What is K-Means Clustering?

K-Means is an unsupervised machine learning algorithm that groups data into clusters, with each cluster centered around a centroid. The aim is to minimize the distance between data points and their cluster centroids.

Why is K-Means called an unsupervised algorithm?

It's unsupervised because it doesn't rely on labeled data; it identifies natural groupings within data without predefined categories.

What is the significance of 'K' in K-Means?

'K' represents the number of clusters into which the data will be divided.

Explain the Elbow Method in K-Means.

The Elbow Method helps to find the optimal number of clusters by plotting the Within Cluster Sum of Squared Errors (WCSS) against the number of clusters. The "elbow" point, where the WCSS curve starts flattening, suggests the ideal number of clusters.

What is Within Cluster Sum of Squares (WCSS)?

WCSS measures the sum of squared distances between each data point and its corresponding cluster centroid.

What is a centroid in K-Means clustering?

A centroid is the center of a cluster, calculated as the mean position of all points within that cluster.

How does K-Means clustering work?

It initializes 'K' centroids randomly, assigns each data point to the nearest centroid, recalculates centroids based on assigned points, and repeats until centroids stabilize.

What type of data is suitable for K-Means clustering?

K-Means works well with numerical data where calculating the mean and distances between points is meaningful.

List some applications of K-Means clustering.

Applications include customer segmentation in marketing, property classification in real estate, book categorization in libraries, and document analysis in research.

What is the importance of normalization in K-Means clustering?

Normalization scales data, making distances meaningful and preventing features with larger ranges from dominating.

How do you handle the problem of choosing 'K'?

Methods like the Elbow Method and the Silhouette Score can guide the choice of an appropriate number of clusters.

What are some limitations of K-Means clustering?

Limitations include sensitivity to initial centroid selection, difficulty with clusters of varying densities or shapes, and a need to specify 'K' in advance.

Why is K-Means sensitive to the initial choice of centroids?

Poor initialization can lead to convergence at local minima, resulting in suboptimal clusters.

What is the K-Means++ initialization?

K-Means++ initializes centroids in a way that improves convergence speed and clustering quality by spreading initial centroids across data points.

Explain how clusters are updated in K-Means.

After assigning points to centroids, the centroids are recalculated as the mean of all points in each cluster, repeating until convergence.

What are the convergence criteria in K-Means?

Convergence happens when centroids no longer move significantly or when a maximum number of iterations is reached.

How can we evaluate the quality of clusters in K-Means?

Cluster quality can be assessed by the WCSS, Silhouette Score, and visual inspection of clusters.

Can K-Means handle categorical data?

K-Means isn’t suited for categorical data, as it requires mean calculations, which are only meaningful for numeric data.

Describe a real-world scenario where K-Means clustering is useful.

In marketing, K-Means can segment customers based on purchasing behavior, allowing targeted promotions and marketing strategies.

What are some libraries used in Python for implementing K-Means?

Libraries include Scikit-Learn (`KMeans`), NumPy for data manipulation, and Matplotlib and Seaborn for visualization.

Explain the role of the `fit` and `predict` functions in K-Means implementation.

`fit` trains the K-Means model on data, while `predict` assigns cluster labels to data points.

What does the ‘inertia’ attribute in Scikit-Learn’s KMeans represent?

Inertia is the WCSS for all clusters, representing the compactness of the clusters.

What is a Silhouette Score, and how is it used?

A Silhouette Score measures cluster separation; a higher score indicates well-separated clusters. It helps to choose the optimal ‘K'.

What is a dendrogram, and is it used in K-Means?

A dendrogram is a tree-like diagram representing hierarchical clustering, not directly used in K-Means.

What happens if you choose too high or too low a value for ‘K’?

Too high a 'K' may create too many small, irrelevant clusters; too low a 'K' might oversimplify groupings.

How does increasing the number of features affect K-Means clustering?

With more features, the algorithm may struggle to find stable clusters due to the "curse of dimensionality.”

What are the main steps in visualizing K-Means clusters?

Steps include scatter plots for 2D clusters, 3D plots for three features, and color-coding clusters for differentiation.

Describe a scenario where K-Means would not be effective.

K-Means may fail for non-spherical or unevenly sized clusters due to its reliance on distance and centroid calculation.