Cluster Analysis – Quiz

Probability & Statistics

Question 1

What is the main goal of K-means clustering?

A. To classify new observations into known categories
B. To partition data into K clusters by minimizing within-cluster variance
C. To predict a continuous outcome variable
D. To reduce the number of variables

Question 1

What is the main goal of K-means clustering?

A. To classify new observations into known categories
B. To partition data into K clusters by minimizing within-cluster variance
C. To predict a continuous outcome variable
D. To reduce the number of variables

Answer: B
K-means aims to partition observations into K clusters such that each observation belongs to the cluster with the nearest mean, minimizing within-cluster sum of squares.

Question 2

The elbow method is used to:

A. Choose the optimal number of clusters K
B. Select the distance metric
C. Determine the linkage method
D. Initialize cluster centroids

Question 2

The elbow method is used to:

A. Choose the optimal number of clusters K
B. Select the distance metric
C. Determine the linkage method
D. Initialize cluster centroids

Answer: A
The elbow method plots within-cluster sum of squares against K. The "elbow" point where improvements slow down suggests the optimal number of clusters.

Question 3

In hierarchical clustering, a dendrogram shows:

A. The final cluster assignments only
B. The nested hierarchy of clusters at different levels
C. The variance explained by each cluster
D. The optimal value of K

Question 3

In hierarchical clustering, a dendrogram shows:

A. The final cluster assignments only
B. The nested hierarchy of clusters at different levels
C. The variance explained by each cluster
D. The optimal value of K

Answer: B
A dendrogram displays the hierarchical relationship between observations and clusters. Cutting it at different heights gives different numbers of clusters.

Question 4

What distinguishes DBSCAN from K-means?

A. DBSCAN requires specifying K in advance
B. DBSCAN can find clusters of arbitrary shape and identify noise points
C. DBSCAN only works with categorical data
D. DBSCAN always produces the same clusters as K-means

Question 4

What distinguishes DBSCAN from K-means?

A. DBSCAN requires specifying K in advance

B. DBSCAN can find clusters of arbitrary shape and identify noise points

C. DBSCAN only works with categorical data

D. DBSCAN always produces the same clusters as K-means

Answer: B

DBSCAN is density-based, can find non-spherical clusters, automatically determines the number of clusters, and identifies noise points (outliers).

Question 5

The silhouette score measures:

A. The number of clusters
B. How similar an object is to its own cluster compared to other clusters
C. The distance between cluster centroids
D. The variance within clusters

Question 5

The silhouette score measures:

A. The number of clusters
B. How similar an object is to its own cluster compared to other clusters
C. The distance between cluster centroids
D. The variance within clusters

Answer: B
Silhouette score ranges from -1 to 1. Values near 1 mean objects are well-matched to their cluster and poorly-matched to neighboring clusters.

Question 6

Which linkage method in hierarchical clustering tends to create compact, spherical clusters?

A. Single linkage
B. Complete linkage
C. Average linkage
D. Ward's method

Question 6

Which linkage method in hierarchical clustering tends to create compact, spherical clusters?

A. Single linkage
B. Complete linkage
C. Average linkage
D. Ward's method

Answer: D
Ward's method minimizes total within-cluster variance, producing compact, spherical clusters. Single linkage can create elongated "chaining" effects.

Question 7

Why is data standardization important before clustering?

A. It speeds up the algorithm
B. It prevents variables with larger scales from dominating the distance calculation
C. It is required by all clustering algorithms
D. It increases the number of clusters found

Question 7

Why is data standardization important before clustering?

A. It speeds up the algorithm

B. It prevents variables with larger scales from dominating the distance calculation

C. It is required by all clustering algorithms

D. It increases the number of clusters found

Answer: B

Without standardization, variables with larger scales contribute more to distance calculations, potentially biasing cluster formation regardless of their actual importance.

Question 8

A limitation of K-means clustering is:

A. It can only find two clusters
B. It assumes spherical clusters of similar size
C. It requires labeled data
D. It cannot handle continuous variables

Question 8

A limitation of K-means clustering is:

A. It can only find two clusters
B. It assumes spherical clusters of similar size
C. It requires labeled data
D. It cannot handle continuous variables

Answer: B
K-means assumes clusters are spherical and of similar size. It struggles with elongated clusters, clusters of very different sizes, or non-convex shapes.

Question 9

In K-means, the algorithm converges when:

A. All observations are assigned to one cluster
B. The number of iterations reaches K
C. Cluster assignments no longer change between iterations
D. The silhouette score reaches 1

Question 9

In K-means, the algorithm converges when:

A. All observations are assigned to one cluster
B. The number of iterations reaches K
C. Cluster assignments no longer change between iterations
D. The silhouette score reaches 1

Answer: C
K-means iterates between assigning points to nearest centroids and updating centroids. It converges when assignments stabilize (no points change clusters).

Question 10

Which distance metric is most commonly used in K-means clustering?

A. Manhattan distance
B. Euclidean distance
C. Cosine similarity
D. Jaccard distance

Question 10

Which distance metric is most commonly used in K-means clustering?

A. Manhattan distance
B. Euclidean distance
C. Cosine similarity
D. Jaccard distance

Answer: B
K-means typically uses Euclidean distance (L2 norm) to measure the distance between points and cluster centroids. This is why it tends to find spherical clusters.