Study Guide: K-Means Clustering

Definition and Characteristics

* K-means clustering is an unsupervised learning algorithm that partitions the data into K clusters based on their similarities.
* It represents each cluster by its cluster center (mean).
* The algorithm assigns each data point to its nearest cluster center.

How it Works

1. Initialize cluster centers randomly.
2. Assign each data point to the closest cluster center.
3. Recompute the cluster centers as the mean of all data points assigned to the same cluster.
5. Repeat steps 2-3 until the clusters converge (i.e., the cluster assignments no longer change).

Mathematical Representation

* The objective function to optimize is the sum of squared distances from each data point to its assigned center.
* The K-means algorithm converges in a finite number of iterations, where convergence is measured by the objective function ceasing to change.

Theorem 16 (K-Means Convergence Theorem)

Proof: The proof works by showing that the algorithm changes the values of μ or z only in two points, and both changes lead to convergence.

Applications and Variations

* K-means clustering can be used for data points in Rd (high-dimensional space) or for nodes in a graph with distances on edges.
* There are two variations of the clustering problem for each of the criteria:
+ Require each cluster center to be a data point.
+ Allow a cluster center to be any point in space.
* High-density clusters: An alternative assumption often made is that clusters consist of high-density regions surrounded by low-density "moats" between them.

Examples and Diagrams

* Figure 3.14: First few iterations of K-means running on a previous data set.
* Figure 7.1: Clustering of high-density regions surrounded by low-density "moats" between them.

Summary of Key Points

* K-means clustering is an unsupervised learning algorithm that partitions the data into K clusters based on their similarities.
* The algorithm represents each cluster by its cluster center (mean) and assigns each data point to the nearest cluster center.
* The algorithm converges in a finite number of iterations.

Flashcards

1. What is K-means clustering?

Answer: An unsupervised learning algorithm that partitions the data into K clusters based on their similarities.

2. How does K-means clustering?

Answer: By representing each cluster by its cluster center (mean) and assigning each data point to the nearest cluster center.

3. What is the objective function optimized in K-means clustering?

Answer: The sum of squared distances from each data point to its assigned center.