Cluster Analysis – Quiz

Statistical Data Analysis

Question 1

What is the Manhattan distance between the points $(1, 2)$ and $(4, 6)$?

A. 25
B. 5
C. 4
D. 7

Question 1

What is the Manhattan distance between the points $(1, 2)$ and $(4, 6)$?

A. 25

B. 5

C. 4

D. 7

Answer: D

Manhattan distance is the sum of absolute differences: $|4 - 1| + |6 - 2| = 3 + 4 = 7$. Unlike Euclidean distance, which gives $\sqrt{9 + 16} = 5$, Manhattan distance measures the grid-path (taxicab) distance.

Question 2

For points $(0,0)$ and $(3,4)$, which ordering of Minkowski distances is correct as p increases from 1 to 2 to ∞?

A. $d_1 = 7,\ d_2 = 5,\ d_\infty = 4$
B. $d_1 = 4,\ d_2 = 5,\ d_\infty = 7$
C. $d_1 = 5,\ d_2 = 7,\ d_\infty = 4$
D. $d_1 = 7,\ d_2 = 5,\ d_\infty = 3$

Question 2

For points $(0, 0)$ and $(3, 4)$, which ordering of Minkowski distances is correct as p increases from 1 to 2 to ∞?

A. $d_1 = 7,\ d_2 = 5,\ d_\infty = 4$
B. $d_1 = 4,\ d_2 = 5,\ d_\infty = 7$
C. $d_1 = 5,\ d_2 = 7,\ d_\infty = 4$
D. $d_1 = 7,\ d_2 = 5,\ d_\infty = 3$

Answer: A
For $p = 1$: $|3| + |4| = 7$. For $p = 2$: $\sqrt{9 + 16} = 5$. For $p = \infty$: $\max(3, 4) = 4$. As p increases, the Minkowski distance decreases because larger p values are increasingly dominated by the single largest coordinate difference.

Question 3

Which distance metric produces diamond-shaped equidistance contours in 2D and is also called the L_1 norm?

A. Mahalanobis distance
B. Chebyshev distance
C. Manhattan distance
D. Euclidean distance

Question 3

Which distance metric produces diamond-shaped equidistance contours in 2D and is also called the L_1 norm?

A. Mahalanobis distance
B. Chebyshev distance
C. Manhattan distance
D. Euclidean distance

Answer: C
Manhattan distance (L_1 norm) produces diamond-shaped equidistance contours in 2D. This is because all points satisfying $|x_1| + |x_2| = c$ lie on a rotated square (diamond), unlike the circular contours of Euclidean distance or the square contours of Chebyshev distance.

Question 4

In the k-Means algorithm, what does the update step compute for each cluster?

A. The medoid (most central actual data point)
B. The mean of all points assigned to that cluster
C. The point with the maximum distance to all others
D. The median of all points assigned to that cluster

Question 4

In the k-Means algorithm, what does the update step compute for each cluster?

A. The medoid (most central actual data point)
B. The mean of all points assigned to that cluster
C. The point with the maximum distance to all others
D. The median of all points assigned to that cluster

Answer: B
The update step recalculates each centroid as the arithmetic mean of all points currently assigned to that cluster: $\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$.
This is what distinguishes k-Means from k-Medoids, where the center must be an actual data point.

Question 5

Given point $x = (2, 3)$ and two centroids $\mu_1 = (1, 2)$ and $\mu_2 = (5, 6)$, the squared Euclidean distances are $||x - \mu_1||^2 = 2$ and $||x - \mu_2||^2 = 18$. To which cluster is x assigned?

A. Both clusters share this point equally

B. Cluster 2, because $18 > 2$

C. Neither, because the point lies on the boundary

D. Cluster 1, because $2 < 18$

Question 5

Given point $x = (2, 3)$ and two centroids $\mu_1 = (1, 2)$ and $\mu_2 = (5, 6)$, the squared Euclidean distances are $||x - \mu_1||^2 = 2$ and $||x - \mu_2||^2 = 18$. To which cluster is x assigned?

A. Both clusters share this point equally

B. Cluster 2, because $18 > 2$

C. Neither, because the point lies on the boundary

D. Cluster 1, because $2 < 18$

Answer: D

In k-Means, each point is assigned to the cluster with the nearest centroid using $c_i = \arg\min_j ||x_i - \mu_j||^2$. Since the squared distance to μ_1 is 2, which is smaller than 18, point x is assigned to Cluster 1.

Question 6

The k-Means objective function $J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$ is guaranteed to:

A. Find the global optimum in every run
B. Oscillate between increasing and decreasing values
C. Decrease (or stay equal) with each iteration until convergence
D. Increase with each iteration until convergence

Question 6

The k-Means objective function $J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||^2$ is guaranteed to:

A. Find the global optimum in every run

B. Oscillate between increasing and decreasing values

C. Decrease (or stay equal) with each iteration until convergence

D. Increase with each iteration until convergence

Answer: C

Both the assignment step (assigning each point to its nearest centroid) and the update step (moving centroids to cluster means) can only decrease or maintain WCSS. However, because the algorithm may converge to a local minimum rather than the global one, different initializations can yield different final J values.

Question 7

In k-Means++ initialization, how is the second centroid chosen after the first is selected uniformly at random?

A. A random point with probability proportional to $D(x)^2$, where $D(x)$ is the distance to the nearest existing centroid
B. The point closest to the first centroid
C. A random point chosen uniformly from the remaining data
D. The point that minimizes the total WCSS

Question 7

In k-Means++ initialization, how is the second centroid chosen after the first is selected uniformly at random?

A. A random point with probability proportional to $D(x)^2$, where $D(x)$ is the distance to the nearest existing centroid
B. The point closest to the first centroid
C. A random point chosen uniformly from the remaining data
D. The point that minimizes the total WCSS

Answer: A
k-Means++ selects each subsequent centroid with probability proportional to the squared distance from the nearest existing centroid, $D(x)^2$. This spreads the initial centroids apart, making it provably better than purely random initialization and reducing the chance of poor local optima.

Question 8

What is a medoid in the context of the PAM (Partitioning Around Medoids) algorithm?

A. The arithmetic mean of all points in a cluster
B. The actual data point that minimizes the sum of distances to all other points in its cluster
C. The geometric center of the bounding box around a cluster
D. The point with the highest density in a cluster

Question 8

What is a medoid in the context of the PAM (Partitioning Around Medoids) algorithm?

A. The arithmetic mean of all points in a cluster
B. The actual data point that minimizes the sum of distances to all other points in its cluster
C. The geometric center of the bounding box around a cluster
D. The point with the highest density in a cluster

Answer: B
A medoid is the most centrally located data point in a cluster, defined as the point m_j that minimizes $\sum_{x_i \in C_j} d(x_i, m_j)$. Unlike the mean (centroid) in k-Means, a medoid must be an actual observation, which makes it more robust to outliers and easier to interpret.

Question 9

Consider the dataset $\{1, 2, 3, 4, 100\}$. The mean is 22 and the median is 3. Why does PAM typically produce better cluster centers than k-Means when outliers are present?

A. PAM uses the Euclidean distance exclusively, which is outlier-resistant

B. PAM removes outliers before clustering

C. PAM always assigns outliers to a separate noise cluster

D. PAM's medoid-based centers resist being pulled toward extreme values, unlike the mean-based centroids of k-Means

Question 9

Consider the dataset $\{1, 2, 3, 4, 100\}$. The mean is 22 and the median is 3. Why does PAM typically produce better cluster centers than k-Means when outliers are present?

A. PAM uses the Euclidean distance exclusively, which is outlier-resistant

B. PAM removes outliers before clustering

C. PAM always assigns outliers to a separate noise cluster

D. PAM's medoid-based centers resist being pulled toward extreme values, unlike the mean-based centroids of k-Means

Answer: D

Because PAM selects actual data points as cluster representatives (medoids), a single extreme outlier cannot pull the center to an unrepresentative location the way it can with the arithmetic mean in k-Means. Additionally, PAM can work with any distance metric, not just Euclidean, providing further flexibility in handling unusual data distributions.

Question 10

In agglomerative hierarchical clustering, what is the very first step of the algorithm?

A. Treat each individual data point as its own cluster
B. Merge all points into one cluster
C. Randomly select k initial centroids
D. Compute the silhouette coefficient for each point

Question 10

In agglomerative hierarchical clustering, what is the very first step of the algorithm?

A. Treat each individual data point as its own cluster

B. Merge all points into one cluster

C. Randomly select k initial centroids

D. Compute the silhouette coefficient for each point

Answer: A

Agglomerative (bottom-up) hierarchical clustering begins by treating each of the n data points as a singleton cluster. It then iteratively merges the two closest clusters until a single cluster containing all points remains, producing a dendrogram that records the full merge history.

Question 11

Which statement about agglomerative hierarchical clustering is TRUE?

A. It requires specifying k before the algorithm runs
B. It always produces the same result regardless of the linkage method used
C. Once two clusters are merged, that merge cannot be undone in later steps
D. It is faster than k-Means for large datasets

Question 11

Which statement about agglomerative hierarchical clustering is TRUE?

A. It requires specifying k before the algorithm runs

B. It always produces the same result regardless of the linkage method used

C. Once two clusters are merged, that merge cannot be undone in later steps

D. It is faster than k-Means for large datasets

Answer: C

Agglomerative clustering is a greedy algorithm: once two clusters are merged, that decision is permanent and cannot be reversed in subsequent iterations. This is a key disadvantage because a locally optimal merge at one step may turn out to be globally suboptimal.

Question 12

Hierarchical clustering has space complexity $O(n^2)$ for the distance matrix. For a dataset with $n = 50{,}000$ points, approximately how many entries are in the full $n \times n$ distance matrix?

A. 250 million
B. 50,000
C. 2,500
D. 2.5 billion

Question 12

Hierarchical clustering has space complexity $O(n^2)$ for the distance matrix. For a dataset with $n = 50{,}000$ points, approximately how many entries are in the full $n \times n$ distance matrix?

A. 250 million

B. 50,000

C. 2,500

D. 2.5 billion

Answer: D

The full distance matrix has $n^2 = 50{,}000^2 = 2.5 \times 10^9$ entries (or about 1.25 billion unique pairs in the symmetric upper triangle). This quadratic memory requirement is why hierarchical clustering becomes impractical for very large datasets, unlike k-Means which scales as $O(nkd \cdot i)$.

Question 13

Single linkage defines the distance between two clusters as $d(A, B) = \min_{a \in A, b \in B} d(a, b)$. What is the main practical problem with this approach?

A. It always produces perfectly spherical clusters

B. The chaining effect, where distinct groups can be connected through a chain of close points

C. It requires Ward's variance formula to compute

D. It cannot be used with Euclidean distance

Question 13

Single linkage defines the distance between two clusters as $d(A, B) = \min_{a \in A, b \in B} d(a, b)$. What is the main practical problem with this approach?

A. It always produces perfectly spherical clusters

B. The chaining effect, where distinct groups can be connected through a chain of close points

C. It requires Ward's variance formula to compute

D. It cannot be used with Euclidean distance

Answer: B

Single linkage is susceptible to the chaining effect: if two separate clusters happen to have a few points that are close to each other (or noise points between them), single linkage will merge them prematurely. This produces elongated, chain-like clusters that can inappropriately connect truly distinct groups.

Question 14

Ward's method defines the merge cost as $\Delta(A, B) = \frac{|A||B|}{|A|+|B|} ||\mu_A - \mu_B||^2$. Which property does this formula share with the k-Means objective?

A. Both use medoids instead of means
B. Both are invariant to feature scaling
C. Both minimize within-cluster variance
D. Both require specifying the number of clusters in advance

Question 14

Ward's method defines the merge cost as $\Delta(A, B) = \frac{|A||B|}{|A|+|B|} ||\mu_A - \mu_B||^2$. Which property does this formula share with the k-Means objective?

A. Both use medoids instead of means

B. Both are invariant to feature scaling

C. Both minimize within-cluster variance

D. Both require specifying the number of clusters in advance

Answer: C

Ward's method merges the pair of clusters that causes the smallest increase in total within-cluster variance (sum of squared deviations from cluster means). This is the same objective that k-Means minimizes, which is why Ward's method tends to produce compact, spherical clusters of similar sizes, just as k-Means does.

Question 15

In a dendrogram, what does the height at which two branches merge represent?

A. The distance (or dissimilarity) at which the two clusters were joined
B. The number of points in each cluster
C. The silhouette score of the resulting merged cluster
D. The iteration number of the algorithm

Question 15

In a dendrogram, what does the height at which two branches merge represent?

A. The distance (or dissimilarity) at which the two clusters were joined
B. The number of points in each cluster
C. The silhouette score of the resulting merged cluster
D. The iteration number of the algorithm

Answer: A
The vertical axis of a dendrogram represents the distance or dissimilarity between the two clusters at the moment they are merged. A merge occurring at a greater height means the two clusters being joined are more dissimilar, which is why large vertical gaps in a dendrogram often suggest natural cluster boundaries.

Question 16

You cut a dendrogram with a horizontal line and the line crosses exactly 4 branches. How many clusters does this cut produce?

A. 3 clusters
B. 4 clusters
C. 5 clusters
D. It depends on the linkage method used

Question 16

You cut a dendrogram with a horizontal line and the line crosses exactly 4 branches. How many clusters does this cut produce?

A. 3 clusters

B. 4 clusters

C. 5 clusters

D. It depends on the linkage method used

Answer: B

Each branch that the horizontal cut line crosses corresponds to one cluster. So crossing 4 branches produces exactly 4 clusters. This is the practical way to extract a flat clustering from a hierarchical dendrogram: a higher cut yields fewer clusters and a lower cut yields more clusters.

Question 17

The elbow method plots WCSS against the number of clusters k. Why does WCSS always decrease as k increases?

A. Because more clusters means fewer outliers are detected

B. Because the algorithm runs more iterations for larger k

C. Because standardization is applied automatically for larger k

D. Because adding more cluster centers allows each point to be closer to its assigned center, reducing the total squared distance

Question 17

The elbow method plots WCSS against the number of clusters k. Why does WCSS always decrease as k increases?

A. Because more clusters means fewer outliers are detected

B. Because the algorithm runs more iterations for larger k

C. Because standardization is applied automatically for larger k

D. Because adding more cluster centers allows each point to be closer to its assigned center, reducing the total squared distance

Answer: D

With more centroids available, each data point can be assigned to a nearer center, which mechanically reduces the within-cluster sum of squares. In the extreme case $k = n$, every point is its own cluster and WCSS equals zero. The elbow method seeks the value of k where the rate of decrease sharply levels off.

Question 18

A point i has silhouette coefficient $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$ where $a(i) = 0.3$ and $b(i) = 0.9$. What is $s(i)$ and what does it indicate?

A. $s(i) = 0.67$; the point is well-assigned to its cluster
B. $s(i) = -0.67$; the point is likely misclassified
C. $s(i) = 0.33$; the point is on the cluster boundary
D. $s(i) = 0.60$; the point is moderately well-assigned

Question 18

A point i has silhouette coefficient $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$ where $a(i) = 0.3$ and $b(i) = 0.9$. What is $s(i)$ and what does it indicate?

A. $s(i) = 0.67$; the point is well-assigned to its cluster

B. $s(i) = -0.67$; the point is likely misclassified

C. $s(i) = 0.33$; the point is on the cluster boundary

D. $s(i) = 0.60$; the point is moderately well-assigned

Answer: A

Computing: $s(i) = (0.9 - 0.3)/\max(0.3, 0.9) = 0.6/0.9 = 0.67$. A value close to 1 indicates the point is much closer to its own cluster ($a(i)$ is small) than to the nearest neighboring cluster ($b(i)$ is large), confirming a good cluster assignment.

Question 19

The gap statistic compares $\log(WCSS_k)$ of the observed data to that of random uniform reference data. What does a large gap value for a given k indicate?

A. The observed data has less structure than random data at that k

B. The algorithm failed to converge for that k

C. The clustering at that k captures genuine structure that would not appear in structureless data

D. The distance metric is inappropriate for the dataset

Question 19

The gap statistic compares $\log(WCSS_k)$ of the observed data to that of random uniform reference data. What does a large gap value for a given k indicate?

A. The observed data has less structure than random data at that k

B. The algorithm failed to converge for that k

C. The clustering at that k captures genuine structure that would not appear in structureless data

D. The distance metric is inappropriate for the dataset

Answer: C

The gap statistic is $Gap(k) = E^*[\log(WCSS_k)] - \log(WCSS_k)$, where E^* is the expected value under a uniform null distribution. A large gap means the observed data achieves a much lower WCSS than random data would at that k, indicating genuine cluster structure. The optimal k is the one that maximizes the gap statistic.

Question 20

A dataset has two features: age (range 20–80) and income (range 20,000–200,000). Before applying k-Means, you apply z-score standardization $z_j = (x_j - \mu_j)/\sigma_j$. What is the primary effect?

A. It converts all features to categorical variables

B. It gives each feature mean 0 and standard deviation 1, so income no longer dominates the distance calculation

C. It removes all outliers from the dataset

D. It guarantees that k-Means will find the global optimum

Question 20

A dataset has two features: age (range 20–80) and income (range 20,000–200,000). Before applying k-Means, you apply z-score standardization $z_j = (x_j - \mu_j)/\sigma_j$. What is the primary effect?

A. It converts all features to categorical variables

B. It gives each feature mean 0 and standard deviation 1, so income no longer dominates the distance calculation

C. It removes all outliers from the dataset

D. It guarantees that k-Means will find the global optimum

Answer: B

Without standardization, the income feature (range 180,000) would dominate Euclidean distance calculations over age (range 60), effectively making age irrelevant. Z-score standardization transforms each feature to have mean 0 and standard deviation 1, ensuring both features contribute equally to the distance computation and resulting cluster assignments.