

Clustering Evaluation

Javier Béjar

UL - 2026 Spring Term

CS - MAI

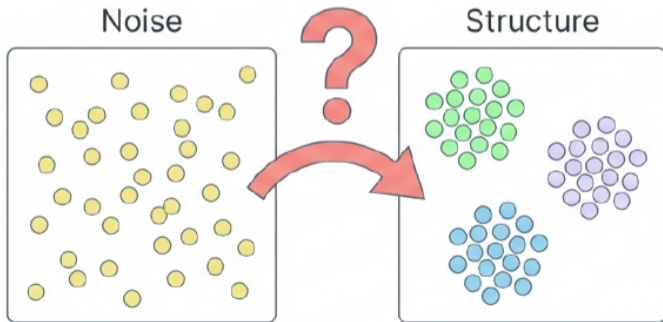


Cluster Evaluation

- ⊙ The **evaluation** of unsupervised learning is **difficult**
- ⊙ There is no goal model to compare with
- ⊙ The true result is unknown, it **may depend on the context**, the task to perform. . .
- ⊙ Why do we want to evaluate them?
 - To avoid finding patterns in noise
 - To compare clustering algorithms
 - To compare different models/parameters

- ⊙ **Cluster tendency**, there are clusters in the data?
- ⊙ **Compare** the clusters to the **true partition** of the data
- ⊙ **Quality** of the clusters without reference to external information
- ⊙ **Compare** the results of **different** clustering **algorithms**
- ⊙ **Evaluate** algorithm **parameters**
 - For instance, to determine the *correct* number of clusters

- ⊙ Before clustering a dataset we can test if there are actually clusters
- ⊙ We have to test the hypothesis of the existence of patterns in the data versus a dataset uniformly distributed (homogeneous distribution)

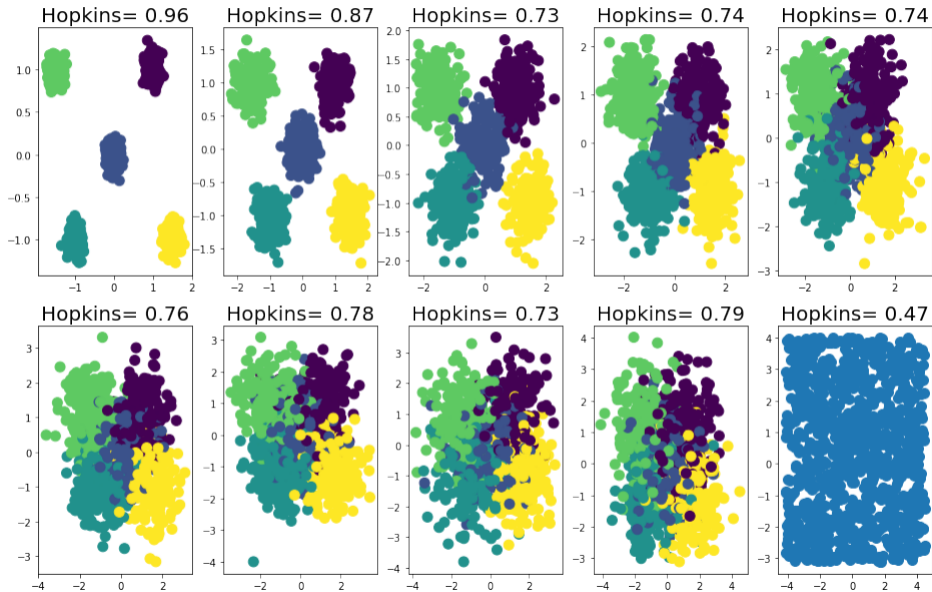


⊙ Hopkins Statistic

1. Sample n points (p_i) from the dataset (D) uniformly and compute the distance to their nearest neighbour ($d(p_i)$)
2. Generate n points (q_i) uniformly distributed in the space of the dataset and compute their distance to their nearest neighbours in D ($d(q_i)$)
3. Compute the quotient:

$$H = \frac{\sum_{i=1}^n d(p_i)}{\sum_{i=1}^n d(p_i) + \sum_{i=1}^n d(q_i)}$$

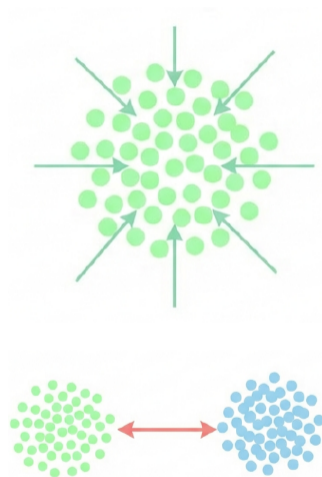
4. If data are uniformly distributed the value of H will be around 0.5



- ⊙ We can use different methodologies/criterion to evaluate the quality of a clustering:
 - **External criteria:** Comparison with a model partition/labelled data
 - **Internal criteria:** Quality measures based on the examples/quality of the partition

Internal criteria

- ⊙ Measure properties expected in a good clustering: Compact groups/ Well separated groups
- ⊙ We can use indices based on the attributes values measuring the properties of a good clustering
- ⊙ These indices are based on statistical properties of the attributes of the model
 - Values distribution
 - Distances distribution



- ⊙ Some indices correspond directly to the objective function optimized:
 - Quadratic error/Distortion (k-means)

$$SSE = \sum_{k=1}^k \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2$$

- Log likelihood (Mixture of gaussians/EM)

- ⊙ **Scatter matrices:** Interclass distance, intraclass distance, separation

$$S_{W_k} = \sum_{\forall x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \quad (\text{Compactness})$$

$$S_{B_k} = |C_k|(\mu_k - \mu)(\mu_k - \mu)^T \quad (\text{Separation})$$

$$S_{M_{k,l}} = \sum_{\forall i \in C_k} \sum_{\forall j \in C_l} (x_i - x_j)(x_i - x_j)^T$$

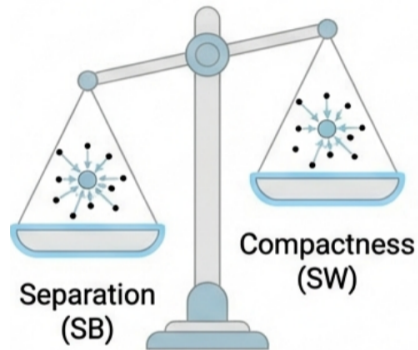
- ⊙ **Trace criteria** (lower overall intraclass distance/higher overall interclass distance)

$$\text{Tr}(S_W) = \frac{1}{K} \sum_{i=1}^K S_{W_k} \quad \text{Tr}(S_B) = \frac{1}{K} \sum_{i=1}^K S_{B_k}$$

- Interclass-intracluster distance ratio

$$CH = \frac{\sum_{i=0}^K |C_i| \times \|\mu_i - \mu\|^2 / (K - 1)}{\sum_{k=1}^K \sum_{i=0}^{|C_i|} \|x_i - \mu_i\|^2 / (N - K)}$$

- Ratio between $Tr(S_B)$ and $Tr(S_W)$ corrected by the number of clusters, **higher** is better
- Biased towards convex clusters, degrades with noise



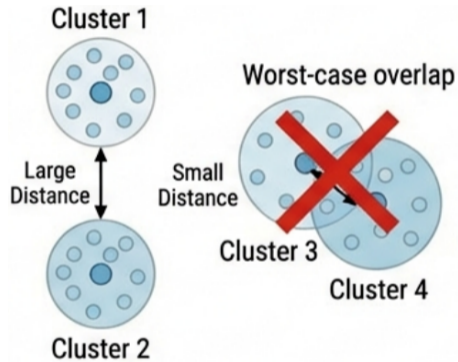
- Maximum interclass-intracluster distance ratio

$$\bar{R} = \frac{1}{K} \sum_{i=1}^K R_i$$

where

$$R_{ij} = \frac{S_{W_i} + S_{W_j}}{S_{M_{ij}}} \quad ; \quad R_i = \max_{j:j \neq i} R_{ij}$$

- Average of the S_W and S_M using the worst clusters pairings, we are measuring possible split clusters, **lower is better**



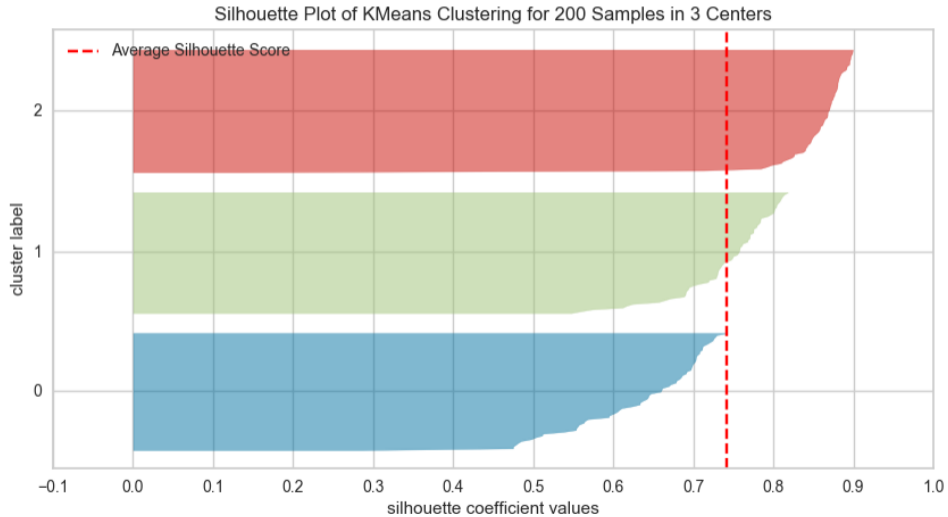
- ⊙ Maximum class spread/variance

$$S = \frac{1}{K} \sum_{i=0}^K \frac{b_i - a_i}{\max(a_i, b_i)}$$

$$a_i = \frac{1}{|C_i| - 1} \sum_{y \in C_i, x \in C_i, y \neq x} \|y - x\|$$

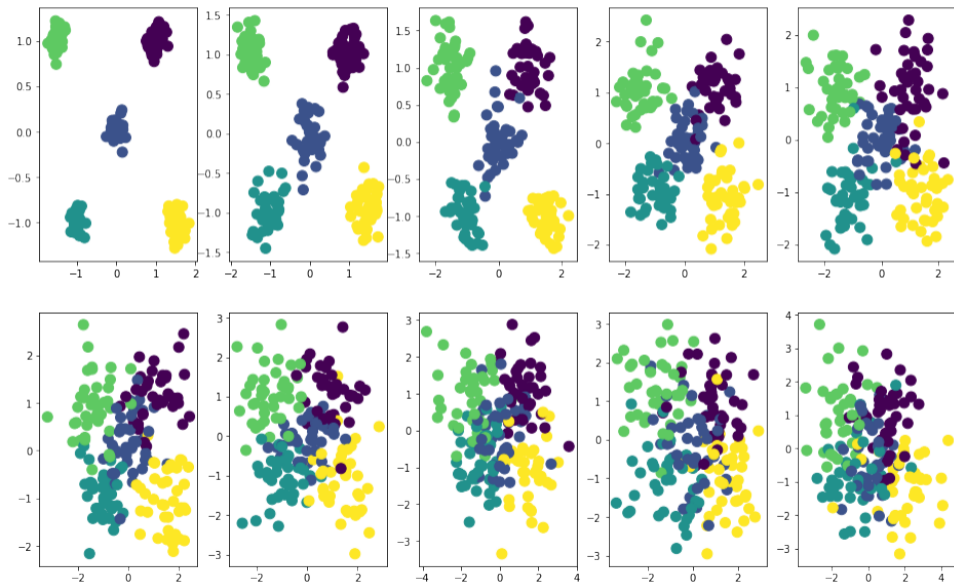
$$b_i = \min_{l \in H, l \neq i} \frac{1}{|C_l|} \sum_{y \in C_l, x \in C_i} \|y - x\| \quad H = \{h : 1 \leq h \leq K\}$$

- ⊙ We are looking for clusters with samples too close to other clusters, **higher is better**

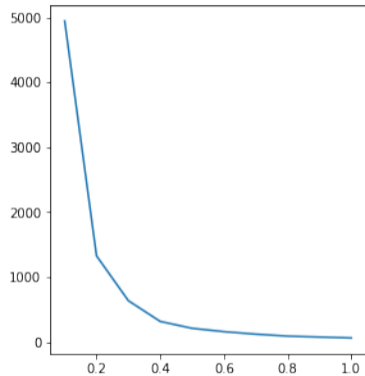


- ⊙ More than 30 indices can be found in the literature
- ⊙ Several studies and comparisons have been performed
- ⊙ Many studies have exhaustively tested these indices and many of its variations, some have a performance significantly better than others
- ⊙ Some indices show a similar performance (not statistically different)
- ⊙ Different studies conclude that Silhouette, Davies-Bouldin and Calinski Harabasz perform well in a wide range of situations, but to use them in combination is better

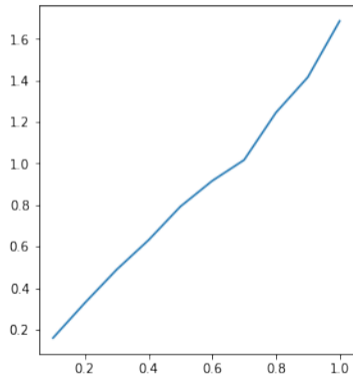
Internal criteria - 5 clusters different variance



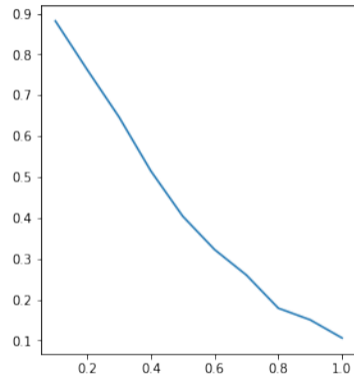
Calinski-Harabaz



Davies-Bouldin



Silhouette



External criteria

- ⊙ These indices measure the similarity of a clustering to a model partition P
- ⊙ Without a model they can be used to compare the results of using different parameters or different algorithms
 - For instance, can be used to assess the sensitivity to initialization
- ⊙ The main advantage is that these indices are independent of the examples/cluster description
- ⊙ That means that they can be used to assess any clustering algorithm

- ⊙ All the indices are based on the coincidence of each pair of examples in the groups of two clusterings
- ⊙ The computations are based on four values:

	Clustering 1	
Clustering 2	Same	Different
Same	a	b
Different	c	d

Rand statistic

$$Rand = \frac{(a + d)}{(a + b + c + d)}$$

Percentage of correct decisions. Measures agreement

Jaccard Coefficient

$$J = \frac{a}{(a + b + c)}$$

Ignores true negatives (d)

Folkes and Mallow

$$FM = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}}$$

Geometric mean of precision and recall

- ⊙ The rand statistic is sensitive to chance and is not 0 for random clusterings
- ⊙ It can be adjusted for chance computing the expected value

$$ARI = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+b)(a+c)}{a+b+c+d}}$$

- ⊙ A value of 1 corresponds to perfect match, a value of 0 is random labelling, it can be negative

- Defining Mutual Information between two partitions as:

$$MI(Y_i, Y_k) = \sum_{X_c^i \in Y_i} \sum_{X_{c'}^k \in Y_k} \frac{|X_c^i \cap X_{c'}^k|}{N} \log_2 \left(\frac{N |X_c^i \cap X_{c'}^k|}{|X_c^i| |X_{c'}^k|} \right)$$

- and Entropy of a partition as

$$H(Y_i) = - \sum_{X_c^i \in Y_i} \frac{|X_c^i|}{N} \log_2 \left(\frac{|X_c^i|}{N} \right)$$

where $|X_c^i \cap X_{c'}^k|$ is the number of samples that are in the intersection of the two groups

⊙ Normalized Mutual Information:

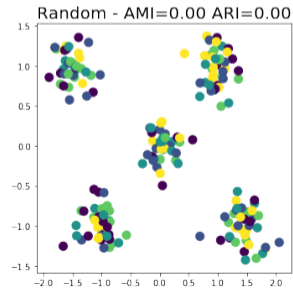
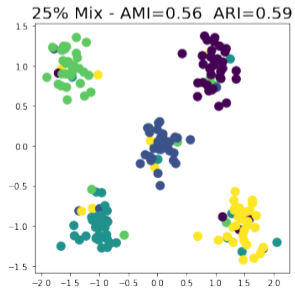
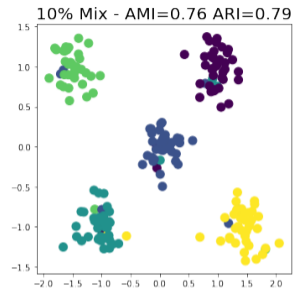
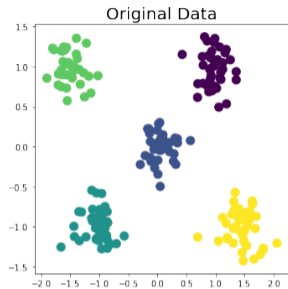
$$NMI(Y_i, Y_k) = \frac{MI(Y_i, Y_k)}{\sqrt{H(Y_i)H(Y_k)}}$$

⊙ Variation of Information:

$$VI(Y_i, Y_k) = H(Y_i) + H(Y_k) - 2MI(Y_i, Y_k)$$

⊙ Adjusted Mutual Information:

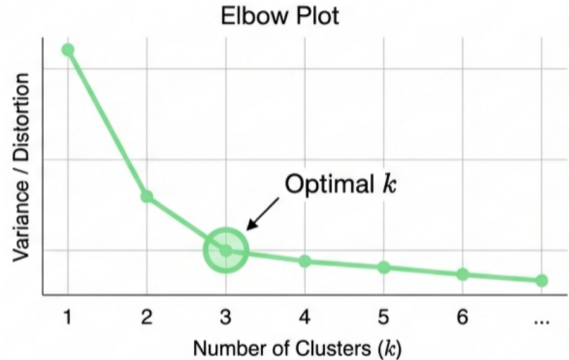
$$AMI(Y_i, Y_k) = \frac{MI(Y_i, Y_k) - E(MI(Y_i, Y_k))}{\max(H(Y_i), H(Y_k)) - E(MI(Y_i, Y_k))}$$



Number of clusters



- ⦿ Related validation is to decide the number of clusters of a partition
- ⦿ We compare clusterings of different sizes
- ⦿ Usually internal criteria indices are used
- ⦿ A plot of these indices can show what number of clusters is more probable (Elbow Plot)



- ⊙ All the internal validity indices can be used for this purpose
- ⊙ Other classical criteria based on the change the within cluster variance can also be used as:
 - Hartigan index:

$$H(k) = \left[\frac{S_W(k)}{S_W(k+1)} - 1 \right] (n - k - 1)$$

- Krzanowski Lai index:

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right|$$

being $DIFF(k) = (k-1)^{2/p} S_W(k-1) - k^{2/p} S_W(k)$

- ⊙ Assess the number of clusters comparing a clustering with the expected distribution of data given the null hypothesis (no clusters)
- ⊙ Computes different clusterings of the data increasing the number of clusters and compare them to clusters of data (B) generated with a **uniform distribution**
- ⊙ The interclass distance matrix S_W is computed for both and compared.
- ⊙ The correct number of clusters is where the **widest gap** appears between the S_W of the data and the uniform data

⊙ The Gap statistic:

$$Gap(k) = (1/B) \sum_b \log(S_W(k)_b) - \log(S_W(k))$$

The first term is the mean of S_W for the clusters obtained from the uniform distributed data

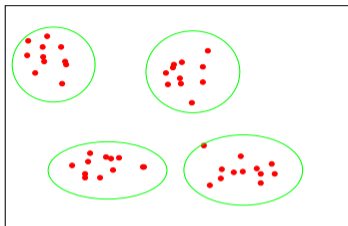
⊙ From the standard deviation of $\sum_b \log(S_W(k)_b)$ is defined the parameter s_k as:

$$s_k = std(\sum_b \log(S_W(k)_b)) \sqrt{1 + 1/B}$$

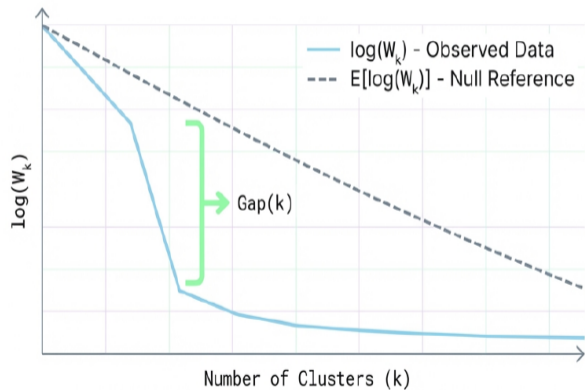
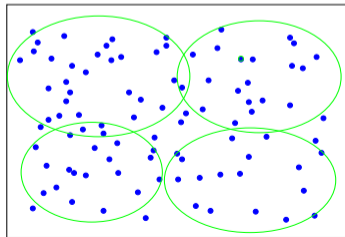
⊙ The probable number of clusters is the smallest number that holds:

$$Gap(k) \geq Gap(k + 1) - s_{k+1}$$

Clustered Data



Uniform Distributed Data





This Python Notebook has examples for Measures of Clustering Validation

- ⦿ Clustering Validation Notebook ([click here](#) to open the notebook in colab)

If you download the notebook you will be able to use it locally (run jupyter notebook to open the notebooks)