manner similar to factor analysis.

(3) It follows 3 basic steps:

- calculate the distances
- link the clusters
- chose a solution by selecting the right number of clusters

(14) How is cluster quality measured?

sol) Cluster quality is measured as following:-

→ Davies-Bouldin Index

The DB Index is calculated by the



following formula:

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left( \frac{b_i + b_j}{d(c_i, c_j)} \right)$$

where  $n$  is number of clusters &  $b_i$  is average distance of cluster points  $p_i$  from cluster centroid  $c_i$

- As the DB widens shrinks, clustering is considered better!

→ Dunn Index

- The formula for Dunn Index is:

$$D = \frac{\min_{1 \leq i < j \leq n} d(i, j)}{\max_{1 \leq k \leq n} d'(k)}$$

where  $i, j, k$  are indices for clusters,  $d$  measures the inter-cluster distance &  $d'$  measure the intra-cluster difference

- While DB index considers dispersion & separation of all clusters, the Dunn index only considers the worst case in clustering: the clusters that are closest together & single most dispersed cluster

- Dunn index increases as performance improves

## → Silhouette coefficient

- It is measure as:

$$S(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}$$

where  $a(i)$  is average distance of  $i$  from all other points in its cluster &  $b(i)$  is smallest average distance of  $i$  to all points in any other cluster.

- It tells us how well-assigned each individual point is.

## ⑮ What is cluster analysis & its types!

- Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters.

- Cluster analysis involves formulating a problem, selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the resulting clusters & finally assessing the



validity of clustering

• Types of cluster analysis are as follows:

- 1- Hierarchical clustering
- 2- Partition clustering
- 3- Exclusive clustering
- 4- Overlapping clustering
- 5- Fuzzy clustering
- 6- Complete clustering -

## Machine Learning - worksheet - 2

① Movie recommendation systems are an example of!

sol > 2 and 3 (D)

② Sentiment analysis is a example of

sol > (e) 1, 2 and 4

③ Can decision trees be used to ~~for~~ perform clustering?

sol > (A) True -

④ Which of the following is the most ~~important~~ appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points?

sol > (A) 1 only



(6) What is the minimum no. of variables / features required to perform clustering?

Sol) (B) 1

(7) For 2 runs of K-means clustering is it expected to get same clustering results?

Sol) (B) No

(8) Is it possible that assignment of ~~obs~~ observations to cluster does not change between successive iterations in K-means?

Sol) (A) Yes

(9) Which of the following can act as a possible termination conditions in K-means?

Sol) (D) All of the above

(9)

which of the following can act as a possible termination conditions in k-means.

Sol

same question solution is as <sup>Q</sup> 8 since question is the same

(10)

which of the following <sup>algorithms</sup> ~~algorithms~~ is most sensitive to outliers?

Sol >

(A) K-means clustering algorithm

(11)

How can clustering be used to improve the accuracy of linear regression model

Sol >

(f) All of the above

(12)

What could be the possible reasons for producing two different dendrograms using agglomerative ~~as~~ clustering algorithms for the same dataset?

Sol >

(E) All of the above



(13) Is  $K^{\text{means}}$  sensitive to outliers?

sol) Yes  $K$  means clustering is sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

(14) Why is  $K$  means better?

- sol)
- Relatively simple to implement
  - Scales to large data sets
  - Guarantees convergence.
  - Can warm start the positions of centroid
  - Easily adapts to new examples
  - Generalizes to clusters of different shapes and sizes, such as elliptical clusters

Q13) Is K means a deterministic algorithm

- Sol) • K-means is a non-deterministic algorithm.
- The non-deterministic nature of K-means is due to its random selection of data points as ~~initial~~ initial centroids