การทำคลัสเตอร์ริ่ง จะสนใจส่วนใหญ่ๆ อยู่ 3 อัน

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**
    - Assess the suitability of clustering, i.e., whether the data has any inherent grouping structure

Measuring Clustering Quality

วิธีวัด

- **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

วัดว่า อกอลิทึมไหนดี นำคำตอบที่แท้จริงมาวัด

เพื่อยืนยันว่า คำตอบที่เอามา จริง หรือเปล่า

วัด ความดี ของ การแบ่งกลุ่ม

Measuring Clustering Quality: External Methods

คำตอบที่แท้จริง ↗ สังเกตลักษณ์

- 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 Q คือ quality
  - **Cluster homogeneity**
    - The purer, the better C = (AAAA)(BABA)
  - **Cluster completeness** (AAAA)(BB)(AA) → วิธีแบบนี้จะดีกว่า ไม่ต้องการให้ตัวที่ไม่เหมือนมาอยู่ด้วยกัน
    - Assign objects belonging to the same category in the ground truth to the same cluster กลุ่มเดียวกันไม่ควรแตกกัน
  - **Rag bag better than alien**
    - 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** ไม่ควรจะตัดกลุ่มเล็กๆ มากเกินไป
    - Splitting a small category into pieces is more harmful than splitting a large category into pieces

Internal Measures (I): BetaCV Measure

- 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