K-Means Clustering

❑Cluster: collection of data objects that are similar to one another within the same cluster and dissimilar to the objects in other cluster

❑Clustering: grouping set of objects into classes of similar objects

❑k means: classical partitioning method

❑Partitioning : A type of clustering approach.

❑Major clustering methods:

- Partitioning
- Hierarchical
- Density
- Grid-based
- Model-based

Partitioning methods

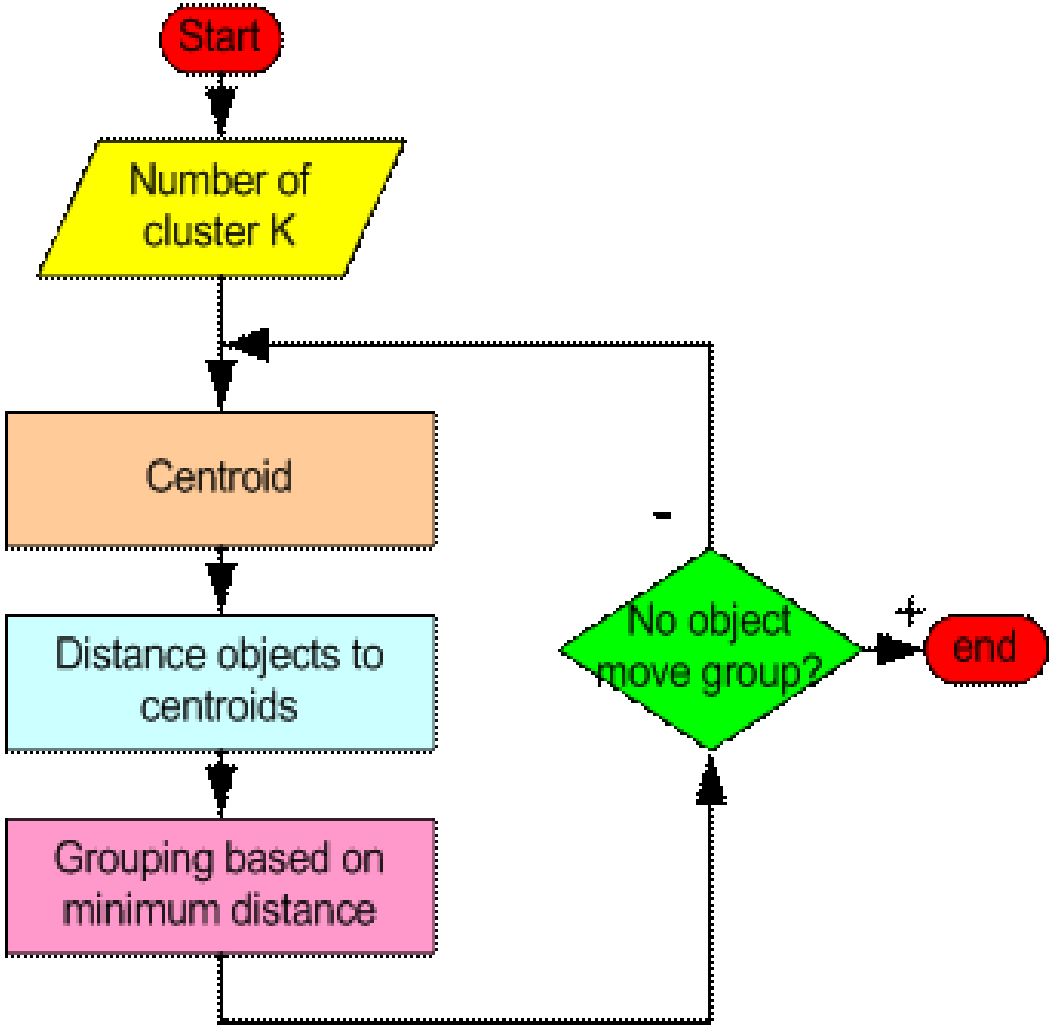
Given : database of n object Construct: k partitions of the data Each partition represents a cluster and k<=n	Classify the data into k groups, which satisfy the following: -Each group must contain at least 1 object -Each object must belong to exactly 1 group	2 heuristic methods: (1) k-means algorithm (2)k-medoids algorithm

Input: The number of clusters k and a database containing n objects

Output: A set of k clusters that minimizes the squared error criterion

Method:

(1) Arbitrarily choose k objects as the initial cluster centers
(2) Repeat
(3) (re)assign each object to the cluster to which the object is most similar based on the mean value if the objects in the cluster
(4) Update the cluster means i.e. calculate the mean value of the objects for each cluster
(5) until no change


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

- High intra-cluster similarity
- Low inter cluster similarity
- Cluster similarity : measured in regard to the mean value of the objects in a cluster

Working:

- Randomly select k of the objects
- Each represents a cluster mean or center
- For each of the remaining objects , an object is assigned to the cluster to which it is most similar, based on distance between object and the cluster mean
- Then compute new mean for each cluster
- Process iterates until the criterion function converges

Squared error criterion

$$E = \sum_{i=1}^{k} \sum_{p \varepsilon Ci} |p - m_i|^2$$

E sum of square error for all objects
P point in space representing a given object
m_i is the mean of the cluster Ci

Square-error : is the sum of the Euclidean distances between each pattern and its cluster center.
The algorithm converges when the criterion function cannot be improved.

Elements to be clustered: 2 3 6 8 12 15 18 22
number of clusters: 3

At this step Value of clusters
K1{ 2 }
K2{ 3 }
K3{ 6 8 12 15 18 22 }
Value of m
m1=2.0 m2=3.0 m3=13.5

At this step
Value of clusters
K1{ 2 }
K2{ 3 6 8 }
K3{ 12 15 18 22 }
Value of m
m1=2.0 m2=5.666666666666667 m3=16.75

At this step Value of clusters
K1{ 2 3 }
K2{ 6 8 }
K3{ 12 15 18 22 }
Value of m
m1=2.5 m2=7.0 m3=16.75

At this step Value of clusters
K1{ 2 3 }
K2{ 6 8 }
K3{ 12 15 18 22 }
Value of m
m1=2.5 m2=7.0 m3=16.75

The Final Clusters By K means are as follows:
K1{ 2 3 }
K2{ 6 8 }
K3{ 12 15 18 22 }