Assignment-2 workset_2

Q.1) Option-B

Movie Recommendation systems are an example of clustering & classification.

Q.2) Option-D

Sentiment Analysis is an example of regression,classfication and reinforcement learning

Q.3) Option- A

Decision trees can also be used to for clusters in the data but clustering often generates natural clusters and is not dependent on any objective function.

Q.4) Option-A

Removal of outliers is not recommended if the data points are few in number. In this scenario, capping and flouring of variables is the most appropriate strategy.

Q.5) Option-B

the minimum no. of variables/ features required to perform clustering is 1.

Q.6) Option-B

Q.7) Option-A

When the K-Means algorithm has reached the local or global minima, it will not alter the assignment of data points to clusters for two successive iterations.

Q.8) Option-D

All four conditions can be used as possible termination condition in K-Means clustering:

Q.9) Option-A

K-means clustering is most sensative to outliers.

Q10) Option-D

all of the above

Q11)Option-D

all of the above

Q12) Kmeans sensative to outliers-

It's sensitive to outliers. Algorithm is sensitive to outliers, since a single mislabeled example dramatically changes the class boundaries. Anomalies affect the method significantly, because k-NN gets all the information from the input, rather than from an algorithm that tries to generalize data.

Q13) K-means is better-

Advantages of k-means-

It is relatively simple to implement.

It Scales to large data sets.

It Guarantees convergence.

It Can warm-start the positions of centroids.

It Easily adapts to new examples.

It Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

Q14) Kmeans is a non deterministic algorithm.

Kmeans clustering ustering algorithms with steps involving randomness usually give different results on different executions for the same dataset. This non-deterministic nature of algorithms such as the K-Means clustering algorithm limits their applicability in areas such as cancer subtype prediction using gene expression data. It is hard to sensibly compare the results of such algorithms with those of other algorithms. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids.