K-Means for Document Clustering

School of Information Studies
Syracuse University

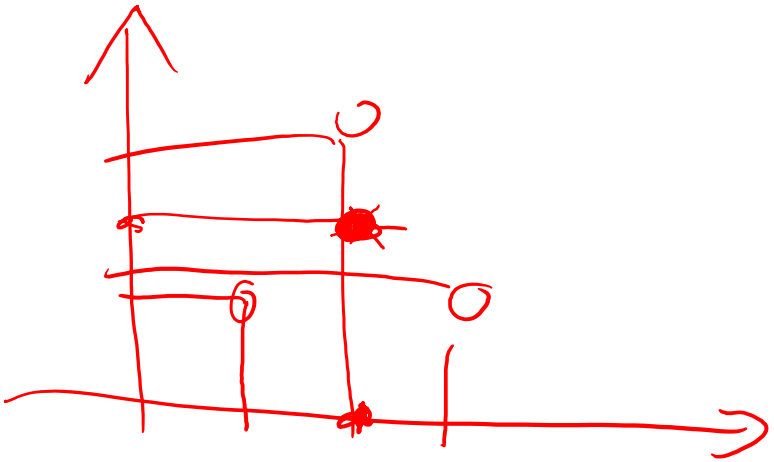
Similarity/Distance Measures

Euclidean distance

Cosine similarity measure

Centroid of a Cluster

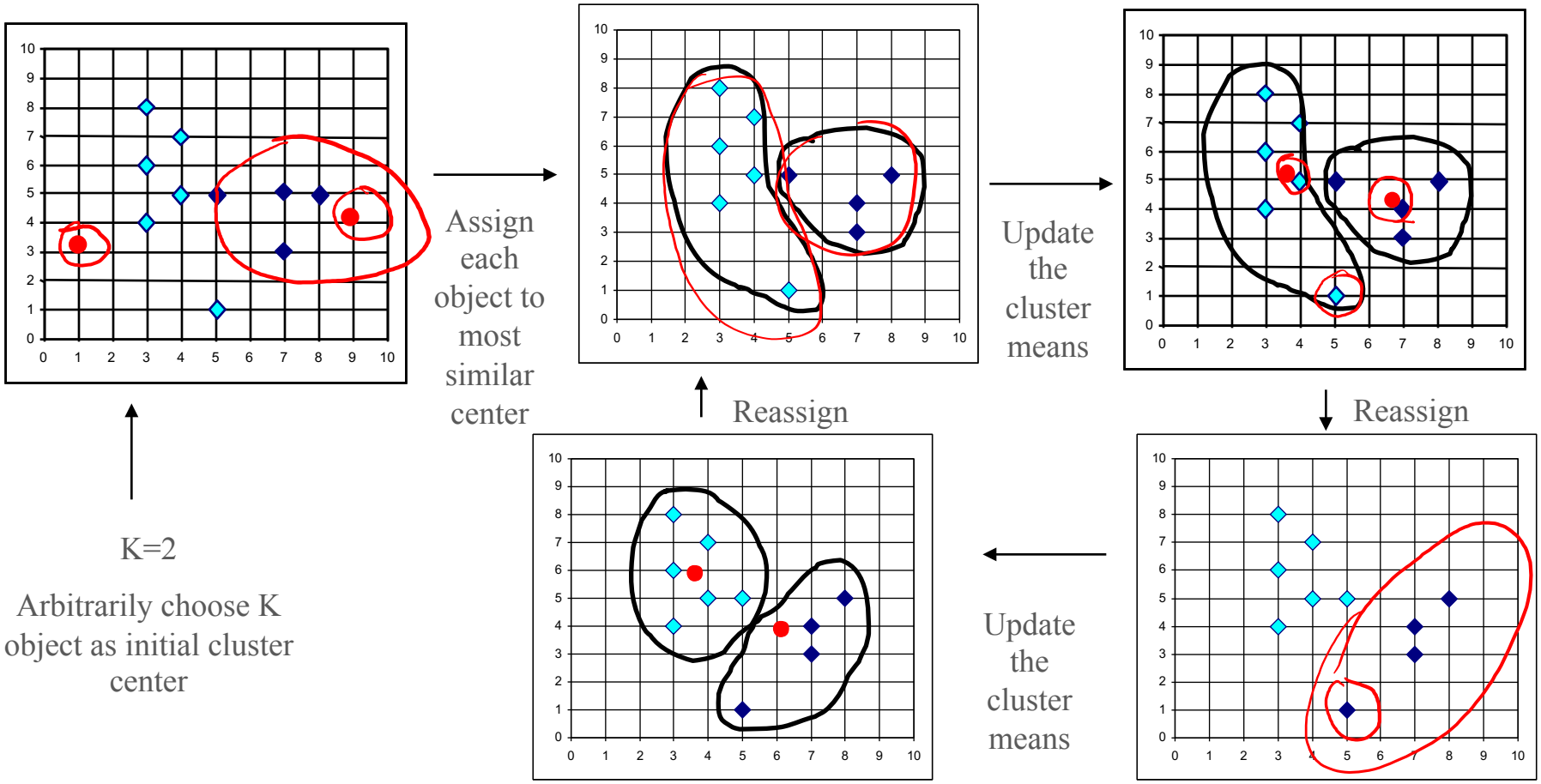
Centroid: The "center" of a cluster is a (pseudo) instance of data in which each attribute is the "mean" of all the attribute values in the cluster.



The *K-Means* Clustering Method

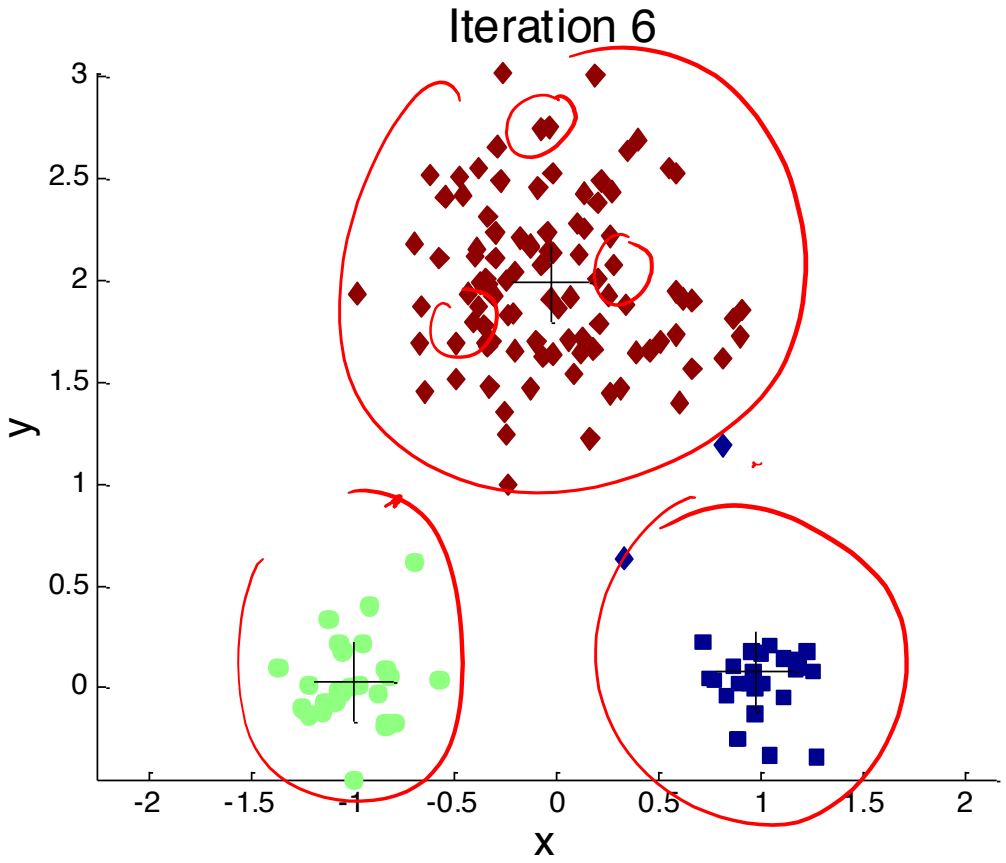
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

The *K-Means* Clustering Method: Example



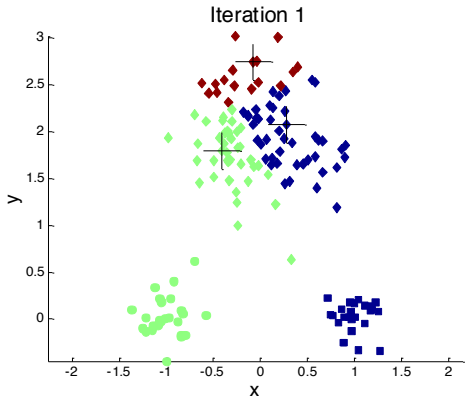
School of Information Studies
Syracuse University

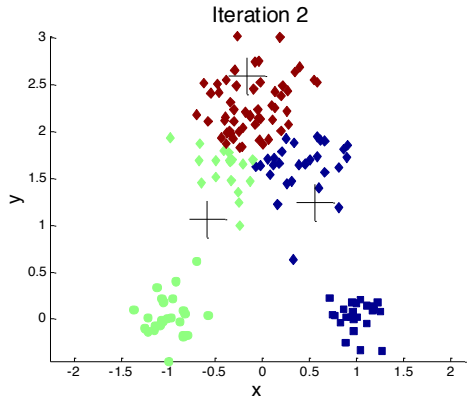
Importance of Choosing Initial Centroids

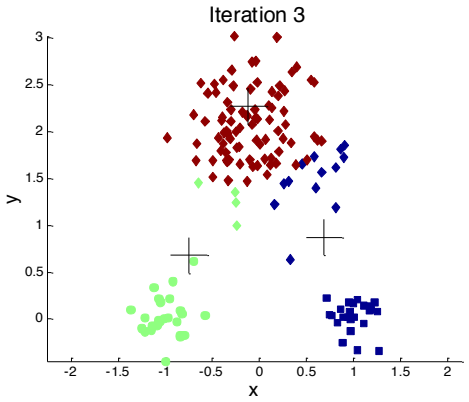


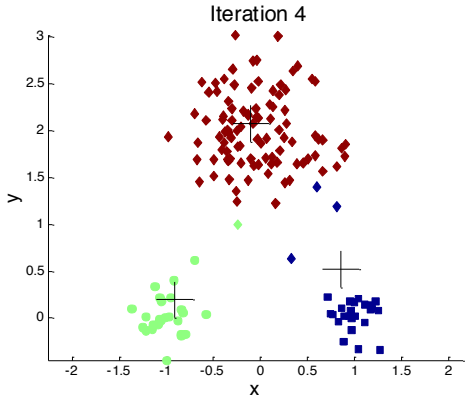
A good clustering result

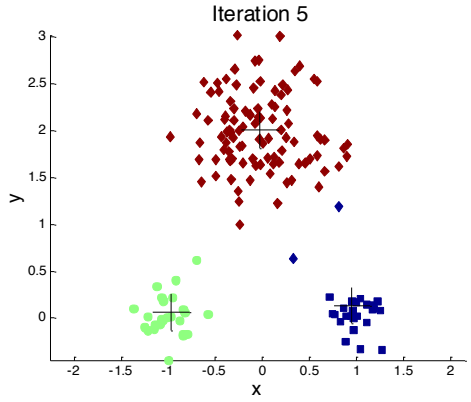
Importance of Choosing Initial Centroids

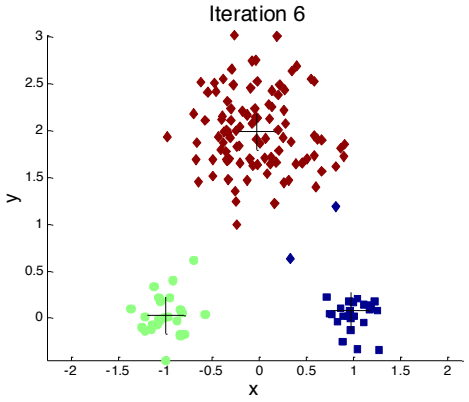




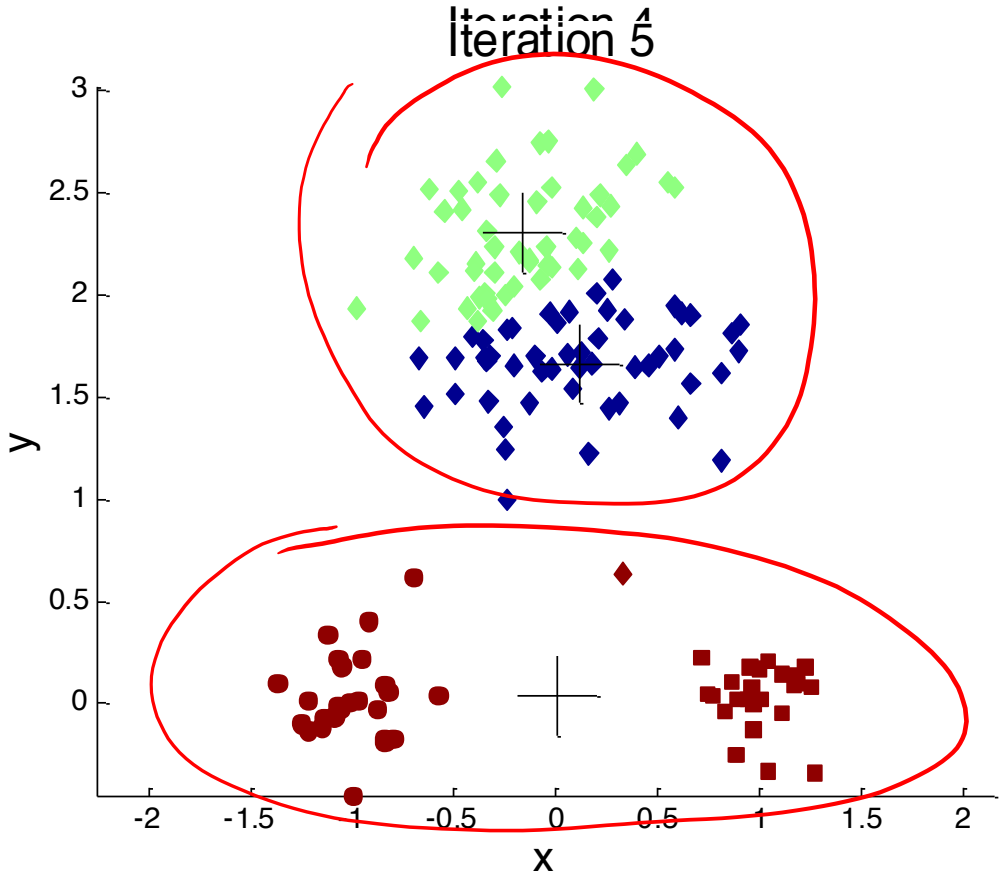






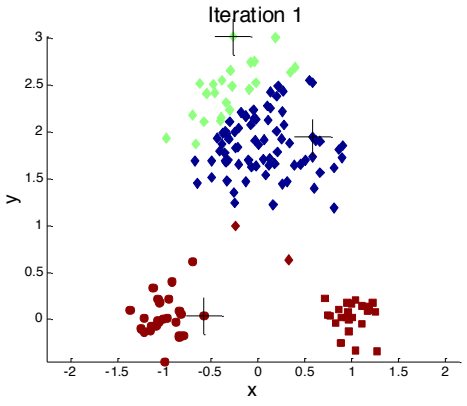


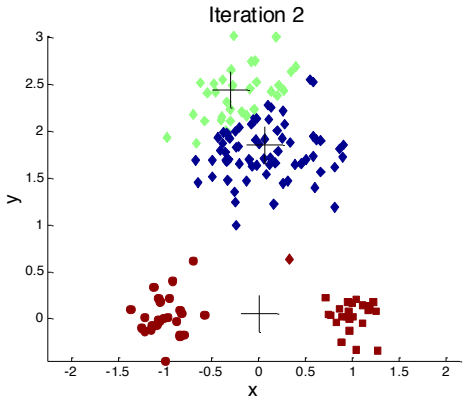
Importance of Choosing Initial Centroids

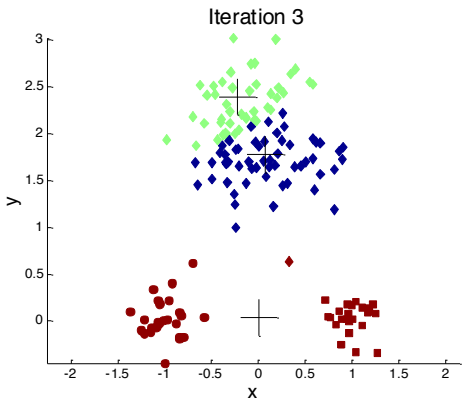


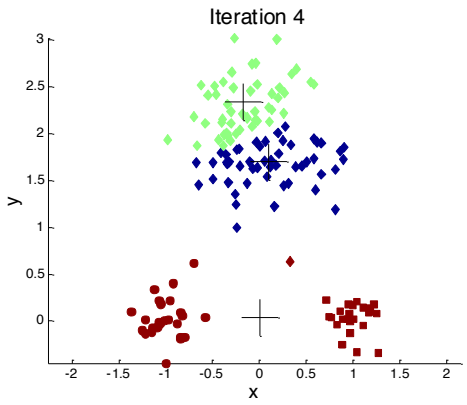
A less meaningful result

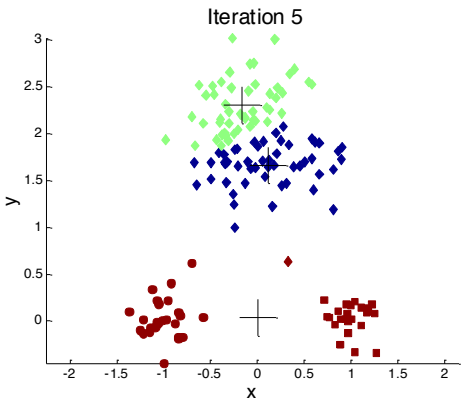
Importance of Choosing Initial Centroids











How to Choose Initial Centroids?

Multiple runs, changing random seeds every time

- Each random seed corresponds to one set of randomly chosen centroids.

Compare *SSE (Sum of Squared Errors)* for each run

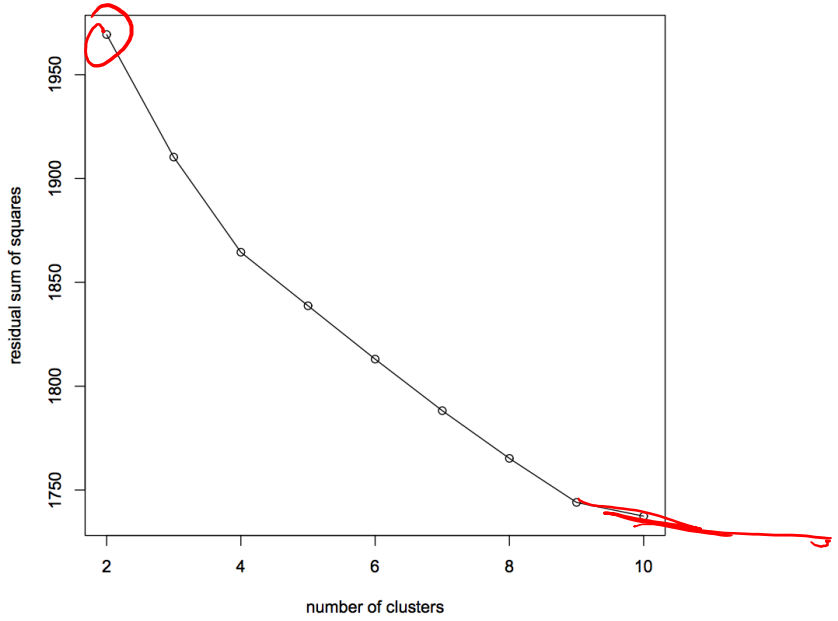
- x is a data point in cluster C_i and m_i is the centroid/medoid for cluster C_i.
- For each point, the error is the distance to the centroid/medoid.
- To get SSE, we square these errors and sum them:

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

- Compare the *SSE* for finding the best initial centroids when k (the number of clusters) is the same.

How to Choose K (Number of Clusters)?

The "elbow" method



► **Figure 16.8** Estimated minimal residual sum of squares as a function of the number of clusters in K-means. In this clustering of 1203 Reuters-RCV1 documents, there are two points where the $\widehat{\text{RSS}}_{\text{min}}$ curve flattens: at 4 clusters and at 9 clusters. The documents were selected from the categories *China*, *Germany*, *Russia* and *Sports*, so the $K = 4$ clustering is closest to the Reuters classification.

What If the Iteration Never Stops?

Set maximum number of iterations

Set minimum value of SSE change

Variations of the *K-Means* Method

One variation is the 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 is.

- 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"!