# Cluster Analysis - Evaluation

# Cluster Validation

- **Cluster validation**
  - Quality: “goodness” of clusters
  - Assess the quality and reliability of clustering results
- **Why validation?**
  - To avoid finding clusters formed by chance
  - To compare clustering algorithms
  - To choose clustering parameters
    - e.g., the number of clusters

# Measures of cluster validity

- Numerical measures used to judge various aspects of cluster validity are classified into the following three types:
  - External index: Measures extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal index: Measures the “goodness” of a clustering structure *without* respect to external information.
    - Correlation
    - Cohesion and Separation
  - Relative index: Compares two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy.

# External Index

# Comparing to Ground Truth

- **Notation**
  - N: number of objects in the data set
  - $P=\{P_1,\ldots,P_s\}$: the set of “ground truth” clusters
  - $C=\{C_1,\ldots,C_t\}$: the set of clusters reported by a clustering algorithm
- **The “incidence matrix”**
  - $N \times N$ (both rows and columns correspond to objects)
  - $P_{ij} = 1$ if $O_i$ and $O_j$ belong to the same “ground truth” cluster in $P$; $P_{ij}=0$ otherwise
  - $C_{ij} = 1$ if $O_i$ and $O_j$ belong to the same cluster in $C$; $C_{ij}=0$ otherwise

# Rand Index and Jaccard Coefficient

- A pair of data object $(O_i, O_j)$ falls into one of the following categories
  - SS: $C_{ij}=1$ and $P_{ij}=1$; (agree)
  - DD: $C_{ij}=0$ and $P_{ij}=0$; (agree)
  - SD: $C_{ij}=1$ and $P_{ij}=0$; (disagree)
  - DS: $C_{ij}=0$ and $P_{ij}=1$; (disagree)
- **Rand index** $$Rand = \frac{|Agree|}{|Agree| + |Disagree|} = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|}$$
  - may be dominated by DD
- **Jaccard Coefficient** $$Jaccard\ \ coefficien\,t = \frac{|SS|}{|SS| + |SD| + |DS|}$$

Clustering

| | g 1 | g 2 | g 3 | g 4 | g 5 |
|---|---|---|---|---|---|
| g 1 | 1 | 1 | 1 | 0 | 0 |
| g 2 | 1 | 1 | 1 | 0 | 0 |
| g 3 | 1 | 1 | 1 | 0 | 0 |
| g 4 | 0 | 0 | 0 | 1 | 1 |
| g 5 | 0 | 0 | 0 | 1 | 1 |

Groundtruth

| | g 1 | g 2 | g 3 | g 4 | g 5 |
|---|---|---|---|---|---|
| g 1 | 1 | 1 | 0 | 0 | 0 |
| g 2 | 1 | 1 | 0 | 0 | 0 |
| g 3 | 0 | 0 | 1 | 1 | 1 |
| g 4 | 0 | 0 | 1 | 1 | 1 |
| g 5 | 0 | 0 | 1 | 1 | 1 |

Clustering

| Ground truth | | Same Cluster | Different Cluster |
|---|---|---|---|
| | Same Cluster | 9 | 4 |
| | Different Cluster | 4 | 8 |

$$Rand = \frac{|SS| + |DD|}{|SS| + |SD| + |DS| + |DD|} = \frac{17}{25}$$

$$Jaccard = \frac{|SS|}{|SS| + |SD| + |DS|} = \frac{9}{17}$$

# Entropy and Purity

- **Notation**
  - $|C_k \cap P_j|$ the number of objects in both the *k*-th cluster of the clustering solution and *j*-th cluster of the groundtruth
  - $|C_k|$ the number of objects in the *k*-th cluster of the clustering solution
  - $|P_j|$ the number of objects in the *j*-th cluster of the groundtruth
- **Purity** $$Purity = \frac{1}{N}\sum_k \max_j |C_k \cap P_j|$$
- **Normalized Mutual Information**

$$NMI = \frac{I(C,P)}{\sqrt{H(C)H(P)}} \qquad I(C,P) = \sum_k \sum_j \frac{|C_k \cap P_j|}{N} \log \frac{N \cdot |C_k \cap P_j|}{|C_k||P_j|}$$

$$H(C) = \sum_k \frac{|C_k|}{N} \log \frac{|C_k|}{N} \qquad H(P) = \sum_j \frac{|P_j|}{N} \log \frac{|P_j|}{N}$$

# Example

| | P 1 | P 2 | P 3 | P 4 | P5 | P6 | Total |
|---|---|---|---|---|---|---|---|
| C1 | 3 | 5 | 40 | 506 | 96 | 27 | 677 |
| C 2 | 4 | 7 | 280 | 29 | 39 | 2 | 361 |
| C 3 | 1 | 1 | 1 | 7 | 4 | 671 | 685 |
| C 4 | 10 | 162 | 3 | 119 | 73 | 2 | 369 |
| C 5 | 331 | 22 | 5 | 70 | 13 | 23 | 464 |
| C 6 | 5 | 358 | 12 | 212 | 48 | 13 | 648 |
| total | 354 | 555 | 341 | 943 | 273 | 738 | 3204 |

$$Purity = \frac{1}{N}\sum_k \max_j |C_k \cap P_j|$$

$$Purity = \frac{506+280+671+162+331+358}{3204}$$
$$= 0.7203$$

$$NMI = \frac{I(C,P)}{\sqrt{H(C)H(P)}} \qquad I(C,P) = \sum_k \sum_j \frac{|C_k \cap P_j|}{N} \log \frac{N \cdot |C_k \cap P_j|}{|C_k||P_j|}$$

$$H(C) = \sum_k \frac{|C_k|}{N} \log \frac{|C_k|}{N} \qquad H(P) = \sum_j \frac{|P_j|}{N} \log \frac{|P_j|}{N}$$

# Internal Index

# Internal Index

- “Ground truth” may be unavailable
- Use only the data to measure cluster quality
  - Measure the “*cohesion*” and “*separation*” of clusters
  - Calculate the *correlation* between clustering results and distance matrix

# Cohesion and Separation

- **Cohesion** is measured by the within cluster sum of squares

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

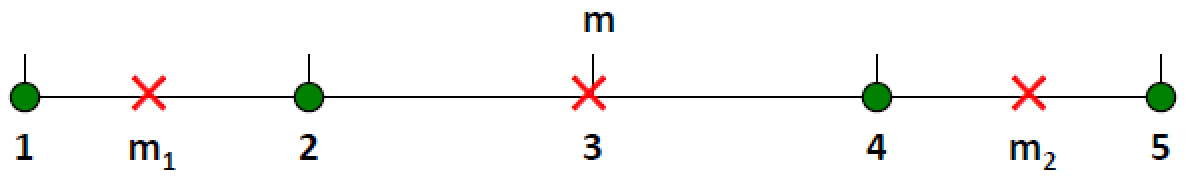
- **Separation** is measured by the between cluster sum of squares

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

where $|C_i|$ is the size of cluster $i$, m is the centroid of the whole data set

- BSS + WSS = constant
- WSS (Cohesion) measure is called Sum of Squared Error (SSE)—a commonly used measure
- A larger number of clusters tend to result in smaller SSE

# Example




**K=1 :**

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$BSS = 4 \times (3-3)^2 = 0$$
$$Total = 10 + 0 = 10$$

**K=2 :**

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$
$$Total = 1 + 9 = 10$$

**K=4:**

$$WSS = (1-1)^2 + (2-2)^2 + (4-4)^2 + (5-5)^2 = 0$$
$$BSS = 1 \times (1-3)^2 + 1 \times (2-3)^2 + 1 \times (4-3)^2 + 1 \times (5-3)^2 = 10$$
$$Total = 0 + 10 = 10$$

# Correlation with Distance Matrix

- Distance Matrix
  - $D_{ij}$ is the similarity between object $O_i$ and $O_j$
- Incidence Matrix
  - $C_{ij}$=1 if $O_i$ and $O_j$ belong to the same cluster, $C_{ij}$=0 otherwise
- Compute the correlation between the two matrices
  - Only $n(n-1)/2$ entries needs to be calculated
- High correlation indicates good clustering

# Correlation with Distance Matrix

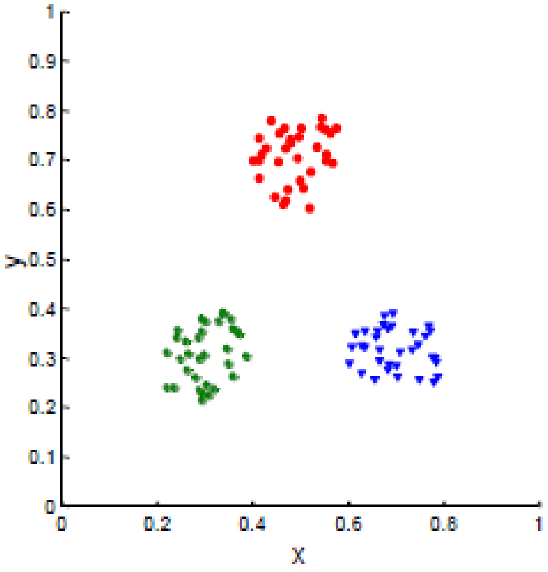
- Given Distance Matrix D = $\{d_{11}, d_{12}, \ldots, d_{nn}\}$ and Incidence Matrix $C = \{ c_{11}, c_{12}, \ldots, c_{nn}\}$ .

- Correlation $r$ between $D$ and $C$ is given by

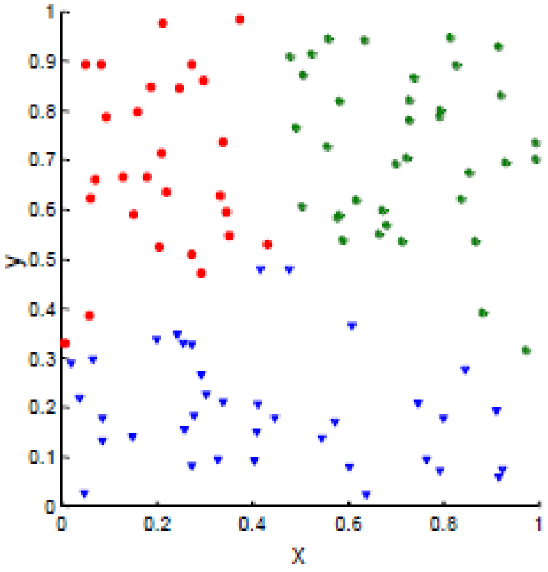
$$r = \frac{\sum_{i=1, j=1}^{n} (d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i=1, j=1}^{n} (d_{ij} - \bar{d})^2} \sqrt{\sum_{i=1, j=1}^{n} (c_{ij} - \bar{c})^2}}$$

# Measuring Cluster Validity Via Correlation

- Correlation of incidence and distance matrices for the K-means clusterings of the following two data sets



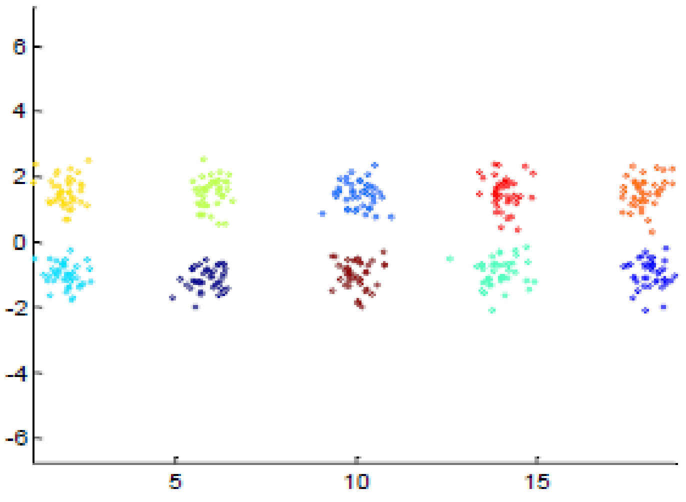

Corr = -0.9235



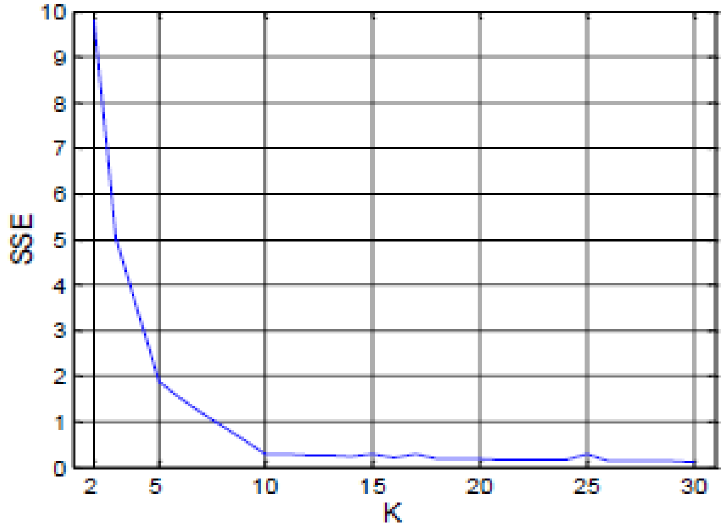

Corr = -0.5810

# Determine the Number of Clusters Using SSE

- SSE curve




Clustering of Input Data




SSE wrt K