CS 412 Intro. to Data Mining

Chapter 10. Cluster Analysis: Basic Concepts and Methods

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

k (ปกติ) คือ จำนวน... พารามิเตอร์

The *K-Means* Clustering Method

จน. cluster ที่เราต้องการ

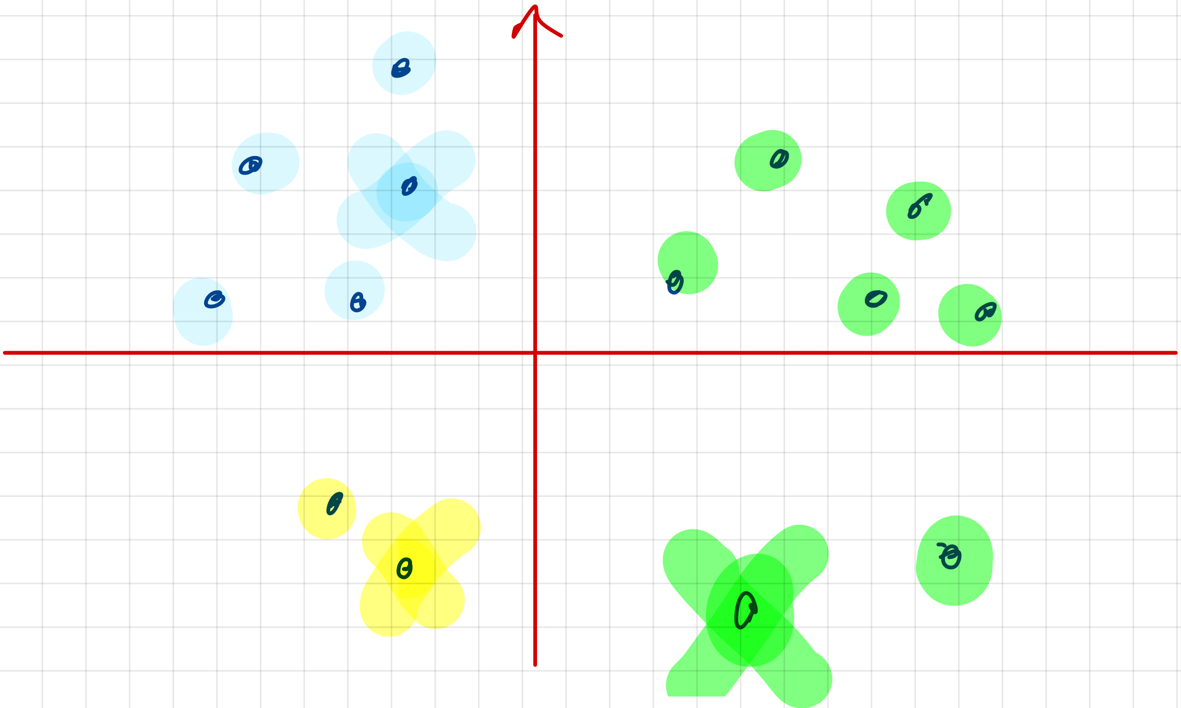
k-nearest → จน.เพื่อนบ้าน

k-item set → ของที่ซื้อพร้อมกัน 3 อัน (k=3)

- *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 → กำหนดค่า k (จะเอากี่กลุ่ม) k=3
  - Select *K* points as initial centroids สุ่มจุด 3 จุด (ตัวแทนของแต่ละกลุ่ม)
  - **Repeat** ทำซ้ำไปเรื่อยๆ จนกว่าจะพบ

    เข้าไปดูทีละตัว ตัวไหนใกล้จุด centroid ไหนมากที่สุด
    - Form *K* clusters by assigning each point to its closest centroid
    - Re-compute the centroids (i.e., *mean point*) of each cluster → จุดตัวแทนของกลุ่ม
  - **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



สีส้มใกล้กลุ่มสีฟ้า

ดูว่าจุดแต่ละจุดใกล้ centriod ไหน

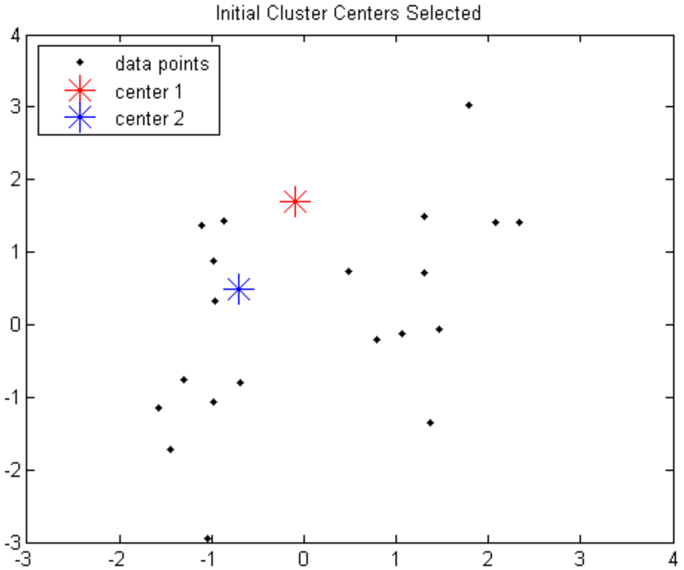
k = 3

เขียว

เหลือง

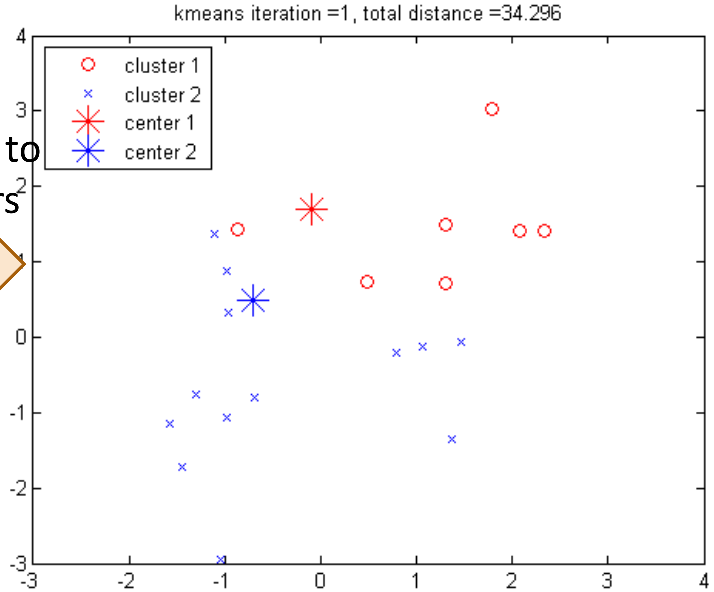
ฟ้า

Example: *K-Means* Clustering

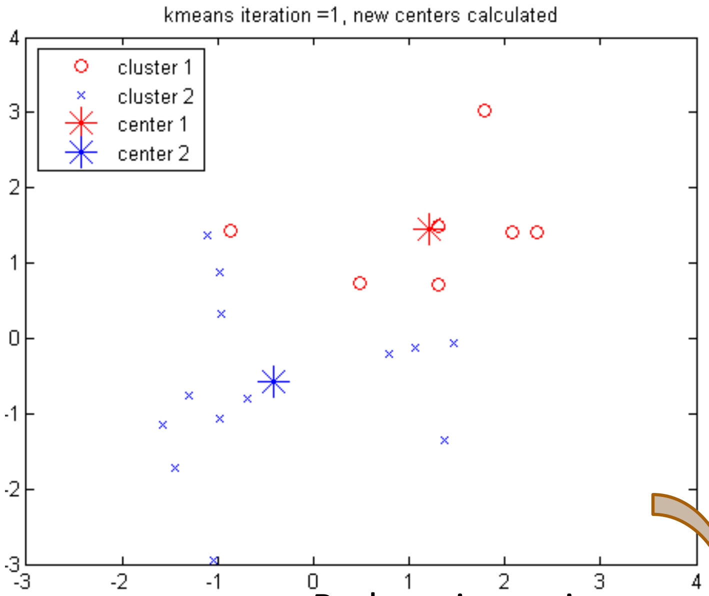


The original data points & randomly select *K* = 2 centroids

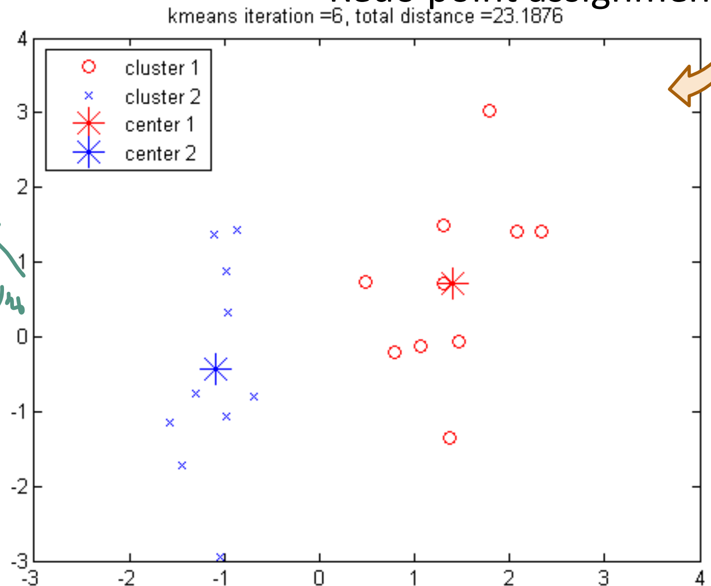
Assign points to clusters



Recompute cluster centers



Redo point assignment



จัดกลุ่มไป
เรื่อยๆจน
มันไม่เปลี่ยน
กลุ่มแล้ว

Execution of the K-Means **Clustering Algorithm**

Select *K* points as initial centroids

Repeat

- 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

Discussion on the *K-Means* Method

- **Efficiency**: $O(tKn)$ where n: # of objects, K: # of clusters, and t: # of iterations
  - Normally, K, t << 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.

Variations of *K-Means*

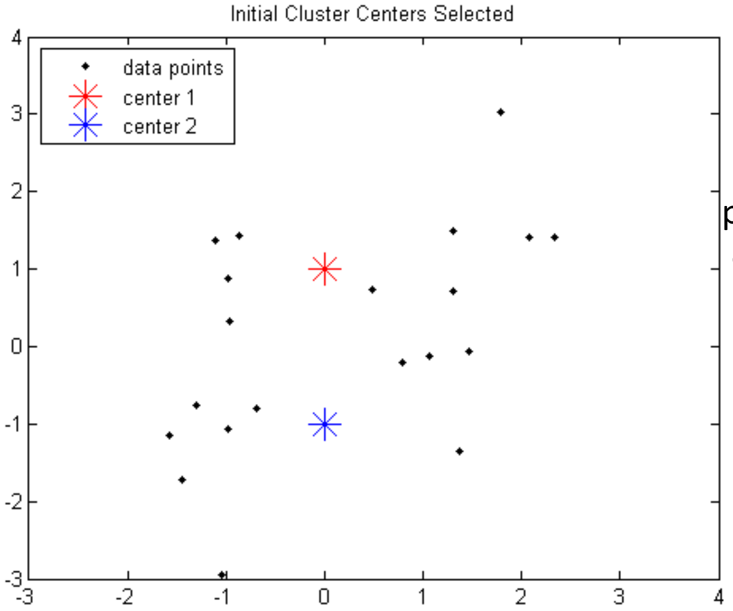
- There are many variants of the *K-Means* method, varying in different aspects
  - Choosing better initial centroid estimates
    - *K-means++, Intelligent K-Means, Genetic K-Means*

      To be discussed in this lecture
  - Choosing different representative prototypes for the clusters
    - *K-Medoids, K-Medians, K-Modes*

      To be discussed in this lecture
  - Applying feature transformation techniques
    - *Weighted K-Means, Kernel K-Means*

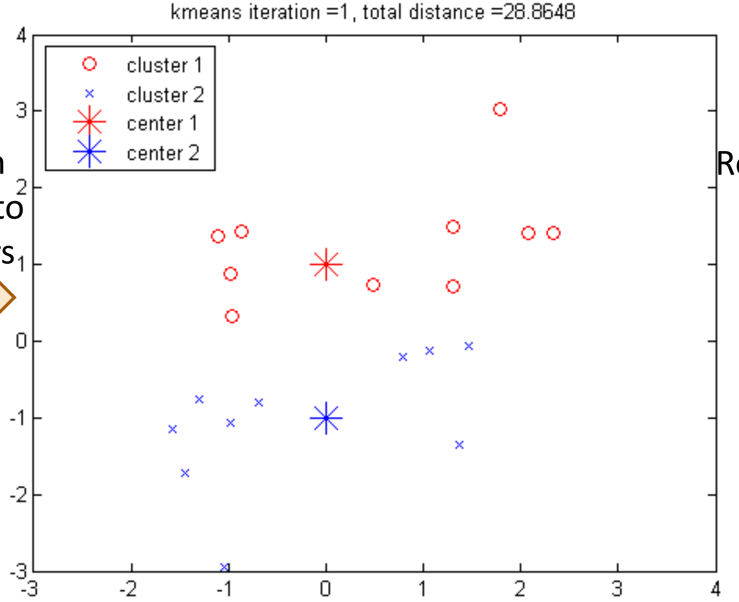
      To be discussed in this lecture

Poor Initialization in K-Means May Lead to Poor Clustering

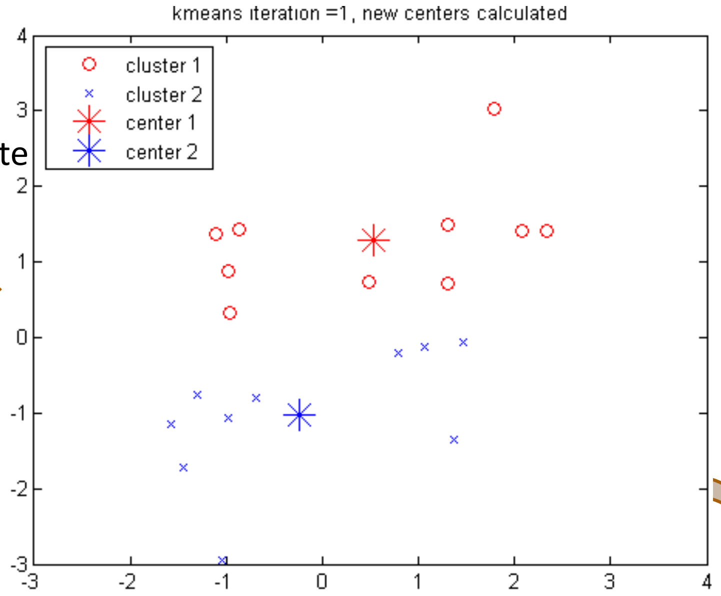


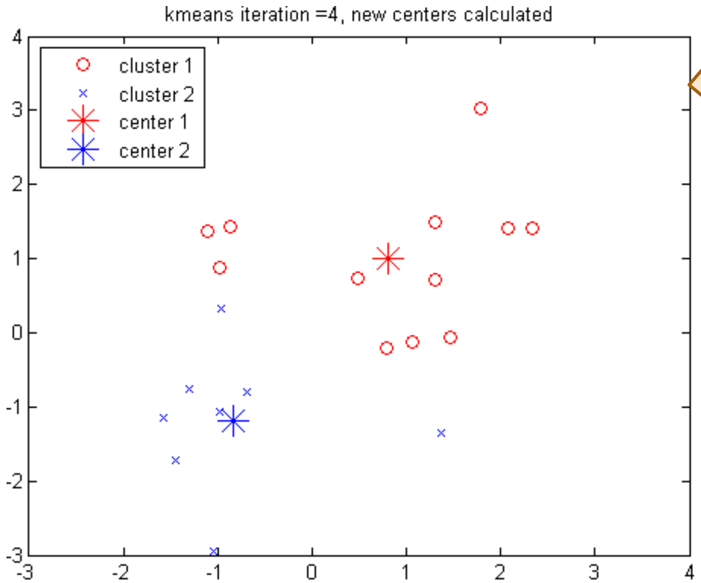
Another random selection of k
centroids for the same data points

Assign
points to
clusters



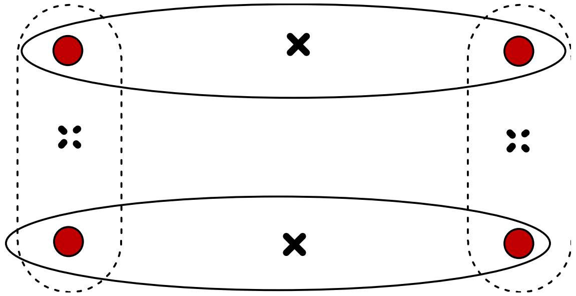
Recompute
cluster
centers





- Rerun of the *K-Means* using another random *K* seeds
- This run of *K*-Means generates a poor quality clustering

Initialization of K-Means: Problem and Solution

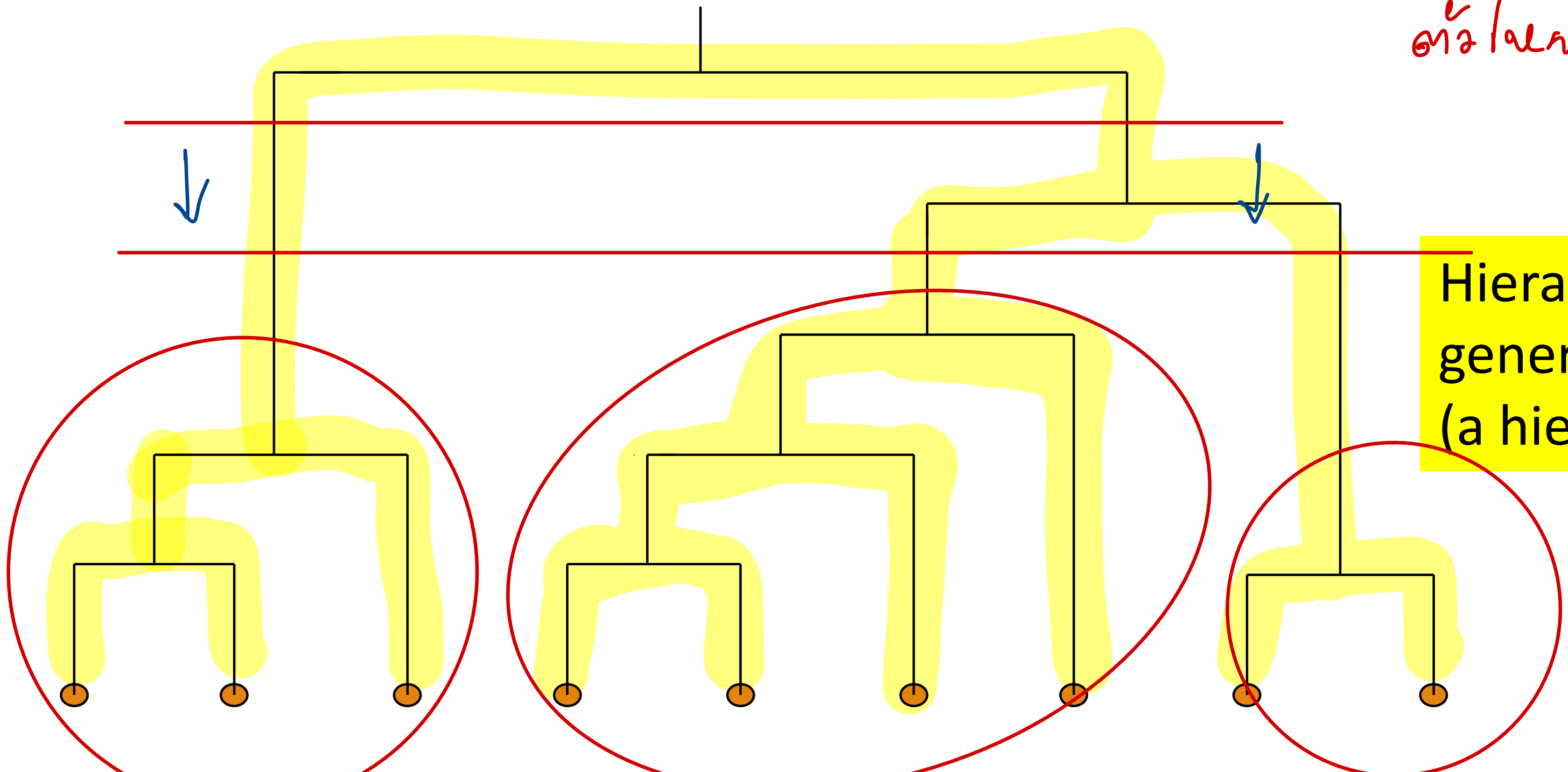


- Different initializations may generate rather different clustering results (some could be far from optimal)
- Original proposal (MacQueen'67): Select *K* seeds randomly
  - Need to run the algorithm multiple times using different seeds
- There are many methods proposed for better initialization of *k* seeds
  - ***K-Means++*** (Arthur & Vassilvitskii'07):
    - The first centroid is selected at random
    - The next centroid selected is the one that is farthest from the currently selected (selection is based on a weighted probability score)
    - The selection continues until *K* centroids are obtained

Dendrogram: Shows How Clusters are Merged

- Dendrogram: Decompose a set of data objects into a tree of clusters by multi-level nested partitioning
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster

ตัวไหนใกล้กันจับรวมกลุ่ม

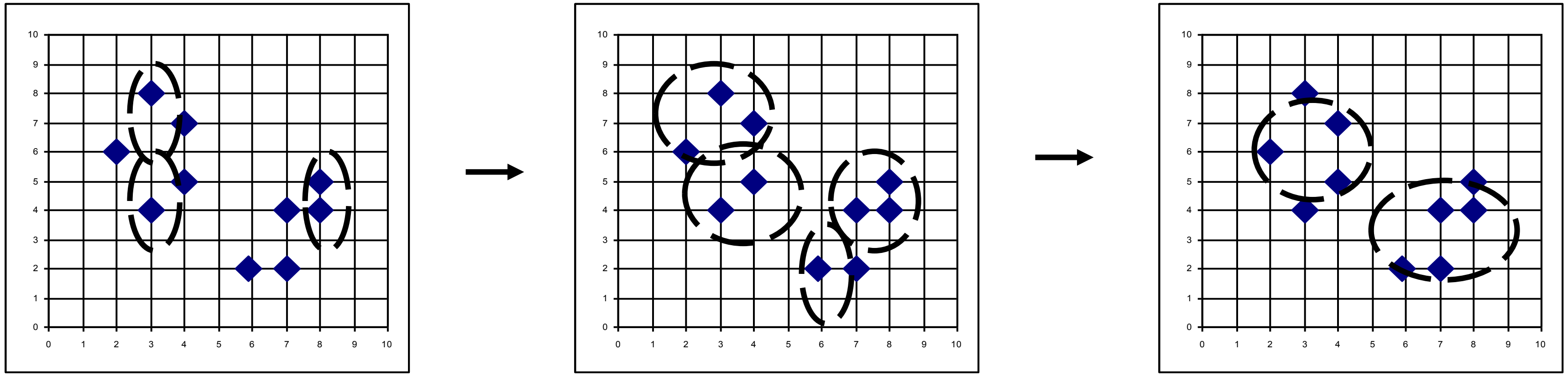


Hierarchical clustering generates a dendrogram (a hierarchy of clusters)

clustering ลดปน.
ทอ random

Agglomerative Clustering Algorithm

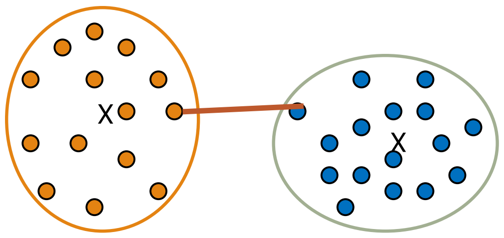
- AGNES (AGglomerative NESting) (Kaufmann and Rousseeuw, 1990)
  - Use the **single-link** method and the dissimilarity matrix
  - Continuously merge nodes that have the least dissimilarity
  - Eventually all nodes belong to the same cluster



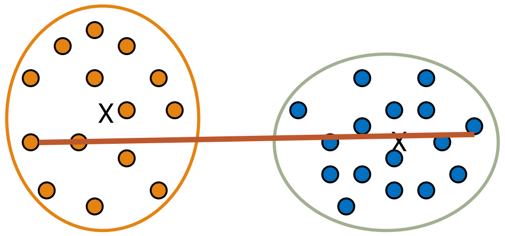
- Agglomerative clustering varies on different similarity measures among clusters
  - Single link (nearest neighbor)
  - Complete link (diameter)
  - Average link (group average)
  - Centroid link (centroid similarity)

Single Link vs. Complete Link in Hierarchical Clustering

- Single link (nearest neighbor)
  - The similarity between two clusters is the similarity between their most similar (nearest neighbor) members
  - Local similarity-based: Emphasizing more on close regions, ignoring the overall structure of the cluster
  - Capable of clustering non-elliptical shaped group of objects
  - Sensitive to noise and outliers

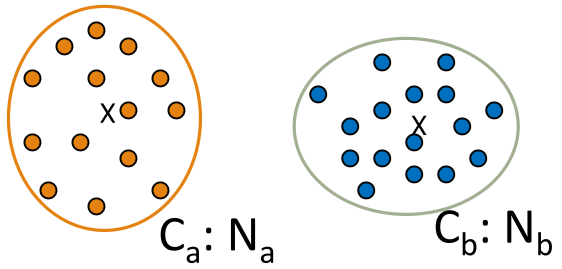


- Complete link (diameter)
  - The similarity between two clusters is the similarity between their most dissimilar members
  - Merge two clusters to form one with the smallest diameter
  - Nonlocal in behavior, obtaining compact shaped clusters
  - Sensitive to outliers

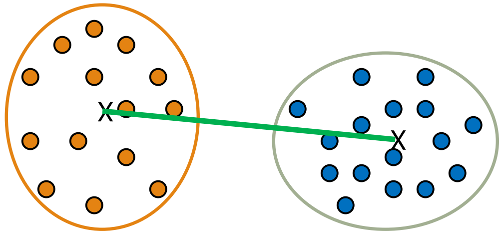


Agglomerative Clustering: Average vs. Centroid Links

- Agglomerative clustering with **average link**
  - **Average link**: The average distance between an element in one cluster and an element in the other (i.e., all pairs in two clusters)
    - Expensive to compute



- Agglomerative clustering with **centroid link**
  - **Centroid link**: The distance between the centroids of two clusters



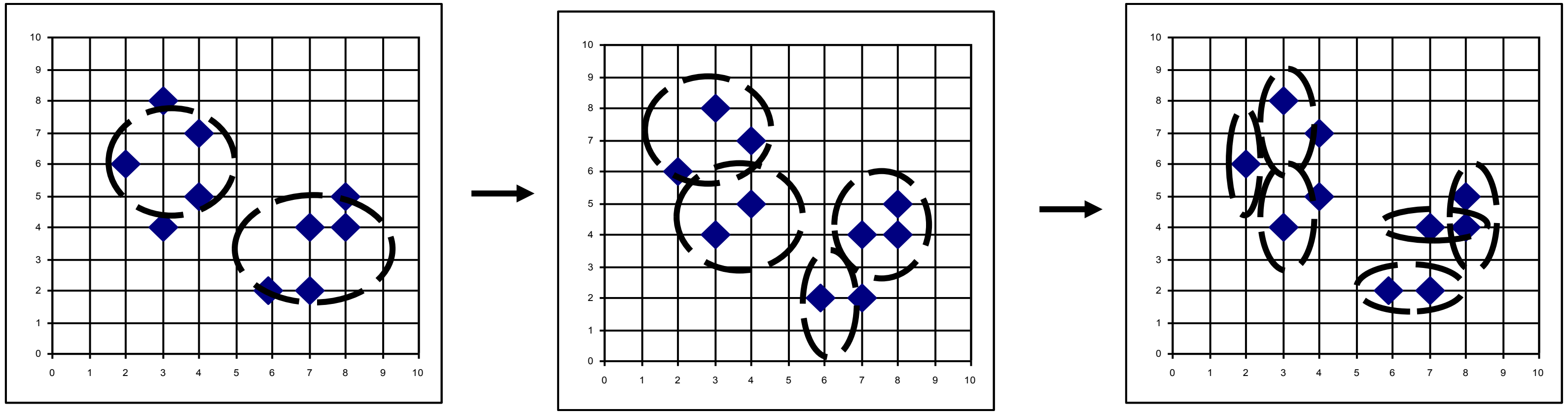
- **Group Averaged Agglomerative Clustering (GAAC)**
  - Let two clusters C_a and C_b be merged into $C_{a \cup b}$. The new centroid is:
    - N_a is the cardinality of cluster C_a, and c_a is the centroid of C_a

$$c_{a \cup b} = \frac{N_a c_a + N_b c_b}{N_a + N_b}$$

  - The similarity measure for GAAC is the average of their distances
- Agglomerative clustering with **Ward's criterion**
  - **Ward's criterion:** The increase in the value of the SSE criterion for the clustering obtained by merging them into $C_a \cup C_b$: $W(C_{a \cup b}, c_{a \cup b}) - W(C, c) = \frac{N_a N_b}{N_a + N_b} d(c_a, c_b)$

Divisive Clustering

- DIANA (Divisive Analysis) (Kaufmann and Rousseeuw,1990)
  - Implemented in some statistical analysis packages, e.g., Splus
- Inverse order of AGNES: Eventually each node forms a cluster on its own



- Divisive clustering is a top-down approach
  - The process starts at the root with all the points as one cluster
  - It recursively splits the higher level clusters to build the dendrogram
  - Can be considered as a global approach
  - More efficient when compared with agglomerative clustering

Clustering Validation

- Clustering Validation: Basic Concepts
- Clustering Evaluation: Measuring Clustering Quality
- External Measures for Clustering Validation
  - I: Matching-Based Measures
  - II: Entropy-Based Measures
  - III: Pairwise Measures
- Internal Measures for Clustering Validation
- Relative Measures
- Cluster Stability
- Clustering Tendency

Clustering Validation and Assessment

- Major issues on clustering validation and assessment
  - **Clustering evaluation** ประเมินค่า
    - Evaluating the goodness of the clustering
  - **Clustering stability** ค.คงที่ เช่น เปลี่ยนพารามิเตอร์นิดหน่อย ผลลัพธ์จะเปลี่ยนมากน้อยแค่ไหน
    - To understand the sensitivity of the clustering result to various algorithm parameters, e.g., # of clusters
  - **Clustering tendency** ค.เหมาะสมในการทำ clustering
    - Assess the suitability of clustering, i.e., whether the data has any inherent grouping structure

Measuring Clustering Quality

- **Clustering Evaluation**: Evaluating the goodness of clustering results
  - No commonly recognized best suitable measure in practice
- **Three categorization of measures**: External, internal, and relative
  - **External**: Supervised, employ criteria not inherent to the dataset
    - Compare a clustering against prior or expert-specified knowledge (i.e., the ground truth) using certain clustering quality measure
  - **Internal**: Unsupervised, criteria derived from data itself
    - Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient
  - **Relative**: Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm

Measuring Clustering Quality: External Methods

คำตอบที่แท้จริง

- Given the **ground truth** *T*, *Q*(*C*, *T*) is the **quality measure** for a clustering *C*
- *Q*(*C*, *T*) is good if it satisfies the following **four** essential criteria
- 1 **Cluster homogeneity** ไม่เอากลุ่มผสม
  - The purer, the better

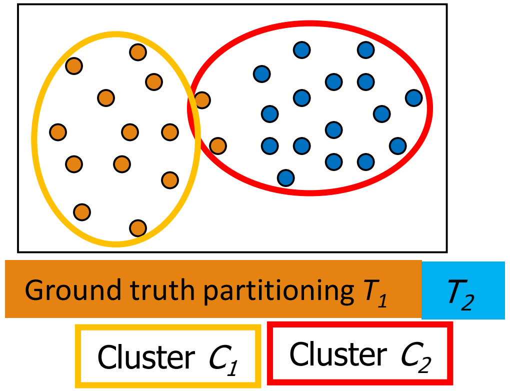
  C = (A A A A)(B A B A) X ไม่ควร

  (A A A A)(B B)(A A) ✓
- 2 **Cluster completeness** กลุ่มเดียวกัน ห้ามแตกกัน
  - Assign objects belonging to the same category in the ground truth to the same cluster
- 3 **Rag bag better than alien** วิธีให้คะแนน (กรณีที่ 1, 2 ควรเลือก กรณี 2)
  - Putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., "miscellaneous" or "other" category)
- 4 **Small cluster preservation** เราไม่ควรจัดเป็นกลุ่มเล็กๆมากเกินไป
  - Splitting a small category into pieces is more harmful than splitting a large category into pieces

Occames razor

Commonly Used External Measures

- Matching-based measures (To be covered)
  - Purity, maximum matching, F-measure
- Entropy-Based Measures
  - Conditional entropy (To be covered)
  - Normalized mutual information (NMI) (To be covered)
  - Variation of information
- Pairwise measures (To be covered)
  - Four possibilities: True positive (TP), FN, FP, TN
  - Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure
- Correlation measures
  - Discretized Huber static, normalized discretized Huber static



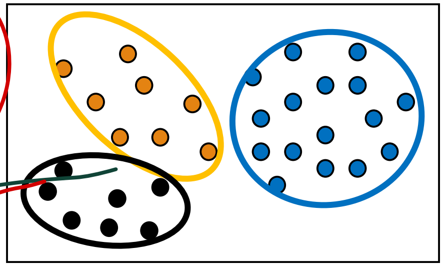
Internal Measures (I): BetaCV Measure

ภายใน Data เกาะกลุ่มกันมากกว่ากระจาย

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation

แยกสองกลุ่มควรห่างกัน

- Given a clustering $C = \{C_1, . . ., C_k\}$ with k clusters, cluster C_i containing $n_i = |C_i|$ points
  - Let $W(S, R)$ be sum of weights on all edges with one vertex in S and the other in R
  - The sum of all the intra-cluster weights over all clusters: $W_{in} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, C_i)$
  - The sum of all the inter-cluster weights: $W_{out} = \frac{1}{2}\sum_{i=1}^{k} W(C_i, \overline{C_i}) = \sum_{i=1}^{k-1}\sum_{j>i} W(C_i, C_j)$
  - The number of distinct intra-cluster edges: $N_{in} = \sum_{i=1}^{k}\binom{n_i}{2}$
  - The number of distinct inter-cluster edges: $N_{out} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} n_i n_j$



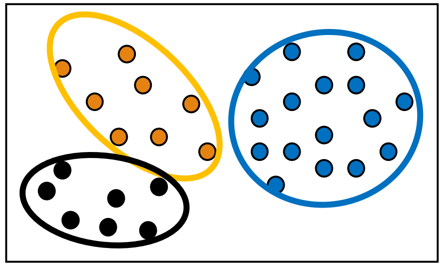
- **Beta-CV measure**: $BetaCV = \dfrac{W_{in} / N_{in}}{W_{out} / N_{out}}$
  - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
  - The smaller, the better the clustering

Internal Measures (II): Normalized Cut and Modularity

- **Normalized cut**: $$NC = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{vol(C_i)} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, \mathrm{V})} = \sum_{i=1}^{k} \frac{W(C_i, \overline{C_i})}{W(C_i, C_i) + W(C_i, \overline{C_i})} = \sum_{i=1}^{k} \frac{1}{\frac{W(C_i, C_i)}{W(C_i, \overline{C_i})} + 1}$$

  where $vol(C_i) = W(C_i, V)$ is the volume of cluster C_i

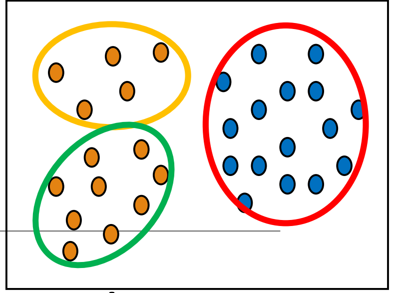
  - The higher normalized cut value, the better the clustering



- **Modularity** (for graph clustering) $$Q = \sum_{i=1}^{k} \left(\frac{W(C_i, C_i)}{W(V, V)} - \left(\frac{W(C_i, V)}{W(V, V)} \right)^2 \right)$$
  - Modularity Q is defined as

    where $$W(V, V) = \sum_{i=1}^{k} W(C_i, V) = \sum_{i=1}^{k} W(C_i, C_i) + \sum_{i=1}^{k} W(C_i, \overline{C_i}) = 2(W_{in} + W_{out})$$
  - Modularity measures the difference between the observed and expected fraction of weights on edges within the clusters.
  - The smaller the value, the better the clustering—the intra-cluster distances are lower than expected

Relative Measure



- Relative measure: Directly compare different clusterings, usually those obtained via different parameter settings for the same algorithm
- **Silhouette coefficient** as an **internal measure**: Check cluster cohesion and separation
  - For each point $\boldsymbol{x}_i$, its silhouette coefficient s_i is: $s_i = \dfrac{\mu_{out}^{\min}(\mathbf{x}_i) - \mu_{in}(\mathbf{x}_i)}{\max\{\mu_{out}^{\min}(\mathbf{x}_i), \mu_{in}(\mathbf{x}_i)\}}$

    where $\mu_{in}(\mathbf{x}_i)$ is the mean distance from $\boldsymbol{x}_i$ to points in its own cluster

    $\mu_{out}^{\min}(\mathbf{x}_i)$ is the mean distance from $\boldsymbol{x}_i$ to points in its closest cluster
  - Silhouette coefficient (*SC*) is the mean values of s_i across all the points: $SC = \dfrac{1}{n}\sum_{i=1}^{n} s_i$
  - *SC* close to +1 implies good clustering
    - Points are close to their own clusters but far from other clusters
- **Silhouette coefficient** as a **relative measure**: Estimate the # of clusters in the data

$$SC_i = \frac{1}{n_i}\sum_{x_j \in C_i} s_j$$

Pick the k value that yields the best clustering, i.e., yielding high values for *SC* and SC_i ($1 \le i \le k$)