## Introduction to Machine Learning -IITKGP

## Assignment - 8

## TYPE OF QUESTION: MCQ/MSQ

Number of questions: 15 Total mark: 2 * 15 = 30

1. What is true about K-Mean Clustering?

   1. K-means is extremely sensitive to cluster center initializations

   2. Bad initialization can lead to Poor convergence speed

   3. Bad initialization can lead to bad overall clustering

   a. 1 and 2

   b. 1 and 3

   c. All of the above

   d. 2 and 3

   **Correct Answer**: c

   **Detailed Solution**: All three of the given statements are true. K-means is extremely sensitive to cluster center initialization. Also, bad initialization can lead to Poor convergence speed as well as bad overall clustering.

---

2. In which of the following cases will K-Means clustering fail to give good results? (Mark all that apply)

   a. Data points with outliers

   b. Data points with round shapes

   c. Data points with non-convex shapes

   d. Data points with different densities

   **Correct Answer**: a, c, d

   **Detailed Solution**: K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.

---

3. Which of the following clustering algorithms suffers from the problem of convergence at local optima? (Mark all that apply)

   a. K- Means clustering algorithm

   b. Agglomerative clustering algorithm

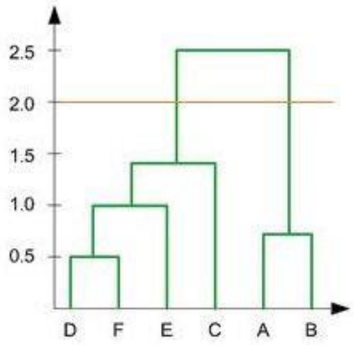
   c. Expectation-Maximization clustering algorithm

   d. Diverse clustering algorithm

**Correct Answer**: a, c

**Detailed Solution**: Out of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

4. In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?



a. 1

b. 2

c. 3

d. 4

**Correct Answer**: b

**Detailed solution**: Since the number of vertical lines intersecting the red horizontal line at

y=2 in the dendrogram are 2, therefore, two clusters will be formed.

5. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration the clusters: C1, C2, C3 has the following observations:
C1: {(1,1), (4,4), (7,7)}
C2: {(0,4), (4,0)}
C3: {(5,5), (9,9)}
What will be the cluster centroids if you want to proceed for second iteration?

a. C1: (4,4), C2: (2,2), C3: (7,7)
b. C1: (2,2), C2: (0,0), C3: (5,5)
c. C1: (6,6), C2: (4,4), C3: (9,9)
d. None of these

**Correct Answer**: a

**Detailed Solution**:

Finding centroid for data points in cluster C1 = ((2+4+6)/3, (2+4+6)/3) = (4, 4)
Finding centroid for data points in cluster C2 = ((0+4)/2, (4+0)/2) = (2, 2)
Finding centroid for data points in cluster C3 = ((5+9)/2, (5+9)/2) = (7, 7)
Hence, C1: (4,4), C2: (2,2), C3: (7,7)

---

6. Following Question 5, what will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in the second iteration?

a. 10
b. 5
c. 6
d. 7

**Correct Answer**: a

**Detailed Solution**: Manhattan distance between centroid C1 i.e. (4, 4) and (9, 9) = (9-4) + (9-4) = 10.

---

7. Which of the following is not a clustering approach?

a. Hierarchical
b. Partitioning
c. Bagging
d. Density-Based

**Correct Answer**: c

**Detailed Solution**: Follow lecture slides.

---

8. Which one of the following is correct?

    a. Complete linkage clustering is computationally cheaper compared to single linkage.
    b. Single linkage clustering is computationally cheaper compared to K-means clustering.
    c. K-Means clustering is computationally cheaper compared to single linkage clustering.
    d. None of the above.

**Correct Answer**: c

**Detailed Solution**: K-Means clustering is generally computationally more efficient than single linkage hierarchical clustering. In K-Means, the main computational cost involves iterating over data points and updating cluster assignments and centroids until convergence, which typically converges in a reasonable number of iterations. In contrast, single linkage hierarchical clustering involves pairwise distance computations for all data points at each step of the hierarchy, making it computationally more expensive, especially for large datasets.

---

9. Considering single-link and complete-link hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster? If so, in which approach will this tend to be observed?
   a. No
   b. Yes, single-link clustering
   c. Yes, complete-link clustering
   d. Yes, both single-link and complete-link clustering.

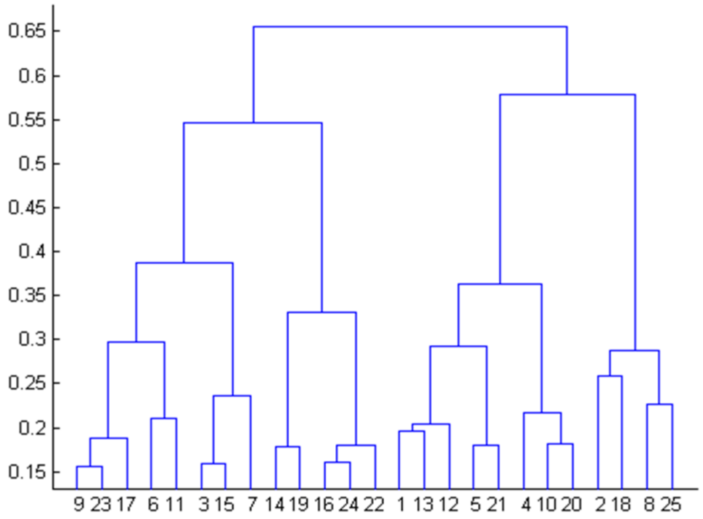
**Correct Answer**: d

**Detailed Solution**: In single-link hierarchical clustering, it is possible for a point to be closer to points in other clusters than to points in its own cluster. This can lead to the phenomenon known as "chaining," where clusters are stretched out because the similarity between two clusters is determined by the closest pair of data points, which can sometimes result in points from different clusters being closer to each other than to points within their own clusters.

In complete-link hierarchical clustering, it is also possible for a point to be closer to points in other clusters than to points in its own cluster. This can lead to clusters being tightly packed together and is sometimes referred to as the "crowding" problem.

---

10. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusions can be drawn from the dendrogram?



a. There were 28 data points in the clustering analysis
b. The best number of clusters for the analyzed data points is 4
c. The proximity function used is Average-link clustering
d. The above dendrogram interpretation is not possible for K-Means clustering analysis

**Correct Answer**: d
**Detailed Solution**:
A dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

---

11. Feature scaling is an important step before applying K-Mean algorithm. What is the reason behind this?
    a. In distance calculation it will give the same weights for all features
    b. You always get the same clusters if you use or don't use feature scaling
    c. In Manhattan distance it is an important step but in Euclidean it is not
    d. None of these

**Correct Answer:** a
**Detailed Solution:**
Feature scaling ensures that all the features get the same weight in the clustering analysis. Consider a scenario of clustering people based on their weights (in KG) with a range 55-110 and height (in inches) with a range 5.6 to 6.4. In this case, the clusters produced without scaling can be very misleading as the range of weight is much higher than that of height. Therefore, its necessary to bring them to same scale so that they have equal weightage on the clustering result.

---

12. Which of the following options is a measure of internal evaluation of a clustering algorithm?
a. Rand Index
b. Jaccard Index
c. Davis-Bouldin Index
d. F-score

**Correct Answer**: c
**Detailed Solution**: Follow lecture slides.

---

13. Given, A= {0,1,2,5,6} and B = {0,2,3,4,5,7,9}, calculate Jaccard Index of these two sets.
    a. 0.50
    b. 0.25
    c. 0.33
    d. 0.41

**Correct Answer:** c

**Detailed Solution:**

To calculate the Jaccard Index for two sets A and B, you need to find the intersection and union of the sets and then divide the size of the intersection by the size of the union.

Let's denote the sets as Set A and Set B:

Set A: {0, 1, 2, 5, 6}

Set B: {0, 2, 3, 4, 5, 7, 9}

Intersection (the elements that are common to both sets): {0, 2, 5}

Union (all unique elements from both sets): {0, 1, 2, 3, 4, 5, 6, 7, 9}

Now, calculate the Jaccard Index:

Jaccard Index = (Size of Intersection) / (Size of Union)

Jaccard Index = (3) / (9) = 1/3 ≈ 0.33

So, the Jaccard Index for these two sets is approximately 0.33.

The correct answer is c. 0.33.

---

14. Suppose you run K-means clustering algorithm on a given dataset. What are the factors on which the final clusters depend?
    I. The value of K
    II. The initial cluster seeds chosen
    III. The distance function used.

    a. I only
    b. II only
    c. I and II only
    d. I, II and III

**Correct Answer**: d

**Detailed Solution:**

The final clusters in the K-means clustering algorithm depend on the following factors:

I. The value of K: The number of clusters (K) is a crucial parameter in K-means, and it significantly affects the final clustering. Different values of K can lead to different clusterings.

II. The initial cluster seeds chosen: The initial placement of cluster centroids or seeds can impact the convergence of the algorithm and the resulting clusters. Different initializations may lead to different final cluster assignments.

III. The distance function used: The choice of distance metric (e.g., Euclidean distance, Manhattan distance, etc.) influences how the algorithm measures the similarity or dissimilarity between data points. Different distance functions can lead to different cluster shapes and assignments.

So, all three factors (I, II, and III) play a role in determining the final clusters in K-means.

The correct answer is d. I, II, and III.

---

15. Consider a training dataset with two numerical features namely, height of a person and age of the person. The height varies from 4-8 and age varies from 1-100. We wish to perform K-Means clustering on the dataset. Which of the following options is correct?

    a. We should use Feature-scaling for K-Means Algorithm.
    b. Feature Scaling can not be used for K-Means Algorithm.
    c. You always get the same clusters if you use or don't use feature scaling.
    d. None of these

**Correct Answer:** a

**Detailed Solution**: In K-Means clustering, the scale of features can affect the clustering results. When features have different scales, K-Means tends to give more weight to features with larger scales. In the given scenario, the "height" feature has a range of 4-8, while the "age" feature has a range of 1-100. Because of this significant difference in scales, it's advisable to use feature scaling to bring both features to a similar scale. Standardization (subtracting the mean and dividing by the standard deviation) or min-max scaling (scaling features to a specific range, like [0, 1]) are common methods for feature scaling in K-Means.

Option b is incorrect because feature scaling can be used for K-Means, and it's often recommended when dealing with features of different scales.

Option c is also incorrect. You do not always get the same clusters if you use or don't use feature scaling. The initial centroids and the clustering results can be influenced by the scale of the features, so feature scaling can lead to different clusters compared to not using it.

So, the correct answer is **a**. We should use Feature-scaling for K-Means Algorithm.

______________________________________________________________________________

**************END**************