K-means clustering

Last Updated: 2023-02-23

12.8

The K-means algorithm is the most widely used clustering algorithm that uses an explicit distance measure to partition the data set into clusters.

The main concept of the K-means algorithm is to represent each cluster by the vector of mean attribute values of all training instances for numeric attributes and by the vector of modal (most frequent) values for nominal attributes that are assigned to that cluster. This cluster representation is called *cluster center*.

The following conditions apply to the cluster center:

- The algorithm handles continuous attributes and nominal attributes.
- You can handle the processes of cluster formation and cluster modeling in a computationally efficient way by applying the distance function to match instances against cluster centers.

- **Background of K-means clustering**
  The algorithm operates by doing several iterations of the same basic process.
- **Usage of K-means clustering**
  The K-means algorithm usually compares well to more refined and computationally expensive clustering algorithms concerning the quality of results.
- **Functions for K-means clustering**
  The K-means algorithm is implemented in the KMEANS stored procedure and the PREDICT_KMEANS stored procedure. To print a K-means model, use the PRINT_MODEL procedure.
- **Examples for creating K-means clustering models**
  This example creates a clustering model for the *customer churn* data set.
- **Model table data formats for K-means clustering**
  The model tables are created in the database where you run the algorithm.

Parent topic:

→ Analytic stored procedures