L7-2

Cluster Validity

Seungchan Kim

Center for Computational Systems Biology

Electrical and Computer Engineering

Clustering Validation

- For supervised classification we have a variety of measures to evaluate how good our model is
    - AUC, accuracy, precision, recall, F1, …

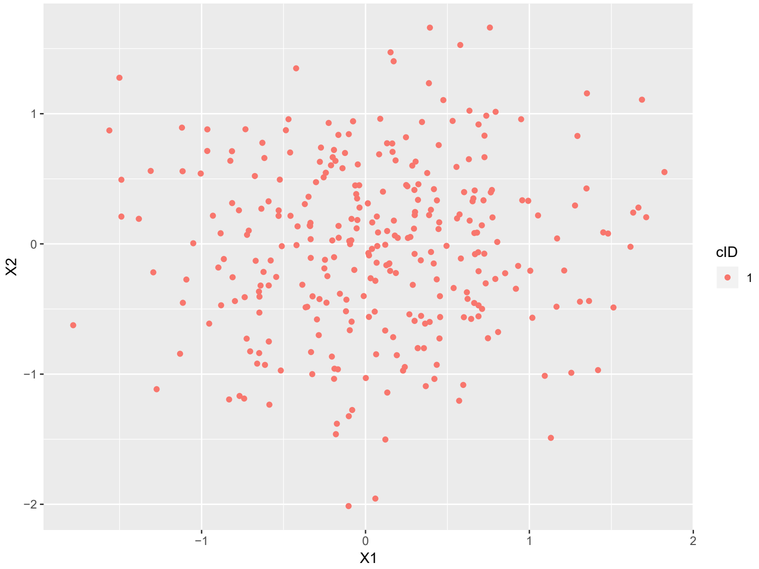
- For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters?

Why Clustering Validation

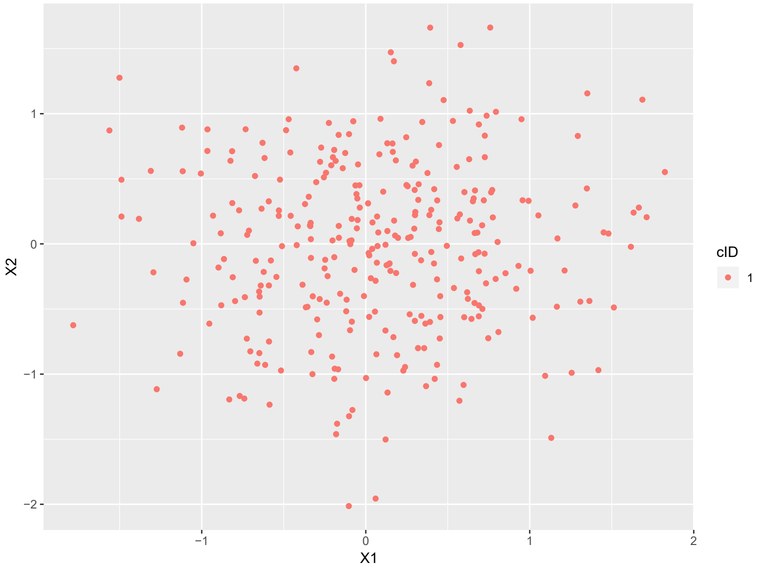
- “Clusters are in the eye of the beholder”!

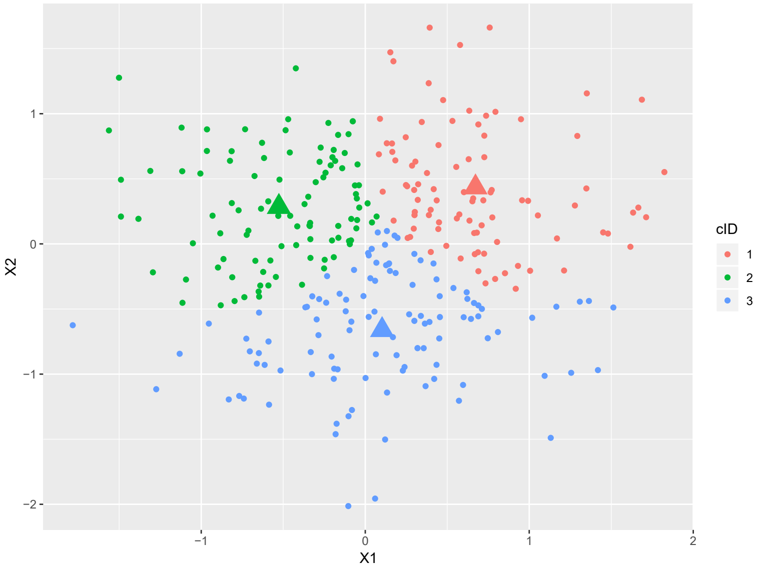
- Then why do we want to evaluate them?
    - To avoid finding patterns in noise
    - To compare clustering algorithms
    - To compare two sets of clusters
    - To compare two clusters

k-means Clustering: Case 1

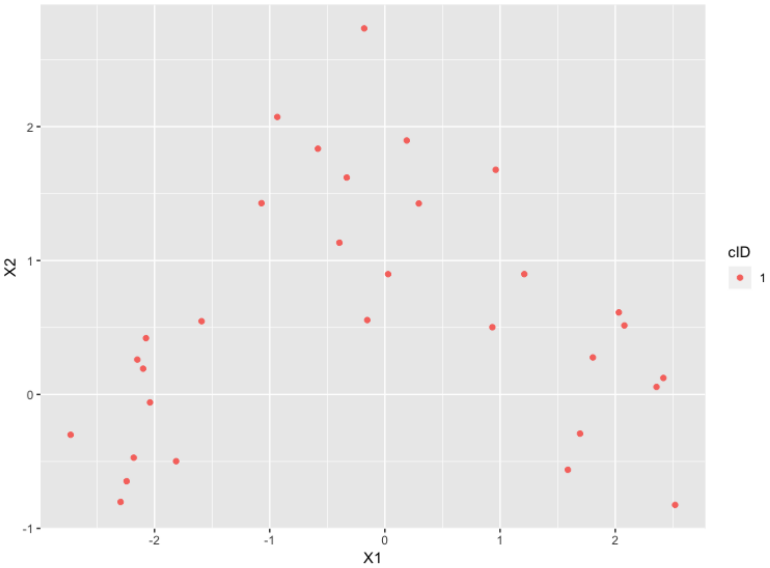


k-means Clustering: Case 1

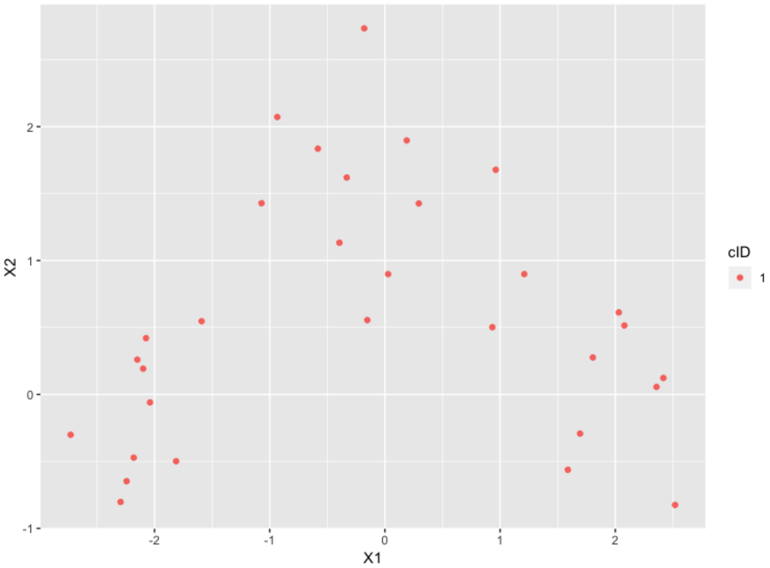


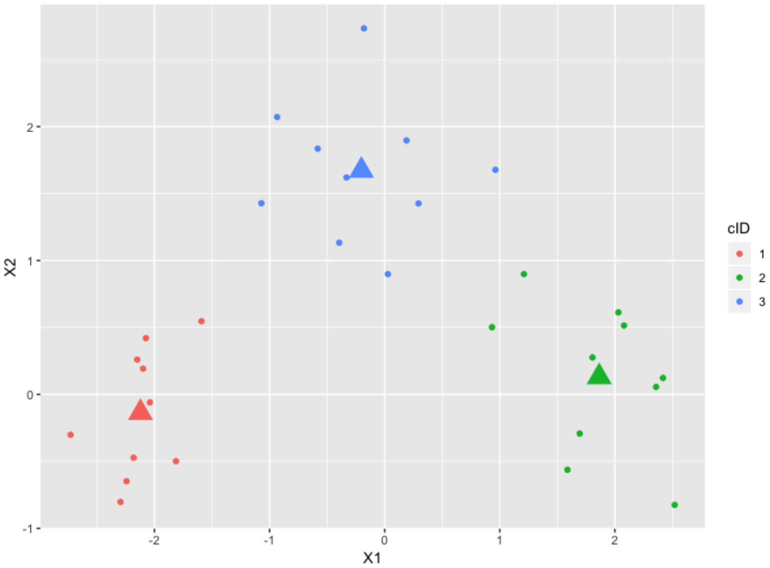


k-means Clustering: Case 2



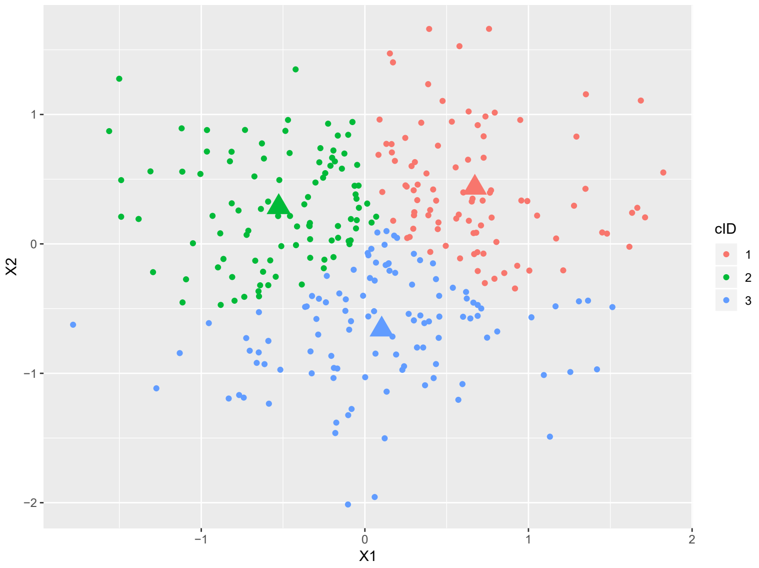
k-means Clustering: Case 2

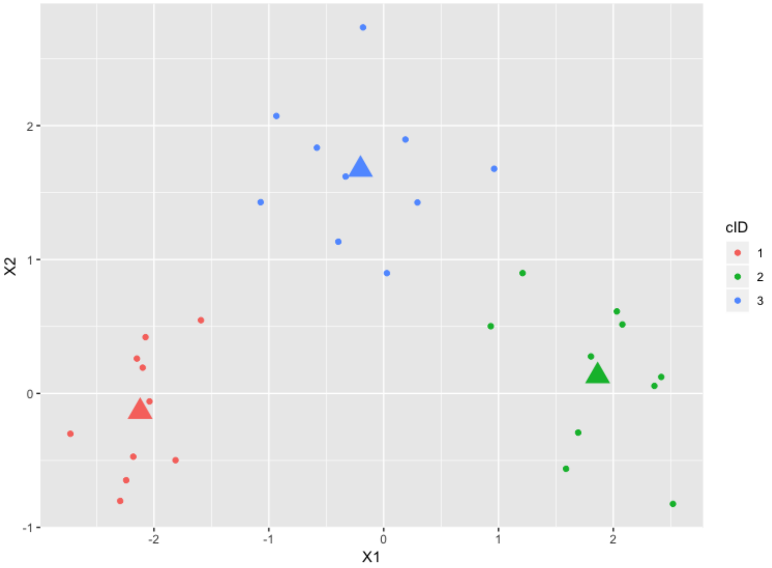




Comparing Sets of Clusters

- Which one is better?





Measures of Cluster Validity

- Numerical measures to judge various aspects of cluster validity

- **Internal Index** measures the goodness of a clustering structure without respect to external information
  - Sum of Squared Error (SSE)
- **External Index** measures the extent to which cluster labels match externally supplied class labels
  - Entropy
- **Relative Index** compares two sets of clustering or clusters
  - Often an external or internal index is used for this function, e.g., SSE or entropy

Internal Measures: SSE

- Sum of Squared Error (**SSE**) measures how closely related objects are in a cluster

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

where m_i is a centroid of a cluster C_i.

- SSE is also known as the within cluster sum of squares (**WSS**)

Internal Measures: Cohesion and Separation

- Cluster Cohesion measures how closely related objects are in a cluster, and the within-cluster sum of square (WSS) can be used to quantify it.

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

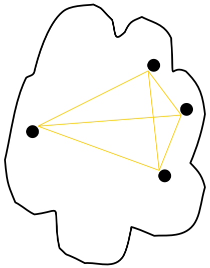
- Cluster Separation measures how distinct or well-separated a cluster is from other clusters, and the between-cluster sum of squares (BSS) can be used to quantify it.

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

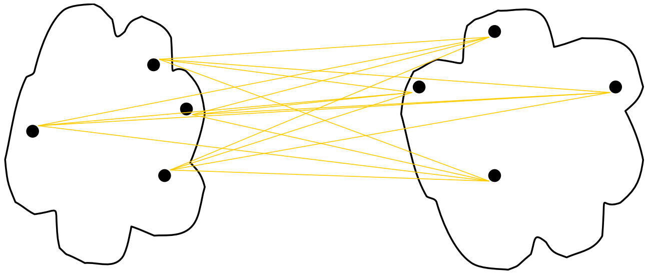
where $|C_i|$ is the size of cluster C_i and m is the centroid of all the samples.

Internal Measures: Cohesion and Separation

- A proximity graph-based approach can be also used to measure cohesion and separation.
  - Cluster cohesion is the sum of the distances of all links within a cluster.
  - Cluster separation is the sum of the distances between nodes in the cluster and nodes outside the cluster.



cohesion



separation

Internal Measures: Silhouette Coefficient

- Silhouette Coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and a set of clusters.
- For an individual sample x_i,

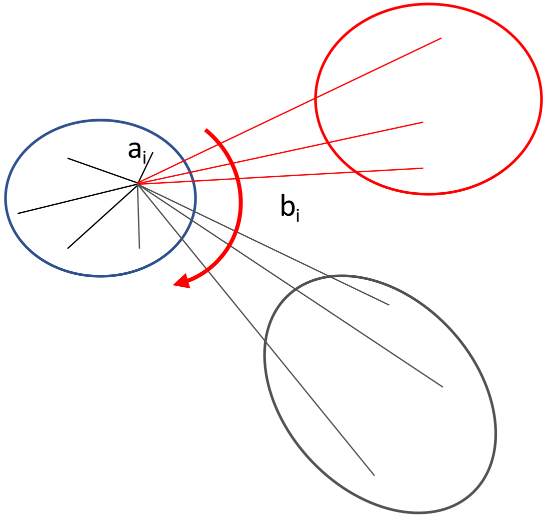
$$a_i = \begin{cases} \frac{1}{|C_I| - 1} \sum_{x \in C_I, x \neq x_i} \|x_i - x\| & \text{if } |C_I| > 1 \\ 0 & \text{if } |C_I| = 1 \end{cases}$$

where $|C_I|$ is the number of samples in the cluster C_I.

$$b_i = \min_{J \neq I} \frac{1}{|C_J|} \sum_{y \in C_J} \|x_i - y\|$$

then, Silhouette Coefficient s_i,

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$



Simplified Silhouette Coefficient

- Instead of computing a_i and b_i with all pair-wise samples, simplified silhouette coefficient can be computed:

$$a_i' = \|x_i - c_I\| \text{ and } b_i' = \min_{C_I \neq C_I} \|x_i - c_J\|$$

$$s_i' = \frac{b_i' - a_i'}{\max\{a_i', b_i'\}}$$

Silhouette Coefficient

- $-1 < s_i < 1$
  - the closer to 1, the better the belongness
  - If $s_i < 0$, there exist a better cluster s_i should be assigned to.
- Silhouette width
  - For cluster C_I,

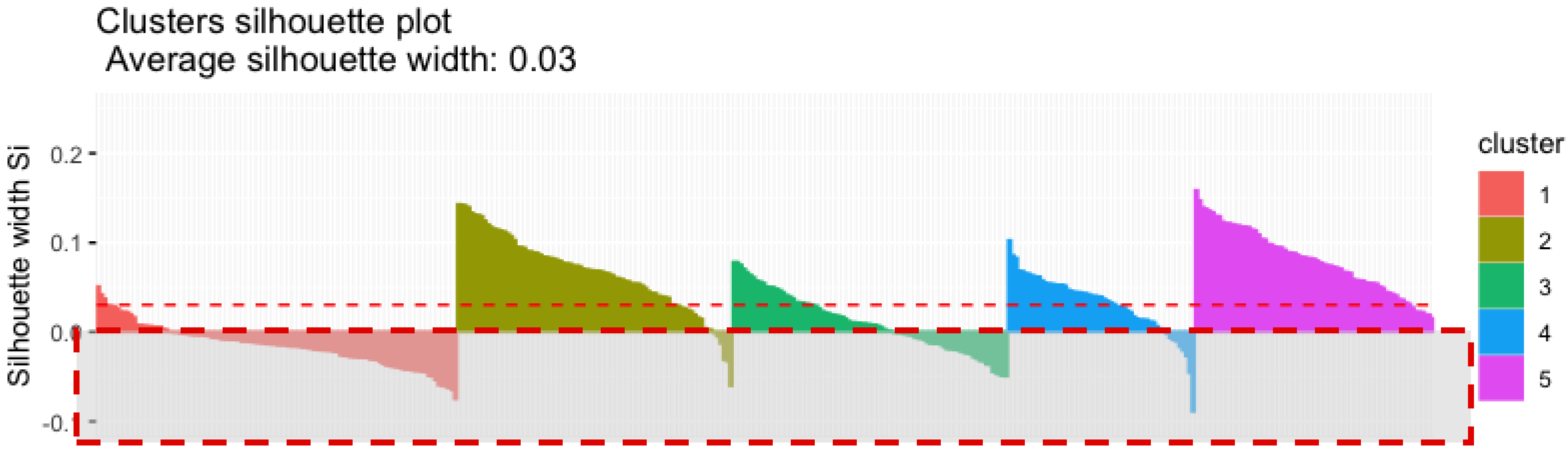
$$\bar{S_I} = \frac{1}{|C_I|}\sum_{C_I} s_i$$

  - For overall clusters

$$\bar{S} = \frac{1}{N}\sum_{\forall C_I} s_i$$

Average Silhouette Coefficient

- Compute Silhouette coefficient s_{ik} for each sample i for each k
- Compute average of s_{ik} for each k, $\bar{s}_k$



Incorrectly assigned

External Measures: Entropy

- **Entropy**: For each cluster, the class *i* distribution of the data is calculated for cluster *j*, p_{ij} is the probability that a member of cluster j belongs to class i.
  - Then the entropy of each cluster j,

$$e_j = -\sum_{i}^{L} p_{ij} \log_2 p_{ij}$$

  - The total entropy for a set of clusters is $e = \sum_j \frac{m_j}{m} e_j$ where m_j is the number of samples in cluster j and m is the total number of samples.

External Measures: Purity

- **Purity**: the purity of cluster *j*,

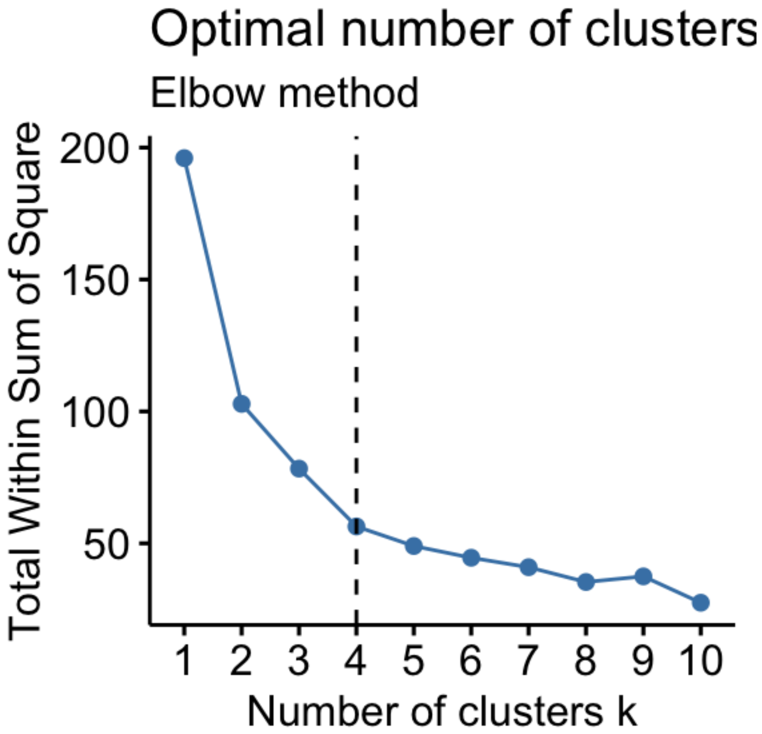
$$\text{purity}_j = \max p_{ij}$$

and the overall purity of a clustering,

$$\text{purity} = \sum_j \frac{m_i}{m_j} \text{purity}_j$$

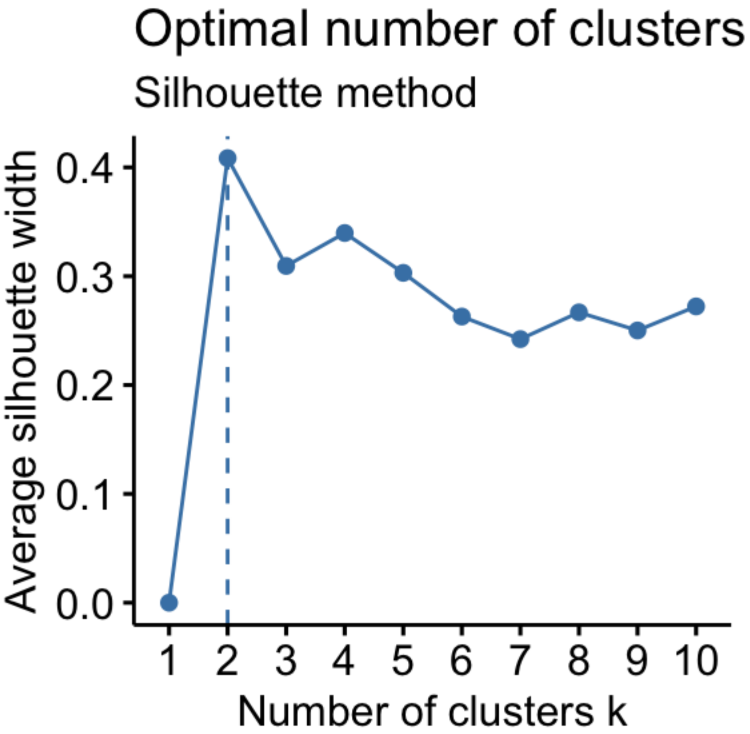
Determine ”Optimal” Number of Clusters

- Elbow method
  - Perform clustering with $k = 2, \ldots, K_{\max}$
  - Compute SSE/WSS for each k
  - Find k where WSS starts to stabilize, hence, ”**elbow**”



Determine ”Optimal” Number of Clusters

- Silhouette method
  - Perform clustering with k = 2, …, K_{max}
  - Compute Silhouette coefficient s_{ik} for each sample i for each k
  - Compute average of s_{ik} for each k, $\bar{s}_k$
  - Pick k where $\bar{s}_k$ peaks.



Consensus Clustering

- Motivation:
  - Assess and Improve the “stability” of discovered clusters
- Assumption:
  - If the data represent a sample of items drawn from distinct sub-populations, and if we were to observe a different sample drawn from the same sub-populations, the induced cluster composition and number should not be radically different.
  - Therefore, **the more the attained clusters are robust to sampling variability, the more we can be confident that these clusters represent real structure.**
- Method:
  - Iteration of clustering with resampling
  - Summarize the results as a **Consensus Matrix**.

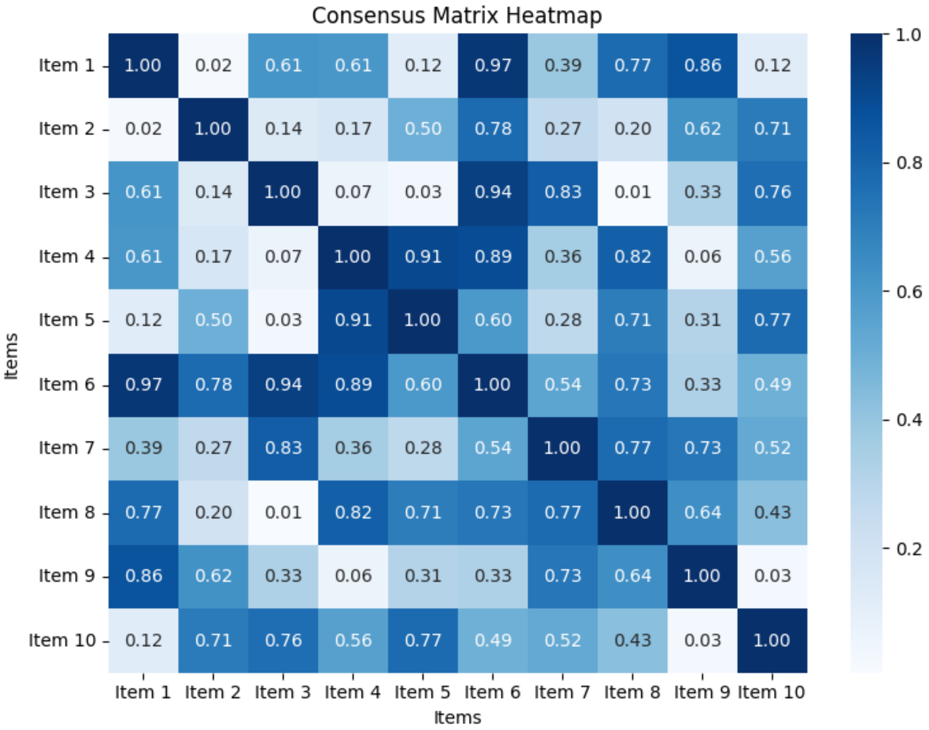
Consensus Matrix

- Let $M^{(h)}$ denote the (N x N) matrix representing a clustering result by applying a clustering algorithm to a resampled data set of $D^{(h)}$ where:
  - $M^{(h)}(i,j) = \begin{cases} 1 & \text{if item } i \text{ and } j \text{ belong to the same cluster} \\ 0 & \text{otherwise} \end{cases}$
- Consensus Matrix, M

$$M(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)}$$

where $I^{(h)}$ is the (NxN) indicator matrix such that its (*i*,*j*)-th entry equals to 1 if both items *i* and *j* are present in the dataset $D^{(h)}$

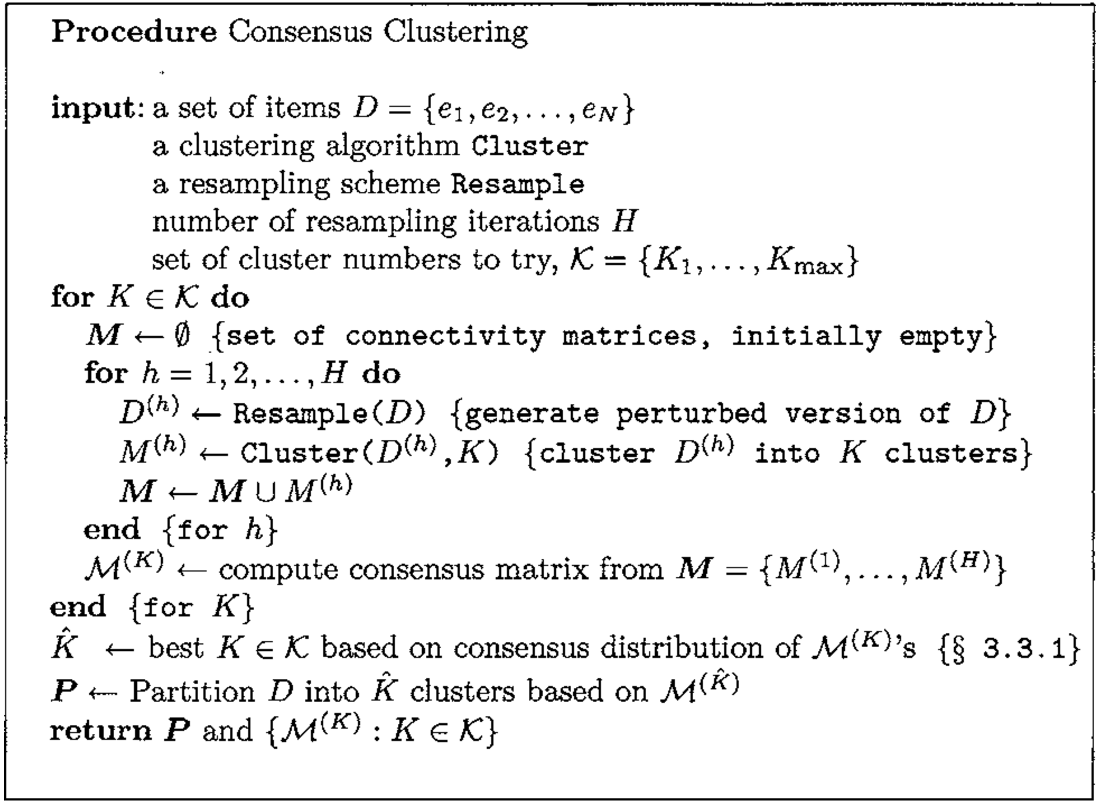
Consensus Matrix



Consensus Matrix Heatmap

Items	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Item 1	1.00	0.02	0.61	0.61	0.12	0.97	0.39	0.77	0.86	0.12
Item 2	0.02	1.00	0.14	0.17	0.50	0.78	0.27	0.20	0.62	0.71
Item 3	0.61	0.14	1.00	0.07	0.03	0.94	0.83	0.01	0.33	0.76
Item 4	0.61	0.17	0.07	1.00	0.91	0.89	0.36	0.82	0.06	0.56
Item 5	0.12	0.50	0.03	0.91	1.00	0.60	0.28	0.71	0.31	0.77
Item 6	0.97	0.78	0.94	0.89	0.60	1.00	0.54	0.73	0.33	0.49
Item 7	0.39	0.27	0.83	0.36	0.28	0.54	1.00	0.77	0.73	0.52
Item 8	0.77	0.20	0.01	0.82	0.71	0.73	0.77	1.00	0.64	0.43
Item 9	0.86	0.62	0.33	0.06	0.31	0.33	0.73	0.64	1.00	0.03
Item 10	0.12	0.71	0.76	0.56	0.77	0.49	0.52	0.43	0.03	1.00

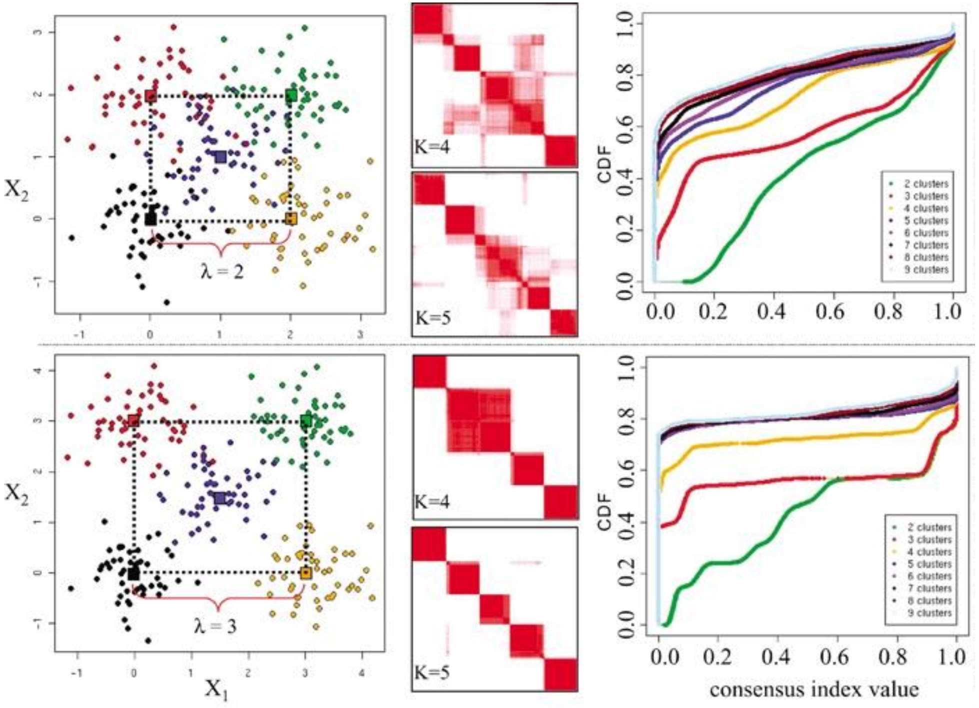
Consensus Clustering



```
Procedure Consensus Clustering

input: a set of items $D = \{e_1, e_2, \ldots, e_N\}$
       a clustering algorithm Cluster
       a resampling scheme Resample
       number of resampling iterations $H$
       set of cluster numbers to try, $\mathcal{K} = \{K_1, \ldots, K_{\max}\}$
for $K \in \mathcal{K}$ do
  $\boldsymbol{M} \leftarrow \emptyset$ {set of connectivity matrices, initially empty}
  for $h = 1, 2, \ldots, H$ do
    $D^{(h)} \leftarrow$ Resample($D$) {generate perturbed version of $D$}
    $M^{(h)} \leftarrow$ Cluster($D^{(h)}$,$K$) {cluster $D^{(h)}$ into $K$ clusters}
    $\boldsymbol{M} \leftarrow \boldsymbol{M} \cup M^{(h)}$
  end {for $h$}
  $\mathcal{M}^{(K)} \leftarrow$ compute consensus matrix from $\boldsymbol{M} = \{M^{(1)}, \ldots, M^{(H)}\}$
end {for $K$}
$\hat{K} \leftarrow$ best $K \in \mathcal{K}$ based on consensus distribution of $\mathcal{M}^{(K)}$'s {§ 3.3.1}
$\boldsymbol{P} \leftarrow$ Partition $D$ into $\hat{K}$ clusters based on $\mathcal{M}^{(\hat{K})}$
return $\boldsymbol{P}$ and $\{\mathcal{M}^{(K)} : K \in \mathcal{K}\}$
```

Consensus Clustering: Example



Reading

- **Introduction to Statistical Learning with Applications in R**
  - http://www-bcf.usc.edu/~gareth/ISL/
- **Cluster Analysis: Basic Concepts and Algorithms:**
  - https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf
- Monti, S., Tamayo, P., Mesirov, J., Golub, T. (2003) **Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data**. *Machine Learning*, 52, 91–118.
  - https://link.springer.com/article/10.1023/A:1023949509487
  - https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
- Tibshirani, R., Walther, G., Hastie, T. (2001) **Estimating the number of clusters in a data set via the gap statistic**. *J. R. Statist. Soc. B* 63, Part 2, pp. 411-423.
  - https://doi.org/10.1111/1467-9868.00293