ASSIGNMENT – 2

# MACHINE LEARNING

**1. Ans – d) 2 and 3**

**2. Ans – d) 1,2 and 4**

**3. Ans – a) True**

**4. Ans – a) 1only**

**5. Ans – b) 1**

**6. Ans – b) No**

**7. Ans – a) Yes**

**8. Ans – d) All of the above**

**9. Ans – a) K-means clustering algorithm**

**10. Ans – d) all of the above**

**11. Ans – d) all of the above**

**12. Ans –**

The *K*-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. *K*-medoids clustering is a variant of *K*-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, *K* medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center.

**13. Ans –**

**Advantages of k-means**

1. Relatively simple to implement.

2. Scales to large data sets.
3. Guarantees convergence.
4. Can warm-start the positions of centroids.
5. Easily adapts to new examples.
6. Generalizes to clusters of different shapes and sizes, such as elliptical clusters

14. Ans –

K-Means is one of the most used algorithms for data clustering and the usual clustering method for benchmarking. Despite its wide application it is well-known that it suffers from a series of disadvantages; it is only able to find local minima and the positions of the initial clustering centres (centroids) can greatly affect the clustering solution. Over the years many K Means variations and initialisation techniques have been proposed with different degrees of complexity. In this study we focus on common K-Means variations along with a range of deterministic and stochastic initialisation techniques. We show that, on average, more sophisticated initialisation techniques alleviate the need for complex clustering methods.
Furthermore, deterministic methods perform better than stochastic methods. However, there is a trade-off: less sophisticated stochastic methods, executed multiple times, can result in better clustering. Factoring in execution time, deterministic methods can be competitive and result in a good clustering solution. These conclusions are obtained through extensive benchmarking using a range of synthetic model generators and real-world data sets.