ML Assignment 2

1. a) 2 Only
2. d) 1, 2 and 4
3. a) True
4. a) 1 only
5. b) 1
6. b) No
7. a) Yes
8. d) All of the above
9. a) K-means clustering algorithm
10. d) All of the above
11. d) All of the above
12. K-means clustering is an unsupervised learning algorithm which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest centroid. The algorithm aims to minimize the squared Euclidean distances between the observation and the centroid of cluster to which it belongs.Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

    But sometime K-Means algorithm does not give best results. It is sensitive to outliers. An outlier is a point which is different from the rest of data points.

    For e.g. Data set point are 1 2 3 7 8 80
    Now 80 is outlier.
    K=2
    Cluster 1=1 Cluster 2=7
    After first iteration
    C1=2 C2=31.67
    As 80 data point which is an outlier come in cluster 2.
    Cluster 2 centroid changes to accommodate 80 .
    Therefore K means is sensitive to outliers

    Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored. Consider removing or clipping outliers before clustering. In K-Means clustering outliers are found by distance based approach and cluster based approach. In case of hierarchical clustering, by using dendrogram outliers are found.

13. k-means is one of the simplest algorithms which uses an unsupervised learning method to solve known clustering issues. It works really well with large datasets It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties.It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters..Other clustering algorithms with better features tend to be harder to implement and more expensive. In this case, k-means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied.

14. The basic k-means clustering is based on a non-deterministic algorithm.This means that running the algorithm several times on the same data, could give different results.This means that a compiler cannot solve the problem in polynomial time and doesn't clearly know the next step. This is because some problems have a great degree of randomness to them. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids. The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids.