Worksheet 1 Machine Learning

1. **B)** 4

   Can be measured by calculating the longest vertical line of the two horizontal lines

2. **D)** 1 , 2 and 4

   K-means clustering fails when data points has outliers , data points with different densities and data points with non convex shapes

3. **D)** formulating the clustering problem

4. **A)** Euclidean Distance

5. **B)** Divisive Clustering

6. **D)** All of the above

   In K means clustering defined distance metric , number of clusters , cluster of centroids all are very important.

7. **A)** To divide the data points into groups

8. **B)** Unsupervised learning

9. **A)** K means clustering algorithm

   It converges at local optima and that is also one of its biggest drawbacks

10. **A)** K means clustering

    It is most sensitive to outliers as it uses data clustering at mean positions

11. **D)** All of the above

12. **A)** Labeled data is not required as it depends on the nature of the data

13. Cluster analysis is calculated by following a few simple steps

    - Distance of the various data points in the data set is calculated; this tells about the data how scattered it is and how we are supposed to proceed with the analysis process.
    - Different clusters are then linked
    - Finally we come upon a solution using the right number of clusters and the right type of clusters for our analysis

14. Quality of the cluster can be measured by calculating the average coefficient value of the different clusters. This coefficient is also known as silhouette coefficient. This coefficient means that the average of the minima value and maxima value is calculated of the cluster for example -5 and +5 and then an average coefficient is calculated. What this implies is that if we get a positive value of this coefficient then the same is far away from the neighboring clusters.

15. Cluster analysis is a data mining technique our goal is to group these values together. The data here in these clusters is user defined data and our goal is to group this data on the basis of multiple user defined characteristics. This technique is very helpful and comes in handy in machine learning etc. There are a few Clustering analysis techniques ,these are as follows
    - Hierarchical clustering has two types -
        1. Agglomerative method - In this method the cluster is analyzed and then added to another cluster to which it is closest in characteristic and then a single big cluster is made. This is done till there is a single big cluster
        2. Divisive method- In this hierarchical type the data is single big data and is then further divided into smaller data types

    - Centroid based clustering - in this type of clustering there is a defined central entity which may or may not be part of the given data. K means clustering is an example of this technique

    - Distribution based clustering - Objects that belong to the same clustering distribution is put into the same cluster and others are not

    - Density based clustering - Clusters are defined by the area of density. The scare data points are not clustered and are considered as noise.