Sachin Boora (DS2302)

https://github.com/sachinboora/Internship

Worksheet_set_1

Machine Learning

Ans 1 : 4 is the best choice for number of clusters

Ans 2 : In the following options, only Data point with round shapes can give better results as rest of them have limitations and will be failed to give better results. Hence Option (D) -- 1,2,4 is the right answer.

Ans 3 : formulating the clustering problem is the right answer

Ans 4 : Euclidean Distance

Ans 5 : Divisive clustering is the procedure where all objects starts in one giant cluster and then this cluster is divided into many smaller clusters. Basically, this is Top-Down Approach while Agglomerative clustering is opposite of Divisive as it is Bottom-Up approach where smaller clusters combines into one giant cluster.

Ans 6 : K-Means Clustering require all of the three options.
Ans 7 : Clustering is unsupervised learning which divides data points into different clusters based on some metrics like similarity or distance measure.

Ans 8 : Clustering is a unsupervised Learning

Ans 9 : Out of the given options , only K-Means Clustering suffers from the problem of convergence at local optima.

Ans 10 : K-Mean Clustering is most affected by Outliers as this algorithm is based on assumption that data points in each cluster are normally distributed around the centroid.

Ans 11 : All of the options in answer are bad characteristics of a dataset for clustering analysis.

Ans 12 : For clustering , we do not require labelling data.

Subjective Questions

Ques 13 : How is Cluster Analysis Calculated?
Ans 13 : Cluster Analysis is used in unsupervised machine learning to group similar data points together.
The process includes several steps including selection of variables and determination of numbers of clusters.

Different Steps included are :

1. Selection of Variables : First, variables are selected , it could be numerical or categorical.

2. Choice of distance measure : It is used to calculate similarity or dissimilarity between different data sets. Euclidean Distance is the most common distance measure used.
3. Choice of Clustering Algorithm : We have different Algorithms to choose from including K-Means Clustering, DBSCAN.
4. Numbers of Clusters : Depending on the Algorithm, we can choose number of clusters or Algorithm can also detect automatically.
5. Clustering : Here we can visualise data points clusters using plots and graphs.
6. Evaluation : Here Final result is calculated to determine how well our algorithm worked.

Ques 14 : How is cluster quality measured?

Ans 14 : We can measure cluster quality by various metrics. Some of them are :

1. Sum of Squared Error(SSE) : It is the sum of squared distance between each data point and centroid. The less the distance, the better the clustering.
2. Silhouette Score : Compares average distance between points within a cluster to average distance between points in different clusters. It ranges from -1 to 1. Higher score means better clustering.
3. Entropy & Purity : Entropy measures randomness of data points to clusters. Purity measures how well-defined clusters are.

Ques 15 : What is cluster analysis and its types?
Ans 15 : It is a statistical technique that involves grouping of data points in order to identify meaningful patterns and relationships in data. There are many type of cluster analysis.

1. Hierarchical clustering : Data points are grouped into hierarchy of nested clusters. Agglomerative and divisive are two common type of hierarchical clustering.
2. Partitional Clustering : Data points are divided into non-overlapping groups or clusters. Most common used partitional clustering is K-Means Clustering.
3. Other cluster analysis includes DB-SCAN.