TUNING K-MEANS

SYRACUSE UNIVERSITY
School of Information Studies

SOLUTIONS TO INITIAL CENTROIDS PROBLEM

Perform multiple runs, changing random seeds every time.

Helps, but probability is not on your side

Sample and use hierarchical clustering to determine initial centroids.

Select more than *k* initial centroids and then select among these initial centroids.

Select most widely separated.

COMPARE SSE OF DIFFERENT INITIAL CENTROIDS

Most common measure is **sum of squared error (SSE).**

x is a data point in cluster C_i and m_i is the centroid or medoid for cluster C_i.

For each point, the error is the distance to the centroid or medoid.

To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

Given two clustering results with **same** number of clusters, we can choose the one with the smallest SSE.

BE CAREFUL WITH SSE

Attention! One easy way to reduce SSE is to increase *k*, the number of clusters. When *k* equals the data set size, meaning, each data point is in its own cluster, then SSE = 0.

Don't simply use *k* to reduce SSE. *k* should have a reasonable range of value in real applications.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

WHAT IF THE ITERATION NEVER STOPS?

Set maximum number of iterations.

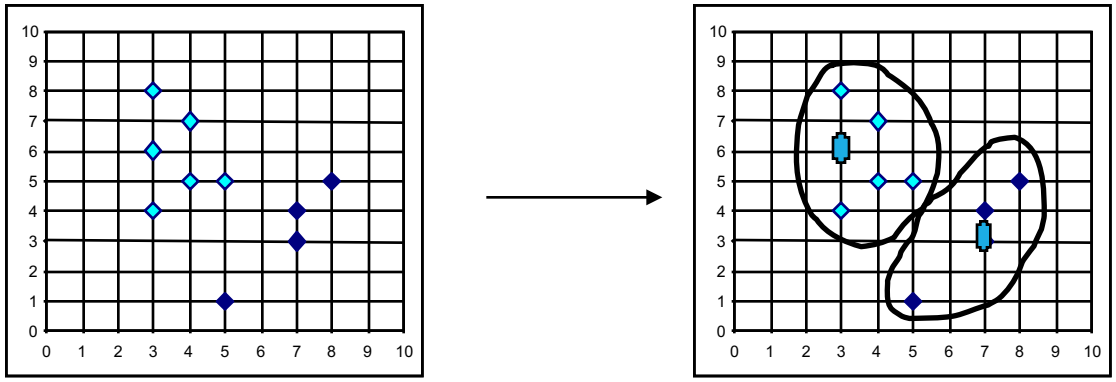
Set minimum value of SSE change.

USE MEDOIDS TO RESIST OUTLIERS IN K-MEANS

The k-means algorithm is sensitive to outliers!

Since an object with an extremely large value may substantially distort the distribution of the data

k-medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.



PAM: A K-MEDOID ALGORITHM

http://www.cs.umb.edu/cs738/pam1.pdf

PAM: Partition Around Medoids

The goal of the algorithm is to minimize the average dissimilarity of objects to their closest selected object. Equivalently, we can minimize the sum of the dissimilarities between object and their closest selected object.

The algorithm has two phases:

(i) In the first phase, BUILD, a collection of k objects are selected for an initial set S.

(ii) In the second phase, SWAP, one tries to improve the quality of the clustering by exchanging selected objects with unselected objects.

VARIATIONS OF THE K-MEANS METHOD

One variation comprises mixture models (soft clustering).

Estimates clusters from probability distributions

Includes the **expectation-maximization** (EM) algorithm

CLUSTER VALIDITY

For supervised classification, we have a variety of measures to evaluate how good our model ism

- Accuracy, precision, recall

For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters.

But "clusters are in the eye of the beholder"!

DIFFERENT METHODS FOR CLUSTER VALIDATION

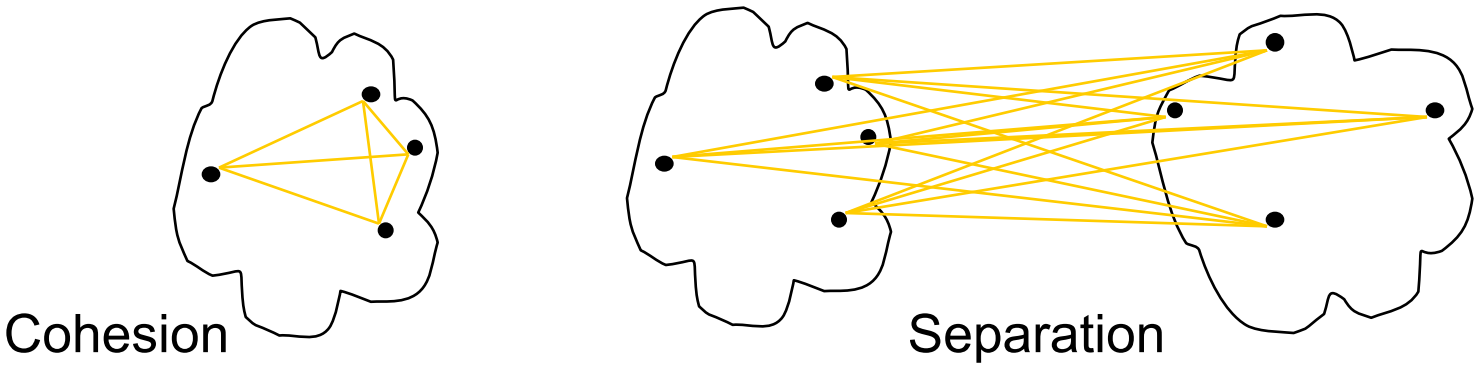
Cluster cohesion: Measures how closely related objects are in a cluster

- High **intraclass** similarity
- SSE as a cohesion measure

Cluster separation: Measure how distinct or well separated a cluster is from other clusters

- Low **interclass** similarity

Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels

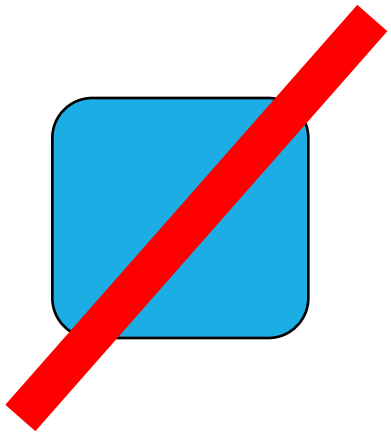


COMMENTS ON THE K-MEANS METHOD

Strength: *Relatively efficient*

Weaknesses:

- Need to specify *k*, the *number* of clusters, in advance
- Unable to handle noisy data and *outliers*
- Not suitable to discover clusters with *nonconvex shapes*



FINAL COMMENT ON CLUSTER VALIDITY

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data (Jain & Dubes)