1. **Group of a like data points is known as**
   * a) Clusters
   * b) Simple points
   * c) Group points
   * d) None

**Answer - a) Clusters**

1. **From the below find the data types on which we can perform clustering**
   * a) Graphs
   * b) Time-series data
   * c) Multimedia data
   * d) All the above

**Answer - d) All the above**

1. **Agglomerative clustering is an example of\_\_\_\_\_\_\_\_\_\_\_\_ .**
   * a) Hierarchical
   * b) Distance-based clustering
   * c) Both a and b
   * d) K means

**Answer - c) Both a and b**

1. **Clustering analysis is unsupervised learning since it does not require \_\_\_\_\_\_ training data.**
   * a) Unlabeled
   * b) Labeled
   * c) Both
   * d) None of the above

**Answer - b) Labeled**

1. **A tree diagram used to show the arrangement of clusters in hierarchical clustering is known as\_\_\_\_\_\_\_.**
   * a) Bar plot diagram
   * b) Histogram
   * c) Dendrogram
   * d) None

**Answer - c) Dendrogram**

1. **From the below, select the types of hierarchical clustering**
   * a) Divisive
   * b) Agglomerative
   * c) K means
   * d) Both a and b

**Answer - d) Both a and b**

1. **Visualization of clustering results will be done only when we have two-dimensional data.**
   * a) True
   * b) False

**Answer - b) False**

1. **We should know the no. of output clusters is a prior for all clustering algorithms.**
   * a) True
   * b) False

**Answer - b) False**

1. **\_\_\_\_\_\_\_\_ is used most frequently to measure similarity**
   * a) Manhattan Distance
   * b) Euclidean distance
   * c) Far by distance
   * d) None of the above

**Answer - b) Euclidean distance**

1. **\_\_\_\_\_\_\_stores the distances between each point.**
   * a) Proximity matrix
   * b) Primary matrix
   * c) Minor data matrix
   * d) Proximity sensor

**Answer - a) Proximity matrix**

1. **Agglomerative Clustering is the \_\_\_\_\_\_\_\_\_\_\_\_.**
   * a) Top-down approach
   * b) Bottom-up approach
   * c) Side to side approach
   * d) Center to end approach

**Answer - b) Bottom up approach**

1. **Divisive Clustering is the \_\_\_\_\_\_\_\_\_\_\_\_**
   * a) Top-down approach
   * b) Bottom-up approach
   * c) Side to side approach
   * d) Center to end approach

**Answer - a) Top-down approach**

1. **From the below find the methods used to measure the similarity**
   * a) Single linkage
   * b) Complete linkage
   * c) Average linkage
   * d) All the above

**Answer - d) All the above**

1. **Hierarchical clustering is also used for \_\_\_\_\_\_\_\_\_\_.**
   * a) Finding na
   * b) Outlier detection
   * c) Fake value detection
   * d) All the above

**Answer - b) Outlier Detection**

## 3. What are the various types of Hierarchical Clustering?

The two different types of Hierarchical Clustering technique are as follows:

**Agglomerative:** It is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

**Divisive:** It is just the opposite of the agglomerative algorithm as it is a top-down approach.

## 4. Explain the Agglomerative Hierarchical Clustering algorithm with the help of an example.

**Step- 1:**In the first step, we compute the proximity of individual observations and consider all the six observations as individual clusters.

**Step- 2:**In this step, similar clusters are merged together and result in a single cluster.

For our example, we consider B, C, and D, E are similar clusters that are merged in this step. Now, we are remaining with four clusters named A, BC, DE, F.

**Step- 3:** We again compute the proximity of new clusters and merge the similar clusters to form new clusters A, BC, DEF.

**Step- 4:** Again, compute the proximity of the newly formed clusters.

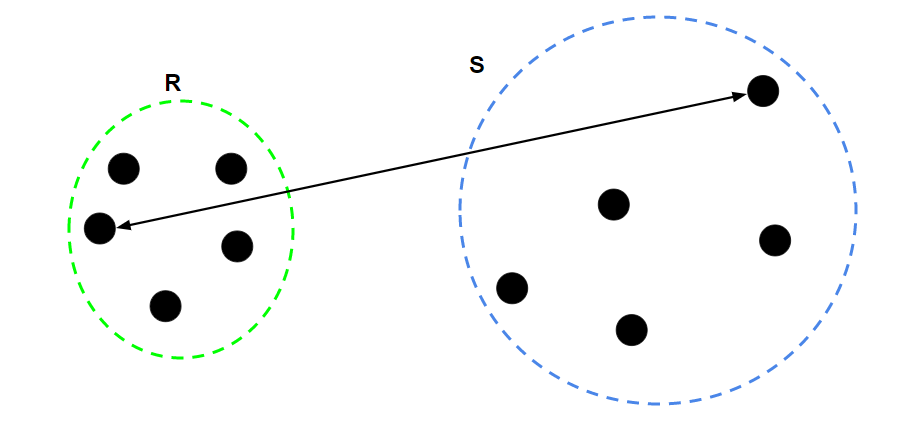
## Explain the different linkage methods used in the Hierarchical Clustering Algorithm.

**Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j

A picture containing diagram

Description automatically generated

**Complete Linkage:** For two clusters R and S, the complete linkage returns the maximum distance between two points i and j



**Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated.

**Centroid-linkage:**In this method, we find the centroid of cluster 1 and the centroid of cluster 2 and then calculate the distance between the two before merging.

**Pros of Single-linkage:**

* This approach can differentiate between non-elliptical shapes as long as the gap between the two clusters is not small.

**Cons of Single-linkage:**

* This approach cannot separate clusters properly if there is noise between clusters.

**Pros of Complete-linkage:**

* This approach gives well-separating clusters if there is some kind of noise present between clusters.

**Cons of Complete-Linkage:**

* This approach is biased towards globular clusters.( A **globular cluster** is a spherical collection of stars. **Globular clusters** are very tightly bound by gravity, with a high concentration of stars)
* It tends to break large clusters.

**Ward’s method** (a.k.a. *Minimum variance method* or *Ward’s Minimum Variance Clustering Method*) is an alternative to single-link [clustering](https://www.statisticshowto.com/clustering/). Popular in fields like linguistics, it’s liked because it usually creates compact, even-sized clusters (Szmrecsanyi, 2012).

Like most other clustering methods, Ward’s method is computationally intensive. However, Ward’s has significantly fewer computations than other methods. **The drawback is this usually results in less than optimal clusters.**

Like other clustering methods, Ward’s method starts with n clusters, each containing a single object. These n clusters are combined to make one cluster containing all objects. At each step, **the process makes a new cluster that minimizes**[**variance**](https://www.statisticshowto.com/probability-and-statistics/variance/)**, measured by an**[**index**](https://www.calculushowto.com/index-number/)**called E**

## **What is a dendrogram in Hierarchical Clustering Algorithm?**

A dendrogram is a tree-like structure that explains the relationship between all the data points in the system.

## **What is Space and Time Complexity of the Hierarchical Clustering Algorithm?**

**Space complexity:** Hierarchical Clustering Technique requires very high space when the number of observations in our dataset is more since we need to store the similarity matrix in the RAM. So, the space complexity is the order of the square of n.

**Space complexity = O(n²) where n is the number of observations.**

**Time complexity:**Since we have to perform n iterations and in each iteration, we need to update the proximity matrix and also restore that matrix, therefore the time complexity is also very high. So, the time complexity is the order of the cube of n.

**Time complexity = O(n³) where n is the number of observations.**

## **List down some of the possible conditions for producing two different dendrograms using an agglomerative Clustering algorithm with the same dataset.**

The given situation happens due to either of the following reasons:

* Change in Proximity function
* Change in the number of data points or variables.

## **How to Find the Optimal Number of Clusters in Agglomerative Clustering Algorithm?**

To find the optimal number of clusters, Silhouette Score is considered to be one of the popular approaches.

In **K-means clustering**, **elbow method** and **silhouette analysis or score** techniques are used to find the number of clusters in a dataset. The elbow method is used to find the “elbow” point, where adding additional data samples does not change cluster membership much. Silhouette score determines whether there are large gaps between each sample and all other samples within the same cluster or across different clusters.

## **How can we measure the goodness of the clusters in the Hierarchical Clustering Algorithm?**

There are many measures to find the goodness of clusters but the most popular one is the **Dunn’s Index.**

Dunn’s index is defined as the ratio of the minimum inter-cluster distances to the maximum intra-cluster diameter and the diameter of a cluster is calculated as the distance between its two furthermost points i.e, maximum distance from each other.

**What is significance testing in clustering?**

**Significance testing** is performed to distinguish between a clustering that reflects meaningful *heterogeneity* in the data and an artificial clustering of *homogeneous* data.

**What is Jaccard index?**

The **Jaccard index**, also known as the Jaccard similarity coefficient, is a statistic used for gauging the similarity and diversity of sample sets. The Jaccard coefficient measures **similarity** between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

**Difference between Gaussian mixture and K-means ?**

Gaussian mixtures seem to be more robust. However, GMs usually tend to be slower than K-Means because it takes more iterations of the EM algorithm to reach the convergence.

**What are advantages of GMM over k-means?**

Gaussian mixture modelling clustering data points have become simpler as they can handle even oblong clusters. It works in the same principle as K-means but has some of the advantages over it. It tells us about which data belongs to which cluster along with the probabilities. In other words, it performs hard classification while K-Means perform soft classification.

**when using various clustering algorithms why does Euclidean distance technique is not suitable for multidimensional dataset**

Euclidean distance is not a good distance in high dimensions'. I guess this statement has something to do with the curse of dimensionality, but what exactly? Besides, what is 'high dimensions'? I have been applying hierarchical clustering using Euclidean distance with 100 features. Up to how many features is it 'safe' to use this metric?

**Can decision trees be used for performing clustering?**

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.

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

At least a single variable is required to perform clustering analysis. Clustering analysis with a single variable can be visualized with the help of a histogram.

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

K-Means clustering algorithm instead converses on local minima which might also correspond to the global minima in some cases but not always. Therefore, it’s advised to run the K-Means algorithm multiple times before drawing inferences about the clusters.

**Why it is necessary to scale the variables?**

While calculating Euclidean distance large scale variable variable will suppressed by small scale hence necessary to scale at same level.

Eg.age and salary change in age by 5 is considerable change but change is 5k in salary is negligible change.

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

1. **For a fixed number of iterations.**
2. **Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.**
3. **Centroids do not change between successive iterations.**
4. **Terminate when RSS falls below a threshold.**

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

1. This condition limits the runtime of the clustering algorithm, but in some cases the quality of the clustering will be poor because of an insufficient number of iterations.
2. Except for cases with a bad local minimum, this produces a good clustering, but runtimes may be unacceptably long.
3. This also ensures that the algorithm has converged at the minima.
4. Terminate when RSS falls below a threshold. This criterion ensures that the clustering is of a desired quality after termination. Practically, it’s a good practice to combine it with a bound on the number of iterations to guarantee termination.

**Q9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?**

1. **K- Means clustering algorithm**
2. **Agglomerative clustering algorithm**
3. **Expectation-Maximization clustering algorithm**
4. **Diverse clustering algorithm**

Options:

A. 1 only

B. 2 and 3

C. 2 and 4

D. 1 and 3

E. 1,2 and 4

F. All of the above

**Solution: (D)**

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

**Q10. Which of the following algorithm 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

**Solution: (A)**

Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

**Q11. After performing K-Means Clustering analysis on a dataset, you observed the following dendrogram. Which of the following conclusion can be drawn from the dendrogram?**

Chart, box and whisker chart

Description automatically generated

A. There were 28 data points in clustering analysis

B. The best no. 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

**Solution: (D)**

A dendrogram is not possible for K-Means clustering analysis. However, one can create a cluster gram based on K-Means clustering analysis.

**What will be accuracy of k-means clustering?**

Since k means clustering is unsupervised learning,it does not have accuracy, however you can have goodness of fit approach which can be dunn’s index or **silhouette analysis or score** 

**Which measures points in the clusters must be closer and distance between the cluster must be higher.**

**Why the elbow curve is called elbow curve only?**

**Ref the video in 8-ibm by aman**

**Q12. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):**

1. **Creating different models for different cluster groups.**
2. **Creating an input feature for cluster ids as an ordinal variable.**
3. **Creating an input feature for cluster centroids as a continuous variable.**
4. **Creating an input feature for cluster size as a continuous variable.**

Options:

A. 1 only

B. 1 and 2

C. 1 and 4

D. 3 only

E. 2 and 4

F. All of the above

**Solution: (F)**

Creating an input feature for cluster ids as ordinal variable or creating an input feature for cluster centroids as a continuous variable might not convey any relevant information to the regression model for multidimensional data. But for clustering in a single dimension, all of the given methods are expected to convey meaningful information to the regression model. For example, to cluster people in two groups based on their hair length, storing clustering ID as ordinal variable and cluster centroids as continuous variables will convey meaningful information.

**Q13. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithm for the same dataset?**

A. Proximity function used

B. of data points used

C. of variables used

D. B and c only

E. All of the above

**Solution: (E)**

Change in either of Proximity function, no. of data points or no. of variables will lead to different clustering results and hence different dendrograms.

**Q14. In the figure below, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?**

Chart, histogram, box and whisker chart

Description automatically generated

A. 1

B. 2

C. 3

D. 4

**Solution: (B)**

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.

**Q15. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:**

Chart, box and whisker chart

Description automatically generated

A. 2

B. 4

C. 6

D. 8

**Solution: (B)**

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

Chart, box and whisker chart

Description automatically generated

In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

**Q16. In which of the following cases will K-Means clustering fail to give good results?**

1. **Data points with outliers**
2. **Data points with different densities**
3. **Data points with round shapes**
4. **Data points with non-convex shapes**

Options:

A. 1 and 2

B. 2 and 3

C. 2 and 4

D. 1, 2 and 4

E. 1, 2, 3 and 4

**Solution: (D)**

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.

Chart, scatter chart

Description automatically generated

**Q17. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?**

1. **Single-link**
2. **Complete-link**
3. **Average-link**

Options:

A. 1 and 2

B. 1 and 3

C. 2 and 3

D. 1, 2 and 3

**Solution: (D)**

All of the three methods i.e. single link, complete link and average link can be used for finding dissimilarity between two clusters in hierarchical clustering.

**Q18. Which of the following are true?**

1. **Clustering analysis is negatively affected by multicollinearity of features**
2. **Clustering analysis is negatively affected by heteroscedasticity**

Options:

A. 1 only

B. 2 only

C. 1 and 2

D. None of them

**Solution: (A)**

Clustering analysis is not negatively affected by heteroscedasticity but the results are negatively impacted by multicollinearity of features/ variables used in clustering as the correlated feature/ variable will carry extra weight on the distance calculation than desired.

**Q19. Given, six points with the following attributes:**

**A picture containing text, receipt

Description automatically generated**

**Which of the following clustering representations and dendrogram depicts the use of MIN or Single link proximity function in hierarchical clustering:**

A. Diagram

Description automatically generated

B. Diagram

Description automatically generated with low confidence

C. Diagram

Description automatically generated

D. Diagram

Description automatically generated

**Solution: (A)**

For the single link or MIN version of hierarchical clustering, the proximity of two clusters is defined to be the minimum of the distance between any two points in the different clusters. For instance, from the table, we see that the distance between points 3 and 6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. As another example, the distance between clusters {3, 6} and {2, 5} is given by dist({3, 6}, {2, 5}) = min(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = min(0.1483, 0.2540, 0.2843, 0.3921) = 0.1483.

**Q20 Given, six points with the following attributes:**

**A picture containing text, receipt

Description automatically generated**

**Which of the following clustering representations and dendrogram depicts the use of MAX or Complete link proximity function in hierarchical clustering:**

A.   Diagram

Description automatically generated

B.  Diagram

Description automatically generated with low confidence

C.  Diagram

Description automatically generated

D.  Diagram

Description automatically generated

**Solution: (B)**

For the single link or MAX version of hierarchical clustering, the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters. Similarly, here points 3 and 6 are merged first. However, {3, 6} is merged with {4}, instead of {2, 5}. This is because the dist({3, 6}, {4}) = max(dist(3, 4), dist(6, 4)) = max(0.1513, 0.2216) = 0.2216, which is smaller than dist({3, 6}, {2, 5}) = max(dist(3, 2), dist(6, 2), dist(3, 5), dist(6, 5)) = max(0.1483, 0.2540, 0.2843, 0.3921) = 0.3921 and dist({3, 6}, {1}) = max(dist(3, 1), dist(6, 1)) = max(0.2218, 0.2347) = 0.2347.

**Q21 Given, six points with the following attributes:**

**A picture containing text, receipt

Description automatically generated**

**Which of the following clustering representations and dendrogram depicts the use of Group average proximity function in hierarchical clustering:**

A.  Diagram

Description automatically generated

B.  Diagram

Description automatically generated with low confidence  
C. Diagram

Description automatically generated

D. Diagram

Description automatically generated

**Solution: (C)**

For the group average version of hierarchical clustering, the proximity of two clusters is defined to be the average of the pairwise proximities between all pairs of points in the different clusters. This is an intermediate approach between MIN and MAX. This is expressed by the following equation:

A picture containing text, screenshot

Description automatically generated

Here, the distance between some clusters. dist({3, 6, 4}, {1}) = (0.2218 + 0.3688 + 0.2347)/(3 ∗ 1) = 0.2751. dist({2, 5}, {1}) = (0.2357 + 0.3421)/(2 ∗ 1) = 0.2889. dist({3, 6, 4}, {2, 5}) = (0.1483 + 0.2843 + 0.2540 + 0.3921 + 0.2042 + 0.2932)/(6∗1) = 0.2637. Because dist({3, 6, 4}, {2, 5}) is smaller than dist({3, 6, 4}, {1}) and dist({2, 5}, {1}), these two clusters are merged at the fourth stage

**Q22. Given, six points with the following attributes:**

**A picture containing text, receipt

Description automatically generated**

**Which of the following clustering representations and dendrogram depicts the use of Ward’s method proximity function in hierarchical clustering:**

A. Diagram

Description automatically generated

B. Diagram

Description automatically generated with low confidence

C.  Diagram

Description automatically generated

D. Diagram

Description automatically generated

**Solution: (D)**

Ward method is a centroid method. Centroid method calculates the proximity between two clusters by calculating the distance between the centroids of clusters. For Ward’s method, the proximity between two clusters is defined as the increase in the squared error that results when two clusters are merged. The results of applying Ward’s method to the sample data set of six points. The resulting clustering is somewhat different from those produced by MIN, MAX, and group average.

**Q23. What should be the best choice of no. of clusters based on the following results:**

Chart, line chart

Description automatically generated

A. 1

B. 2

C. 3

D. 4

**Solution: (C)**

The silhouette coefficient is a measure of how similar an object is to its own cluster compared to other clusters. Number of clusters for which silhouette coefficient is highest represents the best choice of the number of clusters.

**Q24. Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?**

A. Imputation with mean

B. Nearest Neighbor assignment

C. Imputation with Expectation Maximization algorithm

D. All of the above

**Solution: (C)**

All of the mentioned techniques are valid for treating missing values before clustering analysis but only imputation with EM algorithm is iterative in its functioning.

**Q25. K-Mean algorithm  has some limitations. One of the limitation it has is, it makes hard assignments(A point either completely belongs to a cluster or not belongs at all) of points to clusters.**

**Note: Soft assignment can be consider as the probability of being assigned to each cluster: say K = 3 and for some point xn, p1 = 0.7, p2 = 0.2, p3 = 0.1)**

**Which of the following algorithm(s) allows soft assignments?**

1. **Gaussian mixture models**
2. **Fuzzy K-means**

Options:

A. 1 only

B. 2 only

C. 1 and 2

D. None of these

**Solution: (C)**

Both, Gaussian mixture models and Fuzzy K-means allows soft assignments.

**Q26. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:**

**C1: {(2,2), (4,4), (6,6)}**

**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: (6,6), C2: (4,4), C3: (9,9)

C. C1: (2,2), C2: (0,0), C3: (5,5)

D. None of these

**Solution: (A)**

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)

**Q27. Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations:**

**C1: {(2,2), (4,4), (6,6)}**

**C2: {(0,4), (4,0)}**

**C3: {(5,5), (9,9)}**

**What will be the Manhattan distance for observation (9, 9) from cluster centroid C1. In second iteration.**

A. 10

B. 5\*sqrt(2)

C. 13\*sqrt(2)

D. None of these

**Solution: (A)**

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

**Q28. If two variables V1 and V2, are used for clustering. Which of the following are true for K means clustering with k =3?**

1. **If V1 and V2 has a correlation of 1, the cluster centroids will be in a straight line**
2. **If V1 and V2 has a correlation of 0, the cluster centroids will be in straight line**

Options:

A. 1 only

B. 2 only

C. 1 and 2

D. None of the above

**Solution: (A)**

If the correlation between the variables V1 and V2 is 1, then all the data points will be in a straight line. Hence, all the three cluster centroids will form a straight line as well.

**Q29. Feature scaling is an important step before applying K-Mean algorithm. What is 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 Euclidian it is not

D. None of these

**Solution; (A)**

Feature scaling ensures that all the features get same weight in the clustering analysis. Consider a scenario of clustering people based on their weights (in KG) with range 55-110 and height (in inches) with 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.

**Q30. Which of the following method is used for finding optimal of cluster in K-Mean algorithm?**

A. Elbow method

B. Manhattan method

C. Ecludian mehthod

D. All of the above

E. None of these

**Solution: (A)**

Out of the given options, only elbow method is used  for finding the optimal number of clusters. The elbow method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn’t give much better modeling of the data.

**Q31. 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**

Options:

A. 1 and 3

B. 1 and 2

C. 2 and 3

D. 1, 2 and 3

**Solution: (D)**

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.

**Q32. Which of the following can be applied to get good results for K-means algorithm corresponding to global minima?**

1. **Try to run algorithm for different centroid initialization**
2. **Adjust number of iterations**
3. **Find out the optimal number of clusters**

Options:

A. 2 and 3

B. 1 and 3

C. 1 and 2

D. All of above

**Solution: (D)**

All of these are standard practices that are used in order to obtain good clustering results.

**Q33. What should be the best choice for number of clusters based on the following results:**

Chart, line chart

Description automatically generated

A. 5

B. 6

C. 14

D. Greater than 14

**Solution: (B)**

Based on the above results, the best choice of number of clusters using elbow method is 6.

**Q34. What should be the best choice for number of clusters based on the following results:**

Chart, line chart

Description automatically generated

A. 2

B. 4

C. 6

D. 8

**Solution: (C)**

Generally, a higher average silhouette coefficient indicates better clustering quality. In this plot, the optimal clustering number of grid cells in the study area should be 2, at which the value of the average silhouette coefficient is highest. However, the SSE of this clustering solution (k = 2) is too large. At k = 6, the SSE is much lower. In addition, the value of the average silhouette coefficient at k = 6 is also very high, which is just lower than k = 2. Thus, the best choice is k = 6.

**Q35. Which of the following sequences is correct for a K-Means algorithm using Forgy method of initialization?**

1. **Specify the number of clusters**
2. **Assign cluster centroids randomly**
3. **Assign each data point to the nearest cluster centroid**
4. **Re-assign each point to nearest cluster centroids**
5. **Re-compute cluster centroids**

Options:

A. 1, 2, 3, 5, 4

B. 1, 3, 2, 4, 5

C. 2, 1, 3, 4, 5

D. None of these

**Solution: (A)**

The methods used for initialization in K means are Forgy and Random Partition. The Forgy method randomly chooses k observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step, thus computing the initial mean to be the centroid of the cluster’s randomly assigned points.

**Q36. If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important:**

A. All the data points follow two Gaussian distribution

B. All the data points follow n Gaussian distribution (n >2)

C. All the data points follow two multinomial distribution

D. All the data points follow n multinomial distribution (n >2)

**Solution: (C)**

In EM algorithm for clustering its essential to choose the same no. of clusters to classify the data points into as the no. of different distributions they are expected to be generated from and also the distributions must be of the same type.

**Q37. Which of the following is/are not true about Centroid based K-Means clustering algorithm and Distribution based expectation-maximization clustering algorithm:**

1. **Both starts with random initializations**
2. **Both are iterative algorithms**
3. **Both have strong assumptions that the data points must fulfill**
4. **Both are sensitive to outliers**
5. **Expectation maximization algorithm is a special case of K-Means**
6. **Both requires prior knowledge of the no. of desired clusters**
7. **The results produced by both are non-reproducible.**

Options:

A. 1 only

B. 5 only

C. 1 and 3

D. 6 and 7

E. 4, 6 and 7

F. None of the above

**Solution: (B)**

All of the above statements are true except the 5th as instead K-Means is a special case of EM algorithm in which only the centroids of the cluster distributions are calculated at each iteration.

**Q38. Which of the following is/are not true about DBSCAN clustering algorithm:**

1. **For data points to be in a cluster, they must be in a distance threshold to a core point**
2. **It has strong assumptions for the distribution of data points in**dataspace
3. **It has substantially high time complexity of order O(n3)**
4. **It does not require prior knowledge of the no. of desired clusters**
5. **It is robust to outliers**

Options:

A. 1 only

B. 2 only

C. 4 only

D. 2 and 3

E. 1 and 5

F. 1, 3 and 5

**Solution: (D)**

* DBSCAN can form a cluster of any arbitrary shape and does not have strong assumptions for the distribution of data points in the dataspace.
* DBSCAN has a low time complexity of order O(n log n) only.

**Q39. Which of the following are the high and low bounds for the existence of F-Score?**

A. [0,1]

B. (0,1)

C. [-1,1]

D. None of the above

**Solution: (A)**

The lowest and highest possible values of F score are 0 and 1 with 1 representing that every data point is assigned to the correct cluster and 0 representing that the precession and/ or recall of the clustering analysis are both 0. In clustering analysis, high value of F score is desired.

**Q40. Following are the results observed for clustering 6000 data points into 3 clusters: A, B and C:**

Table

Description automatically generated

**What is the F1-Score with respect to cluster B?**

A. 3

B. 4

C. 5

D. 6

**Solution: (D)**

Here,

True Positive, TP = 1200

True Negative, TN = 600 + 1600 = 2200

False Positive, FP = 1000 + 200 = 1200

False Negative, FN = 400 + 400 = 800

Therefore,

Precision = TP / (TP + FP) = 0.5

Recall = TP / (TP + FN) = 0.6

Hence,

F1= 2 \* (Precision \* Recall)/ (Precision + recall) = 0.54 ~ 0.5

 ########################################################

The inability of machine learning method to capture the true relationship is called bias

Suppose you have data points which are

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

This straight line has high bias but for testing it has low variance,this is called underfit

A picture containing chart

Description automatically generated Chart

Description automatically generated

This squiggly line has low bias but high variance this is called overfit model.