The *K-Means* Clustering Method

จำนวน

- *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
  - Select *K* points as initial centroids
  - **Repeat** → ซ้ำ
    - 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

The *K-Means* Clustering Method

จำนวน.

- *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
  - Select *K* points as initial centroids
  - **Repeat** → ซ้ำ
    - 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

Variations of *K-Means*

- There are many variants of the *K-Means* method, varying in different aspects
  - Choosing better initial centroid estimates → สุ่มเลือก Centroid ให้มีประสิทธิภาพ
    - *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*

      จุดศูนย์ต่าง

      To be discussed in this lecture
  - Applying feature transformation techniques → ① เลือก k อย่างไร ? ② หาตัวแทนของกลุ่ม ③ ระยะทาง
    - *Weighted K-Means*, *Kernel K-Means*

      To be discussed in this lecture

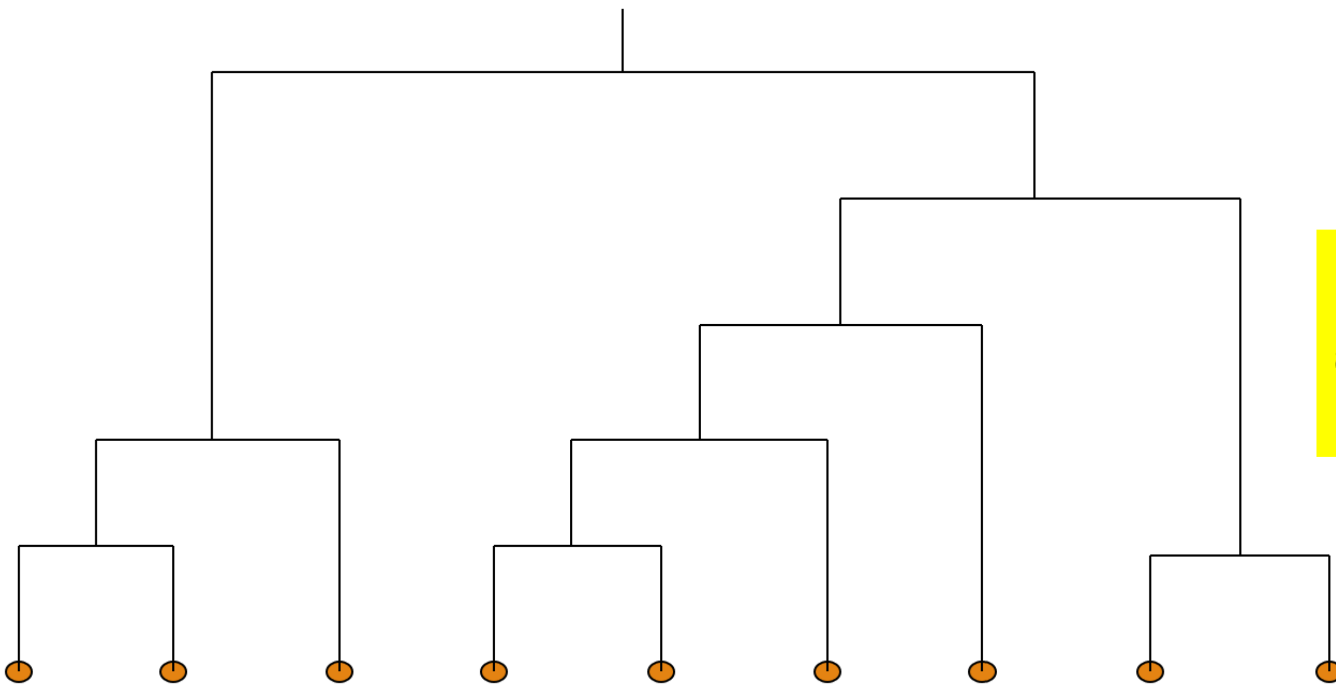
Hierarchical Clustering Methods

Clustering
แบบเป็นลำดับชั้น

- Basic Concepts of Hierarchical Algorithms
- Agglomerative Clustering Algorithms
- Divisive Clustering Algorithms
- Extensions to Hierarchical Clustering
- BIRCH: A Micro-Clustering-Based Approach
- CURE: Exploring Well-Scattered Representative Points
- CHAMELEON: Graph Partitioning on the KNN Graph of the Data
- Probabilistic Hierarchical Clustering

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



Hierarchical clustering generates a dendrogram (a hierarchy of clusters)

Clustering Validation and Assessment

- Major issues on clustering validation and assessment
  - **Clustering evaluation**
    - Evaluating the goodness of the clustering
  - **Clustering stability**
    - To understand the sensitivity of the clustering result to various algorithm parameters, e.g., # of clusters
  - **Clustering tendency**

    ความเหมาะสมของการทำ 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

  → เป็นคำตอบที่รู้อยู่แล้ว → นำคำตอบที่แท้จริงมาวัด *

  1. **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
  2. **Internal**: Unsupervised, criteria derived from data itself → วัดความดีของการแบ่งกลุ่ม
     - 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

ผลการclustering

- 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
  - **Cluster homogeneity** ⇒ กลุ่มที่ไม่มีตัวไม่เหมือนกันมาอยู่กลุ่มเดียวกัน
    - The purer, the better
  - **Cluster completeness** ⇒ กลุ่มที่เป็นตัวเดียวกัน ควรจะเป็นกลุ่มเดียวกัน
    - Assign objects belonging to the same category in the ground truth to the same cluster
  - **Rag bag better than alien** ⇒ วิธีให้คะแนน
    - 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)
  - **Small cluster preservation** ⇒ ไม่ควรจะแตกกลุ่มมากเกินไป
    - Splitting a small category into pieces is more harmful than splitting a large category into pieces

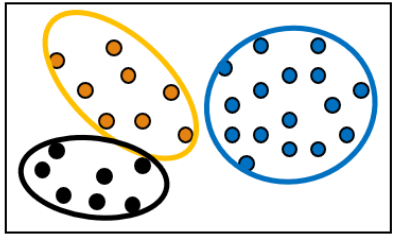
Internal Measures (I): BetaCV Measure

ควรห่างกัน

data ที่เกาะกลุ่มกัน

data ที่แยกกลุ่ม

- A trade-off in maximizing intra-cluster compactness and inter-cluster separation
- Given a clustering $C = \{C_1, \ldots, C_k\}$ 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**: $BetaCV = \frac{W_{in} / N_{in}}{W_{out} / N_{out}}$
  - The ratio of the mean intra-cluster distance to the mean inter-cluster distance
  - The smaller, the better the clustering