香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

Cluster Analysis: Different Evaluation Metrics

CS5483 Data Warehousing and Data Mining

Different aspects for evaluation

- **Tendency**: Do the objects actually form clusters? Statistical significance?
- **Quantity**: What is the number of clusters?
- **Quality**: Are the clusters
  - Correct if ground truth is provided? **E__________ measure**.
  - "Clear" if ground truth is not provided? **I_________ measure**.

Centroid-based methods

- The **within cluster s____ of s_______ error** (WSS) of the clustering solution as a function of k

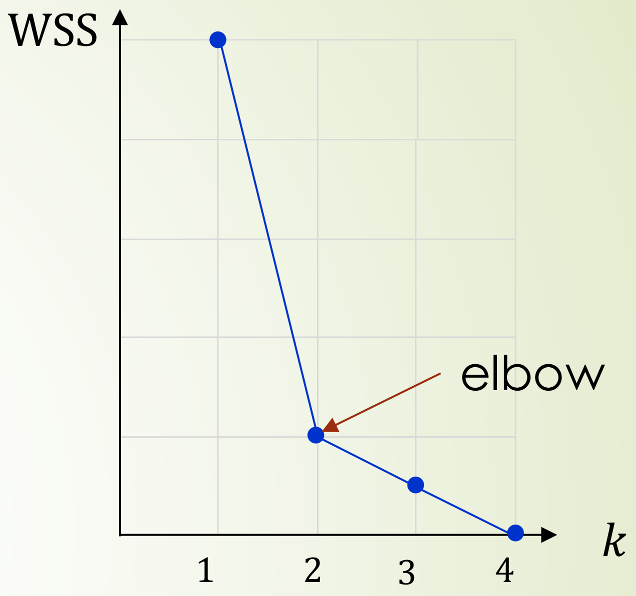
$$\mathrm{WSS}(k) := \sum_{j=1}^{k} \sum_{\boldsymbol{p} \in C_j} \mathrm{dist}(\boldsymbol{p}, \boldsymbol{c}_j)^2,$$

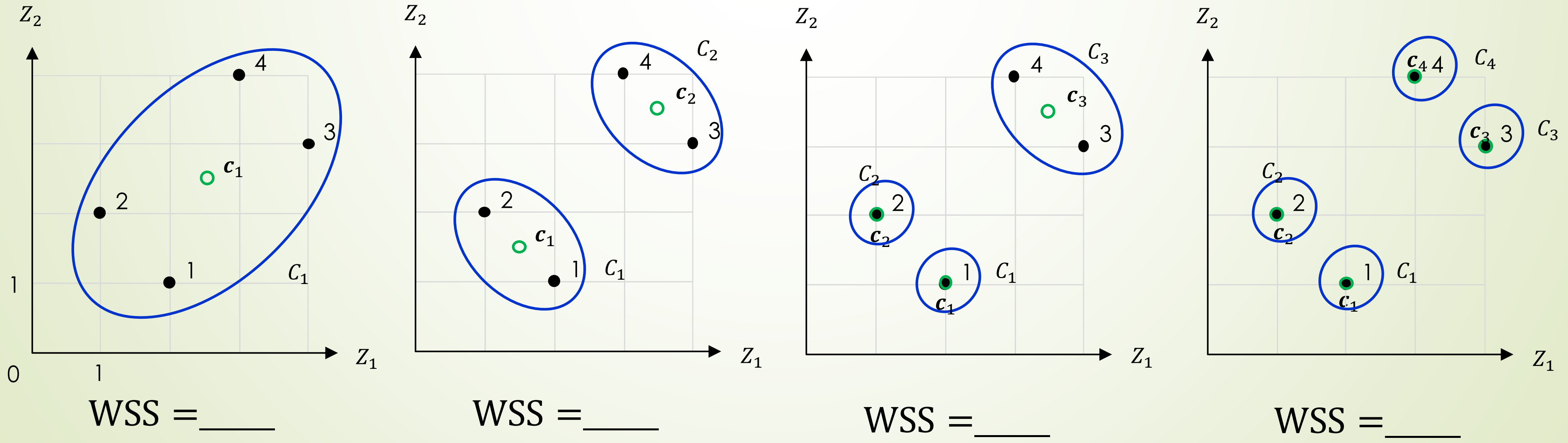
  also denoted as $\mathrm{var}(k)$ in [Han11].

- WSS is an <u>intrinsic/extrinsic</u> quality measure.
- WSS is <u>small/large</u> if the clusters are clear, i.e., sample points deviate less from centroids.
- Should we choose k by minimizing WSS? <u>Yes/No</u> because
  ____.

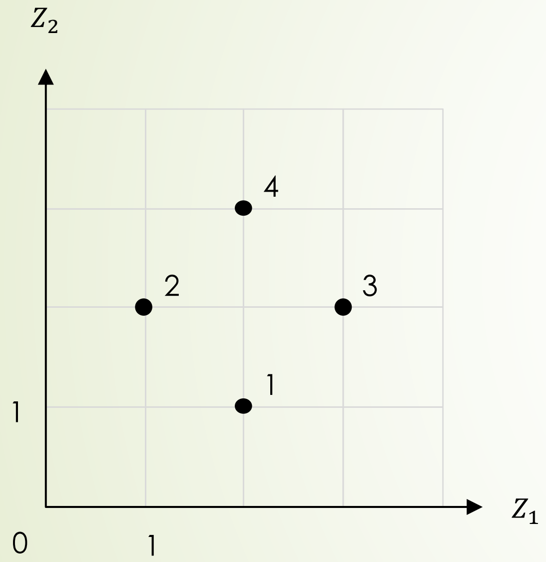
Elbow method

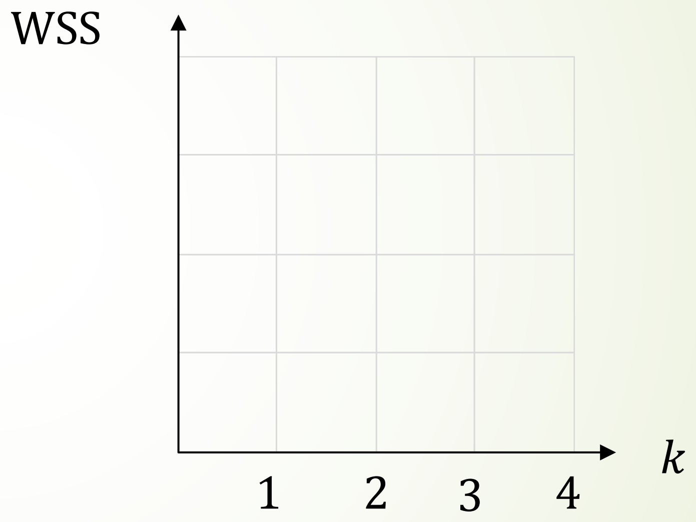
- Choose k that gives large **m**_________ decrease in WSS from $k-1$.
- Any problem of overfitting?





Limitation of the elbow method





- The elbow method returns $k =$______________.
- The elbow method can fail. It is only a **h**______________________________.
- Other intrinsic measures and analysis?

Silhouette coefficient

- For any point $\boldsymbol{p} \in D$ in the i -th cluster C_i :

$$s(\boldsymbol{p}) \coloneqq \begin{cases} \text{undefined}, & k = 1 \\ 0, & |C_i| = 1 \\ \dfrac{b(\boldsymbol{p}) - a(\boldsymbol{p})}{\max\{a(\boldsymbol{p}), b(\boldsymbol{p})\}}, & |C_i| > 1 \end{cases}$$

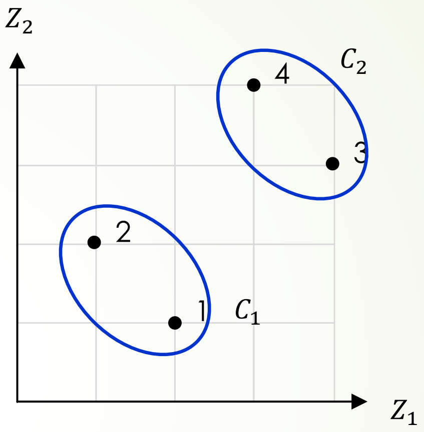
  - **Mean i________-cluster distance**:

$$a(\boldsymbol{p}) \coloneqq \frac{1}{|C_i| - 1} \sum_{\boldsymbol{q} \in C_i : \boldsymbol{q} \neq \boldsymbol{p}} \operatorname{dist}(\boldsymbol{p}, \boldsymbol{q})$$

  - **Mean n________-cluster distance**:

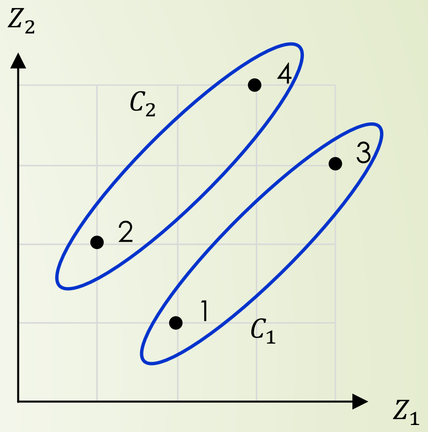
$$b(\boldsymbol{p}) \coloneqq \min_{j : \boldsymbol{p} \notin C_j} \frac{1}{|C_j|} \sum_{\boldsymbol{q} \in C_j} \operatorname{dist}(\boldsymbol{p}, \boldsymbol{q})$$

- Quality of C_i: average $s(\boldsymbol{p})$ over $\boldsymbol{p} \in C_i$.
- Overall quality: average $s(\boldsymbol{p})$ over $\boldsymbol{p} \in D$.



$$a(\boldsymbol{p}) = \sqrt{2}$$
$$b(\boldsymbol{p}) = \sqrt{2} + \sqrt{5/2}$$
$$s(\boldsymbol{p}) = \frac{\sqrt{5}}{2 + \sqrt{5}} \approx 0.53$$

Quality of C_1: 0.53
Quality of C_2: 0.53
Overall quality: 0.53



$$a(\boldsymbol{p}) = 2\sqrt{2}$$
$$b(\boldsymbol{p}) = 1/\sqrt{2} + \sqrt{5/2}$$
$$s(\boldsymbol{p}) = \frac{\sqrt{5} - 3}{4} \approx -0.19$$

Quality of C_1: -0.19
Quality of C_2:-0.19
Overall quality:-0.19

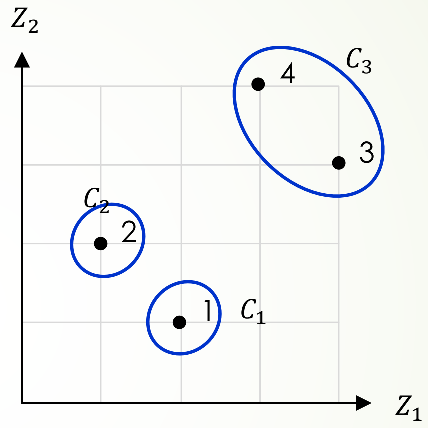
Silhouette coefficient

- For any point $\boldsymbol{p} \in D$ in the i -th cluster C_i :

$$s(\boldsymbol{p}) := \begin{cases} \text{undefined}, & k = 1 \\ 0, & |C_i| = 1 \\ \dfrac{b(\boldsymbol{p}) - a(\boldsymbol{p})}{\max\{a(\boldsymbol{p}), b(\boldsymbol{p})\}}, & |C_i| > 1 \end{cases}$$

Why?

- $s(\boldsymbol{p}) \in [-1,1]$: why?
  - $b(\boldsymbol{p}) > a(\boldsymbol{p})$: $s(\boldsymbol{p}) =$_______________
  - $b(\boldsymbol{p}) < a(\boldsymbol{p})$: $s(\boldsymbol{p}) =$_______________
- Sometimes we need a more detailed analysis by a silhouette plot.
- The method can fail on non-s____________ clusters.



$s(\boldsymbol{p}_1) = s(\boldsymbol{p}_2) =$____

$a(\boldsymbol{p}_3) = a(\boldsymbol{p}_4) =$____

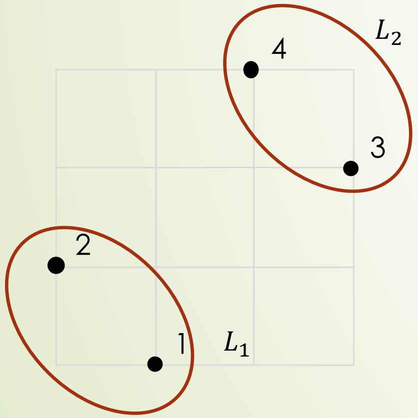
$b(\boldsymbol{p}_3) = b(\boldsymbol{p}_4) =$____

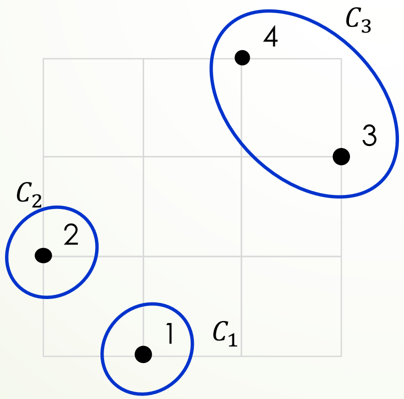
$s(\boldsymbol{p}_3) = s(\boldsymbol{p}_4) =$____

Quality of C_1: __________
Quality of C_2: __________
Quality of C_3: __________
Quality of the clustering: __________

Extrinsic cluster quality measures

- Setting: Ground truth is available
  - $L(\boldsymbol{p})$: **C**______________ (class) of $\boldsymbol{p}$. (Ground truth.)
  - $C(\boldsymbol{p})$: Cluster index of $\boldsymbol{p}$.
- How to compare the clustering solution to the ground truth?



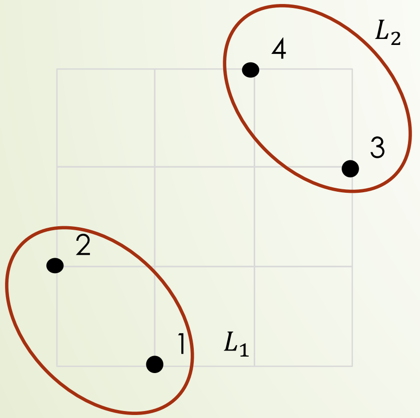


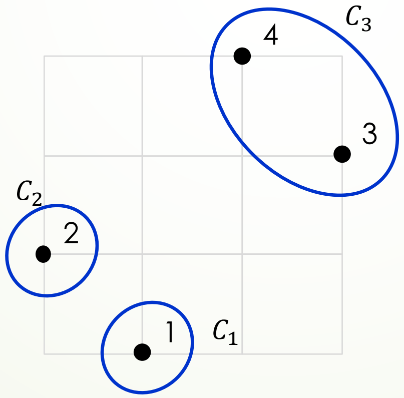
$$L(\boldsymbol{p}_1) = L(\boldsymbol{p}_2) = 1$$
$$L(\boldsymbol{p}_3) = L(\boldsymbol{p}_4) = 2$$

$$C(\boldsymbol{p}_1) = 1$$
$$C(\boldsymbol{p}_2) = 2$$
$$C(\boldsymbol{p}_3) = C(\boldsymbol{p}_4) = 3$$

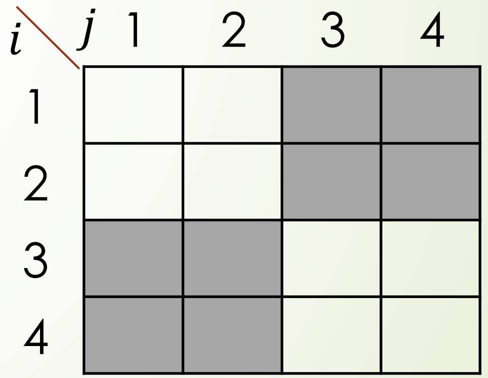
Extrinsic cluster quality measures

- For two points $\boldsymbol{p}, \boldsymbol{q} \in D$ to be clustered correctly:
  - $\boldsymbol{p}, \boldsymbol{q}$ should belong to the same cluster if they are in the same category.
  - $\boldsymbol{p}, \boldsymbol{q}$ should NOT belong to the same cluster if they are NOT in the same category.
  - Success indicator: $\text{correctness}(\boldsymbol{p}, \boldsymbol{q}) = \mathbb{1}\big(L(\boldsymbol{p}) = L(\boldsymbol{q}) \leftrightarrow C(\boldsymbol{p}) = C(\boldsymbol{q})\big)$





$\text{correctness}(\boldsymbol{p}_i, \boldsymbol{p}_j)$



Accuracy:__________

- Is accuracy a good performance metric? Yes/No because ____________________________

B-Cubed precision and recall

- For $\boldsymbol{p}, \boldsymbol{q} \in D$ (allowing $\boldsymbol{p} = \boldsymbol{q}$), compute the confusion matrix

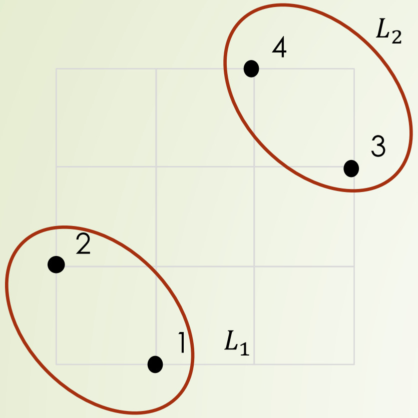
	$C(\boldsymbol{p}) = C(\boldsymbol{q})$	$C(\boldsymbol{p}) \neq C(\boldsymbol{q})$
$L(\boldsymbol{p}) = L(\boldsymbol{q})$	TP	FN
$L(\boldsymbol{p}) \neq L(\boldsymbol{q})$	FP	TN

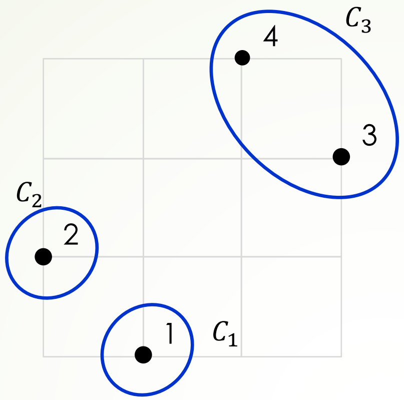
- Sample $\boldsymbol{p}$ and $\boldsymbol{q}$ from D with replacement.
  - If $\boldsymbol{p}, \boldsymbol{q}$ are in the same cluster, what is the chance they are in the same category?

    $$*\text{recall/precision} \coloneqq —$$

  - If $\boldsymbol{p}, \boldsymbol{q}$ are in the same category, what is the chance they are in the same cluster?

    $$*\text{recall/precision} \coloneqq —$$





precision =_______

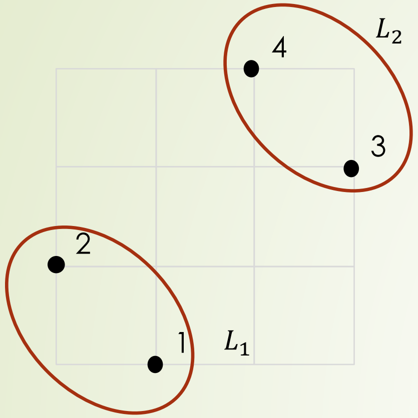
recall =________

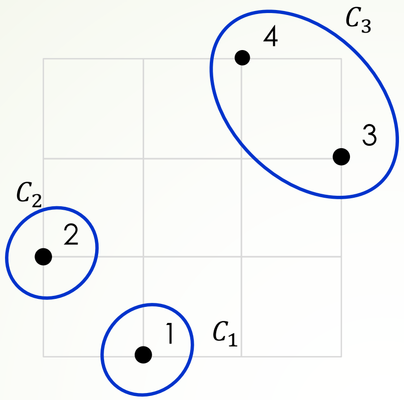
	$C(\boldsymbol{p}) = C(\boldsymbol{q})$	$C(\boldsymbol{p}) \neq C(\boldsymbol{q})$
$L(\boldsymbol{p}) = L(\boldsymbol{q})$	TP=6: (1,1),(2,2),(3,3),(4,4),(3,4),(4,3)	FN=___
$L(\boldsymbol{p}) \neq L(\boldsymbol{q})$	FP=___	TN=8: (1,3),(3,1),(2,3),(3,2),(1,4),(4,1),(2,4),(4,1)

Sanity check: total number of counts should be ____________________.

Alternative formulae

- B−Cubed precision is the average over $\boldsymbol{p} \in D$ of
$$\text{precision}(\boldsymbol{p}) := \frac{|\{\boldsymbol{q} \in D | C(\boldsymbol{p}) = C(\boldsymbol{q}), L(\boldsymbol{p}) = L(\boldsymbol{q})\}|}{|\{\boldsymbol{q} \in D | C(\boldsymbol{p}) = C(\boldsymbol{q})\}|}$$
- B−Cubed recall is the average over $\boldsymbol{p} \in D$ of
$$\text{recall}(\boldsymbol{p}) := \frac{|\{\boldsymbol{q} \in D | C(\boldsymbol{p}) = C(\boldsymbol{q}), L(\boldsymbol{p}) = L(\boldsymbol{q})\}|}{|\{\boldsymbol{q} \in D | L(\boldsymbol{p}) = L(\boldsymbol{q})\}|}$$
- Advantage: Breaks down the quality measure and its computation to individual point.





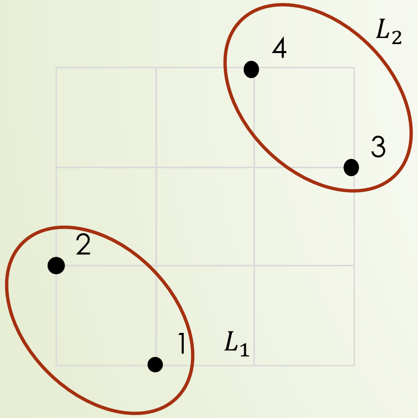
precision =________

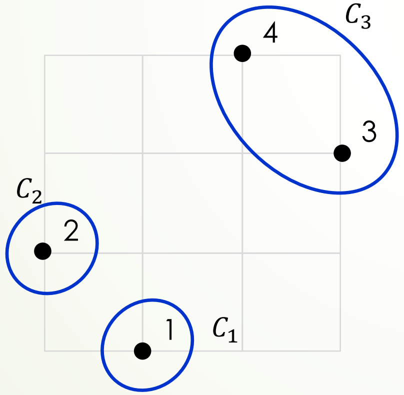
recall =________

i	1	2	3	4
precision($\boldsymbol{p}_i$)			100%	100%
recall($\boldsymbol{p}_i$)			100%	100%

Classes to clusters evaluation (WEKA)

- Match class to cluster labels to maximize correctly classified tuples.
- Use the “classification” error rate as the performance measure.





Number of instances

L \ C	1	2	3
1	1	1	0
2	0	0	2

Class assignment

Cluster ← Class
1←_____
2←_____
3←_____

Accuracy = _____%

- (Optional) There are other extrinsic measures such as the adjusted rand index.

References

- 10.4.2 Ordering Points to Identify the Clustering Structure
- 10.6 Evaluation of Clustering
- Error in Han11:
  - Should remove the constraint $i \neq j$ from (10.29) and (10.30).
- Supplementary readings:
  - Amigó, Enrique, et al. "A comparison of extrinsic clustering evaluation metrics based on formal constraints." Information retrieval 12.4 (2009): 461-486.
  - https://scikit-learn.org/stable/modules/clustering.html#clustering-evaluation