



CS 412 Intro. to Data Mining

Chapter 10. Cluster Analysis: Basic Concepts and Methods



Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

What Is Cluster Analysis?

- ❑ What is a cluster?
 - ❑ A cluster is a collection of data objects which are
 - ❑ Similar (or related) to one another within the same group (i.e., cluster)
 - ❑ Dissimilar (or unrelated) to the objects in other groups (i.e., clusters)
- ❑ Cluster analysis (or *clustering*, *data segmentation*, ...)
 - ❑ Given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible
- ❑ Cluster analysis is **unsupervised learning** (i.e., no predefined classes)
 - ❑ This contrasts with *classification* (i.e., *supervised learning*)
- ❑ Typical ways to use/apply cluster analysis
 - ❑ As a stand-alone tool to get insight into data distribution, or
 - ❑ As a preprocessing (or intermediate) step for other algorithms

5

What Is Good Clustering?

- ❑ A good clustering method will produce high quality clusters which should have
 - ❑ **High intra-class similarity:** **Cohesive** within clusters
 - ❑ **Low inter-class similarity:** **Distinctive** between clusters
- ❑ Quality function
 - ❑ There is usually a separate “quality” function that measures the “goodness” of a cluster
 - ❑ It is hard to define “similar enough” or “good enough”
 - ❑ The answer is typically highly subjective
- ❑ There exist many similarity measures and/or functions for different applications
- ❑ Similarity measure is critical for cluster analysis

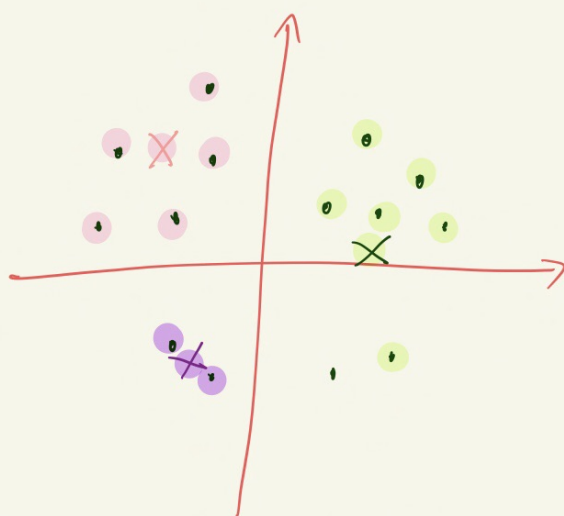
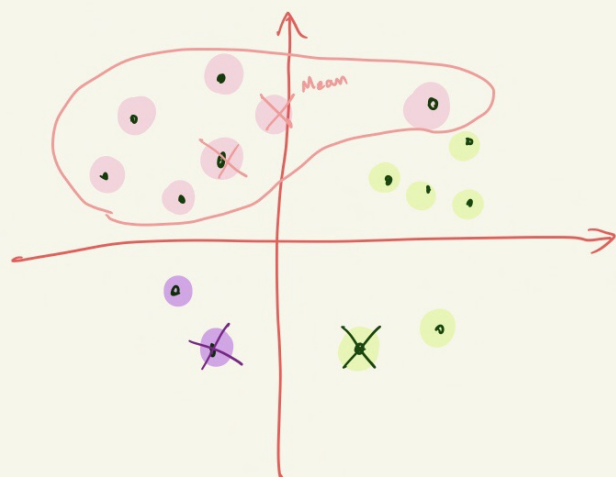
6

↑ จำนวน = 7 รก/ด ที่ไม่กำหนด
ในจ = ทมเพื่อนบ้านใกล้

The *K-Means* Clustering Method

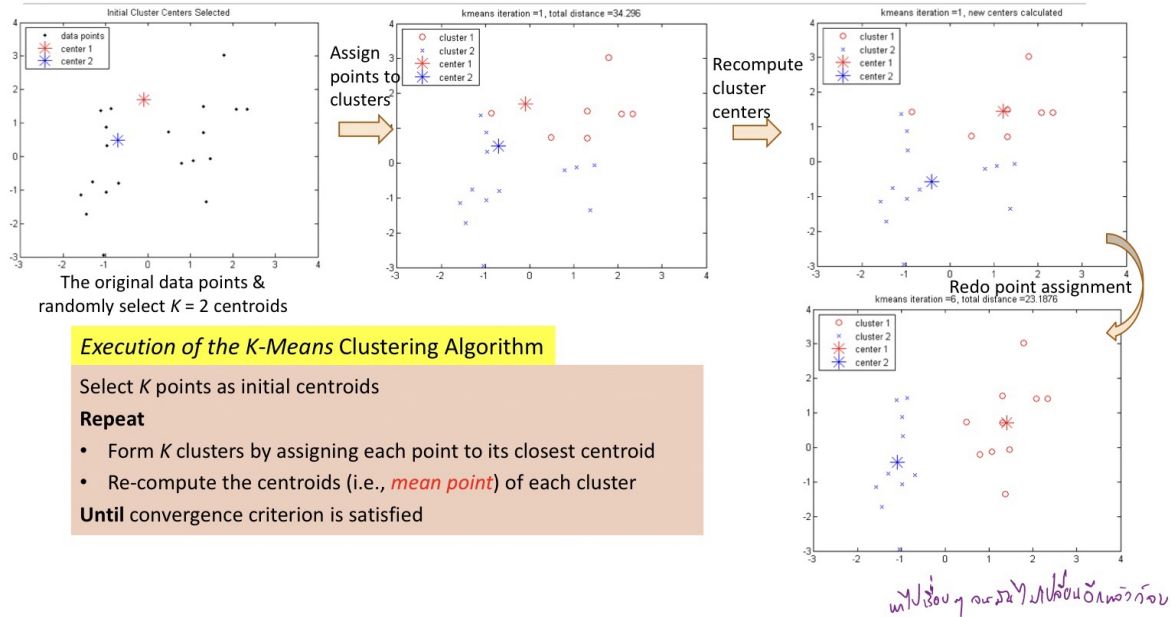
- *K-Means* (MacQueen'67, Lloyd'57/'82)
 - Each cluster is represented by the center of the cluster
- Given K , the number of clusters, the *K-Means* clustering algorithm is outlined as follows
กำหนด K ก่อน ว่ามีจุด = ๗ จุด อยู่
- Select K points as initial centroids → มี ๗ จุด ที่เลือกมา
เริ่มต้น
- Repeat ทำซ้ำ / ๗ รอบ ๆ จนกว่า จ = จุด ที่กำหนด
- Form K clusters by assigning each point to its closest centroid
 - Re-compute the centroids (i.e., *mean point*) of each cluster
- **Until** convergence criterion is satisfied
- Different kinds of measures can be used
 - Manhattan distance (L_1 norm), Euclidean distance (L_2 norm), Cosine similarity

$$K = 3$$



ทำไปเรื่อยๆ จน Mean ไม่เปลี่ยน
จุดเริ่มต้นเป็นสีน้ำเงิน

Example: *K*-Means Clustering



14

Discussion on the *K*-Means Method

- ❑ **Efficiency:** $O(tKn)$ where n : # of objects, K : # of clusters, and t : # of iterations
 - ❑ Normally, $K, t \ll n$; thus, an efficient method
- ❑ *K*-means clustering often **terminates at a local optimal**
 - ❑ Initialization can be important to find high-quality clusters
- ❑ **Need to specify K** , the *number* of clusters, in advance
 - ❑ There are ways to automatically determine the “best” K
 - ❑ In practice, one often runs a range of values and selected the “best” K value
- ❑ **Sensitive to noisy data and outliers**
 - ❑ Variations: Using *K*-medians, *K*-medoids, etc.
- ❑ *K*-means is applicable only to objects in a continuous n -dimensional space
 - ❑ Using the *K*-modes for **categorical data**
- ❑ Not suitable to discover clusters with **non-convex shapes**
 - ❑ Using density-based clustering, kernel *K*-means, etc.

15

Variations of *K-Means*

- There are many variants of the *K-Means* method, varying in different aspects

- ✓ □ Choosing better initial centroid estimates เลือกจุดเริ่มต้นที่ดี
 - *K-means++*, *Intelligent K-Means*, *Genetic K-Means* วิธีเลือกจุดเริ่มต้นที่ดี To be discussed in this lecture
- ✓ □ Choosing different representative prototypes for the clusters เลือกตัวแทนแทนจุดศูนย์กลาง
 - *K-Medoids*, *K-Medians*, *K-Modes* ใช้แทนจุดศูนย์กลาง To be discussed in this lecture
- ✓ □ Applying feature transformation techniques การแปลงคุณลักษณะ
 - *Weighted K-Means*, *Kernel K-Means* ใช้กับข้อมูลที่ไม่เป็นเชิงเส้น To be discussed in this lecture