MACHINE LEARNING

(WORKSHEET-2 ANSWERS)

1. Movie Recommendation systems are an example of:

i) Classification

ii) Clustering

iii) Regression

ANSWER-) b) 1 and 2

2. Sentiment Analysis is an example of:

i) Regression

ii) Classification

iii) Clustering

iv) Reinforcement

ANSWER-) d) 1, 2 and 4

3.Can decision trees be used for performing clustering?

a) True

b) False

ANSWER-) a) True

4.Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:

i) Capping and flooring of variables
ii) Removal of outliers

ANSWER-) a) 1 only

5. What is the minimum no. of variables/ features required to perform clustering?

ANSWER-) b) 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?

ANSWER-) b) No

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means?

ANSWER-) a) Yes

8. Which of the following can act as possible termination conditions in K-Means?

i) For a fixed number of iterations.

ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad local minimum.

iii) Centroids do not change between successive iterations.

iv) Terminate when RSS falls below a threshold.

ANSWER-) d) All of the above

9. Which of the following algorithms is most sensitive to outliers?

a) K-means clustering algorithm

b) K-medians clustering algorithm

c) K-modes clustering algorithm

d) K-medoids clustering algorithm

ANSWER-) a) K-means clustering algorithm

10. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

i) Creating different models for different cluster groups.

ii) Creating an input feature for cluster ids as an ordinal variable.

iii) Creating an input feature for cluster centroids as a continuous variable.

v) Creating an input feature for cluster size as a continuous variable.

ANSWER-) d) All of the above

11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

a) Proximity function used

b) of data points used

c) of variables used

d) All of the above

ANSWER-) d) All of the above

12. Is K sensitive to outliers?

ANSWER-) The k-mean algorithm is sensitive to the outliers, because a mean is easily influenced by extreme values.We propose a robust two-stage k-means clustering algorithm based on the observation point mechanism, which can accurately discover the cluster centers without the disturbance of outliers.

13. Why is K means better?

ANSWER-) Advantages of k-means

Guarantees convergence. Can warm-start the positions of centroids. Easily adapts to new examples. Generalizes to clusters of different shapes and sizes, such as elliptical clusters.

14. Is K means a deterministic algorithm ?

ANSWER-) 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.