Clustering Evaluation

Javier Béjar

URL - 2024 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

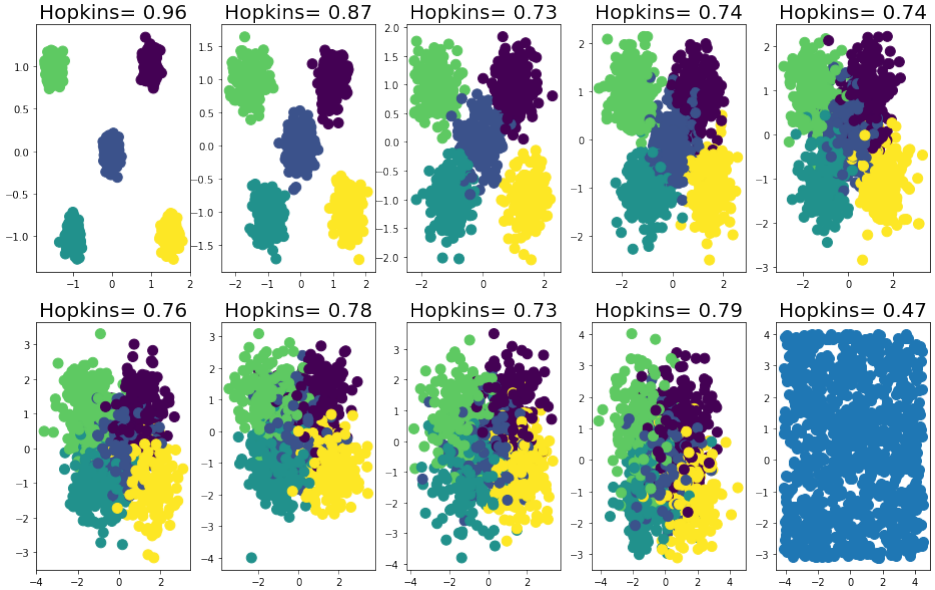
- Before clustering a dataset we can test if there are actually clusters
- We have to test the hypothesis of the existence of patterns in the data versus a dataset uniformly distributed (homogeneous distribution)

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


Hopkins= 0.96
Hopkins= 0.87
Hopkins= 0.73
Hopkins= 0.74
Hopkins= 0.74
Hopkins= 0.76
Hopkins= 0.78
Hopkins= 0.73
Hopkins= 0.79
Hopkins= 0.47

- 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
  - Relative criteria: Comparison with other clusterings

Internal criteria

- Measure properties expected in a good clustering
  - Compact groups
  - Well separated groups
- The indices are based on the model of the 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)

- For prototype based algorithms several measures can be used to compute quality indices
- Scatter matrices: Interclass distance, intraclass distance, separation

$$
\begin{aligned}
S_{W_k} &= \sum_{\forall x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T \\
S_{B_k} &= |C_k|(\mu_k - \mu)(\mu_k - \mu)^T \\
S_{M_{k,l}} &= \sum_{\forall i \in C_k} \sum_{\forall j \in C_l} (x_i - x_j)(x_i - x_j)^T
\end{aligned}
$$

- Trace criteria (lower overall intracluster distance/higher overall intercluster distance)

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

- Calinski-Harabasz index (interclass-intraclass distance ratio)

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

- Davies-Bouldin criteria (maximum interclass-intraclass distance ratio)

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

where

$$\begin{aligned} R_{ij} &= \frac{S_{W_i} + S_{W_j}}{S_{M_{ij}}} \\ R_i &= \max_{j:j \neq i} R_{ij} \end{aligned}$$

- Silhouette index (maximum class spread/variance)

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

Where

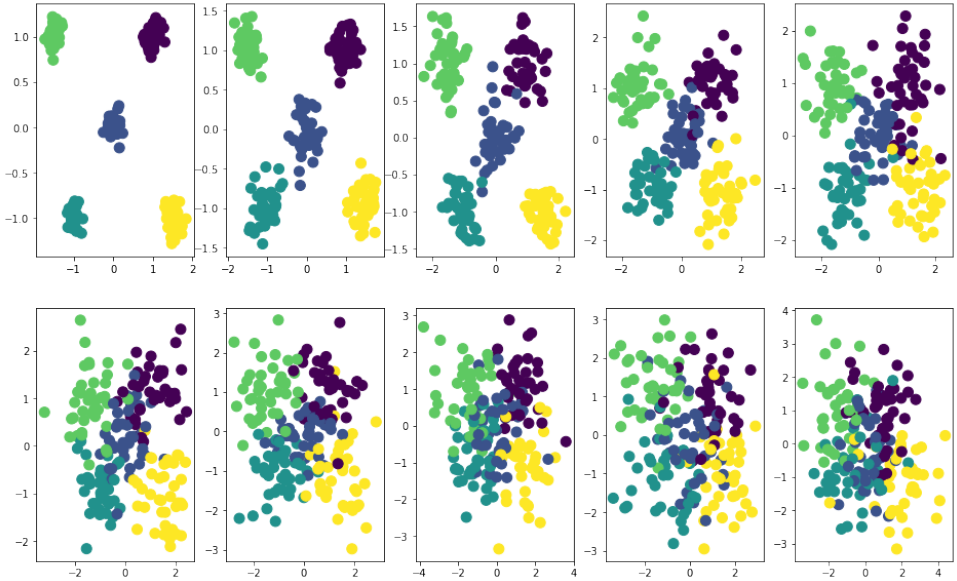
$$a_i = \frac{1}{|C_i| - 1} \sum_{y \in C_i, x \in C_i, y \neq x} \|y - x\|$$

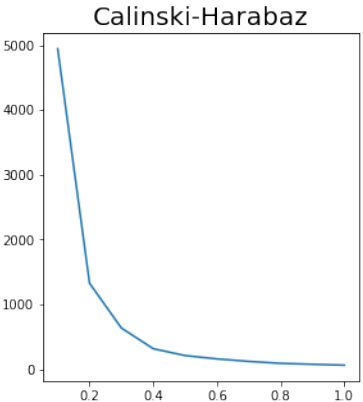
$$b_i = \min_{l \in H, l \neq i} \frac{1}{|C_l|} \sum_{y \in C_l, x \in C_i} \|y - x\|$$

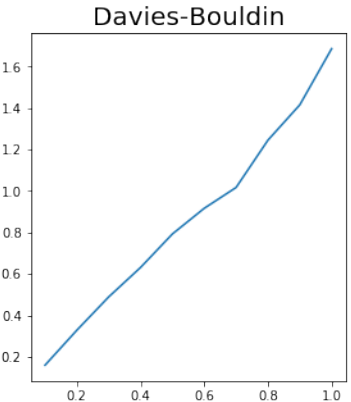
with $H = \{h : 1 \leq h \leq K\}$

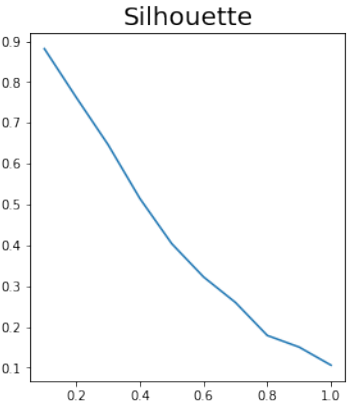
- More than 30 indices can be found in the literature
- Several studies and comparisons have been performed
- Recent studies (Arbelatiz et al, 2013) have exhaustively tested these indices, some have a performance significantly better that others
- Some indices show a similar performance (not statistically different)
- The study concludes that Silhouette, Davies-Bouldin and Calinski Harabasz perform well in a wide range of situations

Internal criteria - 5 clusters different variance




Calinski-Harabaz
5000
4000
3000
2000
1000
0
0.2
0.4
0.6
0.8
1.0


Davies-Bouldin
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0


Silhouette
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.2
0.4
0.6
0.8
1.0

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/Adjusted Rand statistic:

$$Rand = \frac{(a+d)}{(a+b+c+d)}; \quad ARand = \frac{a - \frac{(a+c)(a+b)}{a+b+c+d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+b)(a+c)}{a+b+c+d}}$$

- Jaccard Coefficient:

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

- Folkes and Mallow index:

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

- 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 objects that are in the intersection of the two groups

- Normalized Mutual Information:

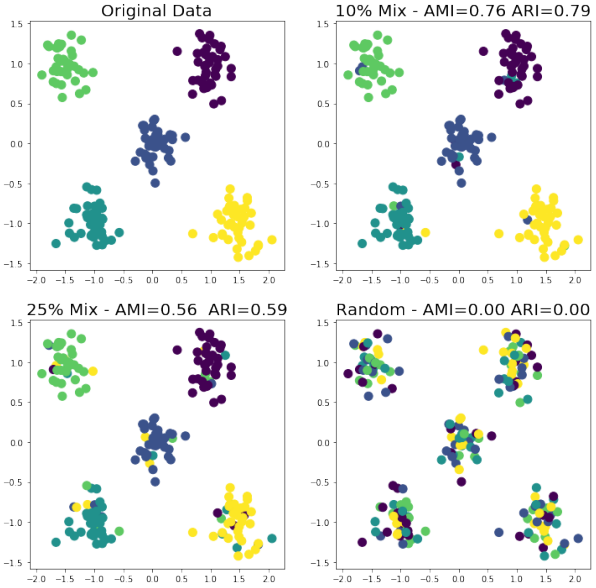
$$NMI(Y_i, Y_k) = \frac{MI(Y_i, Y_k)}{\sqrt{H(Y_i)H(Y_k)}}$$

- Variation of Information:

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

- Adjusted Mutual Information:

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


Original Data
10% Mix - AMI=0.76 ARI=0.79
25% Mix - AMI=0.56 ARI=0.59
Random - AMI=0.00 ARI=0.00

Number of clusters

- A topic related to cluster validation is to decide if the number of clusters obtained is the correct one
- This point is important specially for the algorithms that need this value as a parameter
- The usual procedure is to compare the characteristics of clusterings of different sizes
- Usually internal criteria indices are used in this comparison
- A graphic of these indices for different number of clusters can show what number of clusters is more probable

- Some internal validity indices can be used for this purpose: Calinsky-Harabasz index, Silhouette index
- Using the within class scatter matrix (S_W) other criteria can be defined:
  - 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 st. dev. (sd_k) of $\sum_b log(S_W(k)_b)$ is defined s_k as:

$$s_k = sd_k \sqrt{1 + 1/B}$$

- 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
log(S_W)
2 3 4 5 6 7 8
N Clusters

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)