Clustering validation

Is there any real clustering? How good is it?

- Book: Chapter 6.9
- External material: Halkidi et al. (2002): Cluster Validity Methods: Part I. ACM SIGMOD Record 31(2): 40–45. `https://doi.org/10.1145/565117.565124`

Three similar problems

1. Clustering tendency: is there any clustering in data presented with certain features?
2. Determining number of clusters (or other parameters)
3. Evaluating goodness of clustering
   - compare different methods
   - compare against classification

All three depend on the **clustering objective**!

- assumptions on clusters (e.g., compactness, shape)
- separation between clusters

Evaluating goodness of clustering

1. **Internal criteria**
   - validity indices, similar to objective functions
   - do not work, if clustering had a different objective!
   - can be used to i) evaluate a single clustering or ii) compare clusterings (as **relative indices**)
2. **External criteria**
   - compare clustering to a predefined classification
   - classes may not reflect natural clusters
3. **Statistical hypothesis testing**
   - maybe the most sound approach, but computationally demanding

Internal validity indices

- indices assume some clustering objective $\rightarrow$ reward methods with the same objective
  - even a good clustering can get a bad score if a different objective!
  - many indices assume/favor spherical or convex clusters
- best for comparing similar algorithms and tuning parameters
- Some popular indices:
  - **Average silhouette**
  - **Calinski-Harabasz index**
  - **Davies-Bouldin index**

Silhouette index

Silhouette of a point $\mathbf{x}$ is

$$S(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \text{ a cluster of its own} \\ \frac{b-a}{\max\{a,b\}} & \text{otherwise} \end{cases}$$

$$a = avg\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C, \mathbf{y} \in C\}$$
$$b = \min_q avg\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C, \mathbf{y} \in C_q, C \neq C_q\}$$

$\approx$ how closely $\mathbf{x}$ matches its own cluster and how loosely the neighbouring cluster

- $S(\mathbf{x}) \in [-1, 1]$, **high values good**
- **Average silhouette** describes goodness of entire clustering
- flexible: any distance function d

Example: Silhouette of points

What negative values mean?

$$S(\mathbf{x}) = \begin{cases} 0 & \text{if singleton} \\ \frac{b-a}{\max\{a,b\}} & \text{otherwise} \end{cases}$$

$a = avg\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C, \mathbf{y} \in C\}$
$b = \min_q avg\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in C, \mathbf{y} \in C_q, C \neq C_q\}$

image source `http://www.sthda.com/english/wiki/wiki.php?id_contents=7952`

Calinski-Harabasz index

$$S_{CH} = \frac{(n-K)B}{(K-1)W}$$

- **between-cluster variance** $B = \sum_{i=1}^{K} |C_i| L_2^2(\mathbf{c}_i, \mathbf{m})$, where $\mathbf{m}$ is the mean of the whole data
- **within-cluster variance** $W = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} L_2^2(\mathbf{x}, \mathbf{c}_i)$
- requires $K \geq 2$
- range $[0, \infty[$, **high values good**
- When could you get value 0?

Calinski-Harabasz index (cont'd)

$$S_{CH} = \frac{(n-K)B}{(K-1)W} = \frac{(n-K)\sum_{i=1}^{K}|C_i|L_2^2(\mathbf{c}_i, \mathbf{m})}{(K-1)\sum_{i=1}^{K}\sum_{\mathbf{x}\in C_i} L_2^2(\mathbf{x}, \mathbf{c}_i)}$$

Note: $W = SSE(\mathbf{C})$. K-means criterion minimizes $W \Rightarrow$ maximizes B, because

$$\sum_{\mathbf{x}\in\mathcal{D}} L_2^2(\mathbf{x}, \mathbf{m}) = \sum_{i=1}^{K}\sum_{\mathbf{x}\in C_i} L_2^2(\mathbf{x}, \mathbf{c}_i)^2 + \sum_{i=1}^{K}|C_i|L_2^2(\mathbf{c}_i, \mathbf{m})$$

$\Rightarrow S_{CH}$ favours especially K-means!

Important: need to use L_2 in clustering!

Davies-Bouldin index

$$S_{DB} = \frac{1}{K} \sum_{i=1}^{K} \max_{j \neq i} \frac{S_i + S_j}{D_{ij}} \quad \text{, where}$$

- $S_i = \left(\frac{1}{|C_i|} \sum_{\mathbf{x} \in C_i} L_p^q(\mathbf{x}, \mathbf{c}_i) \right)^{\frac{1}{q}}$ measures dispersion of C_i
  - usually $q = 2$ (stdev of distances)
  - if $q = 1$, average distances
- $D_{ij} = L_p(\mathbf{c}_i, \mathbf{c}_j)$ measures separation between C_i and C_j
- max: for each C_i, evaluate relation to most problematic C_j
- possible to take avg instead of max

Important: use the same L_p as the clustering algorithm!

Davies-Bouldin index (cont'd)

$$S_{DB} = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} \frac{S_i + S_j}{D_{ij}} \quad \text{, where}$$

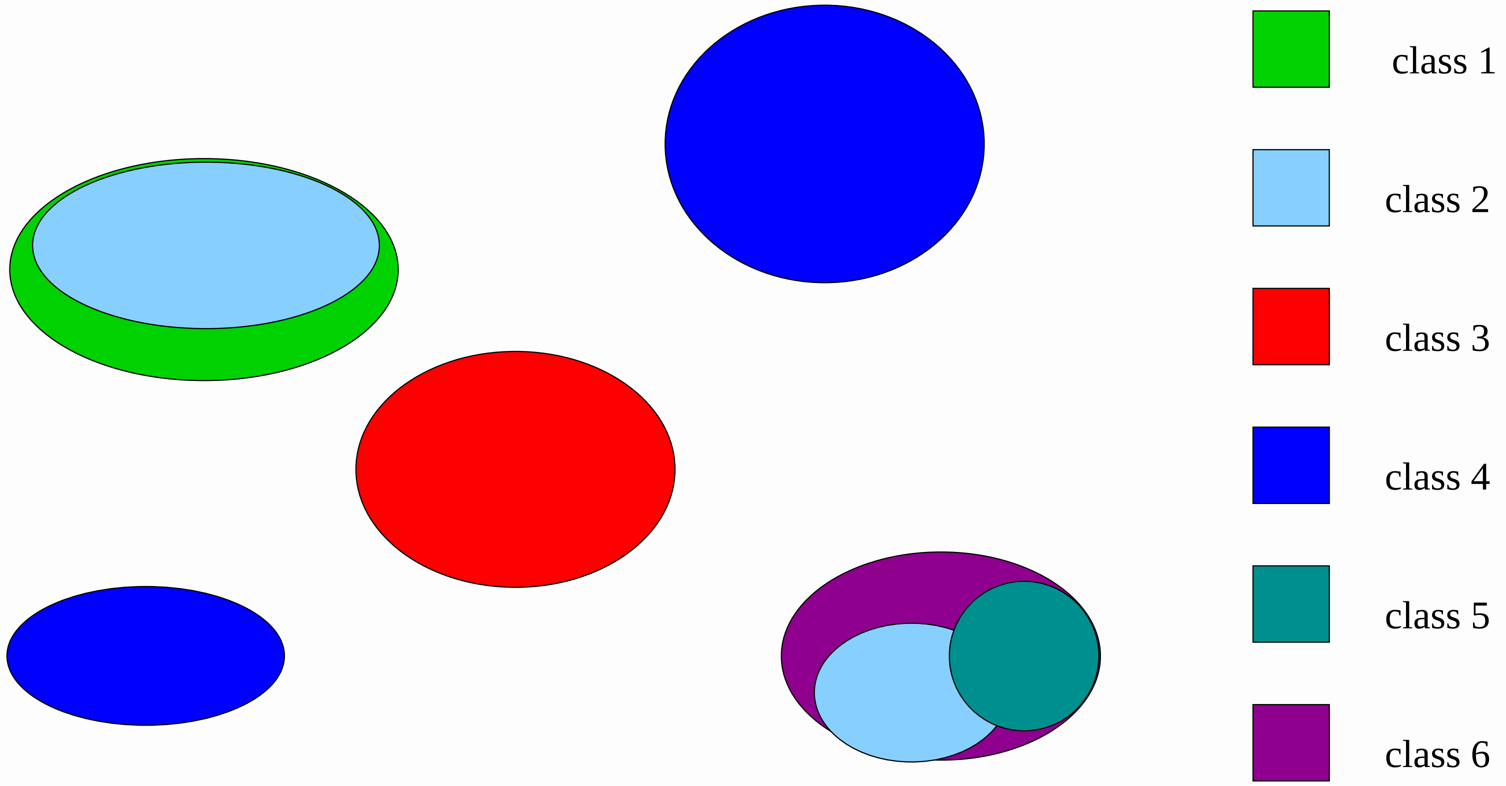
$S_i = \left(\frac{1}{|C_i|}\sum_{\mathbf{x}\in C_i} L_p^q(\mathbf{x}, \mathbf{c}_i)\right)^{\frac{1}{q}}$ and $D_{ij} = L_p(\mathbf{c}_i, \mathbf{c}_j)$

- range $[0, \infty[$, **small values good**
- When could you get value 0?

Possible strategies when S_{DB} used to determine K:

- restrict number of singletons (e.g., 0 or a few)
- define $S_i = a$ for some large a, when $|C_i| = 1$

External validation: Compare clustering against predefined classification



A confusion matrix: clustering vs. classification

	Class 1	Class 2	Class 3	
Cluster 1	n_{11}	n_{12}	n_{13}	m_1
Cluster 2	n_{21}	n_{22}	n_{23}	m_2
Cluster 3	n_{31}	n_{32}	n_{33}	m_3
	c_1	c_2	c_3	n

image source Cunnigham `https://slideplayer.com/slide/14318989/`

External validation

Given clustering $C_1, \ldots, C_K$ and classification $D_1, \ldots, D_q$. Many validation indices! E.g.,

- **purity**

$$Pur(C) = \frac{1}{n} \sum_{i=1}^{K} \max_j |C_i \cap D_j|$$

  - be careful! (increases with K)
- **normalized mutual information** NMI (robust, independent of K)
- **Rand index**

Normalized mutual information

Normalized mutual information by Strehl and Ghosh (2003):

$$NMI = \frac{I(C, D)}{\sqrt{H(C)H(D)}}$$

mutual information $I = \sum_{C_i \in C} \sum_{D_j \in D} P(C_i, D_j) \log \frac{P(C_i, D_j)}{P(C_i)P(D_j)}$

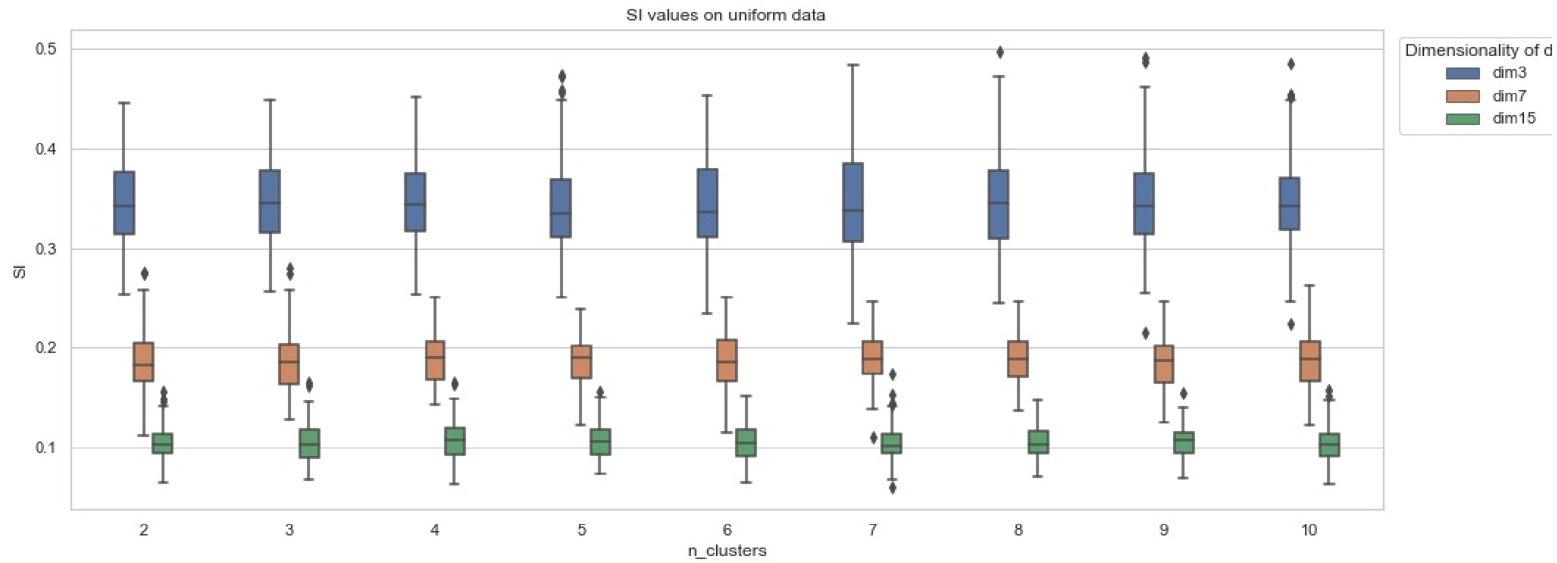
entropy $H(C) = -\sum_{C_i \in C} P(C_i) \log P(C_i)$

+ does not depend on the number of clusters
– many singleton clusters can cause problems

Note: Also other variants of normalized mutual information, give always equation and/or reference what you use!

Statistical hypothesis testing: motivation

SI can be pretty good even for random data!



- each feature generated independently from uniform distribution
- 100 randomizations
- K-means repeated 100 times → best result for each K

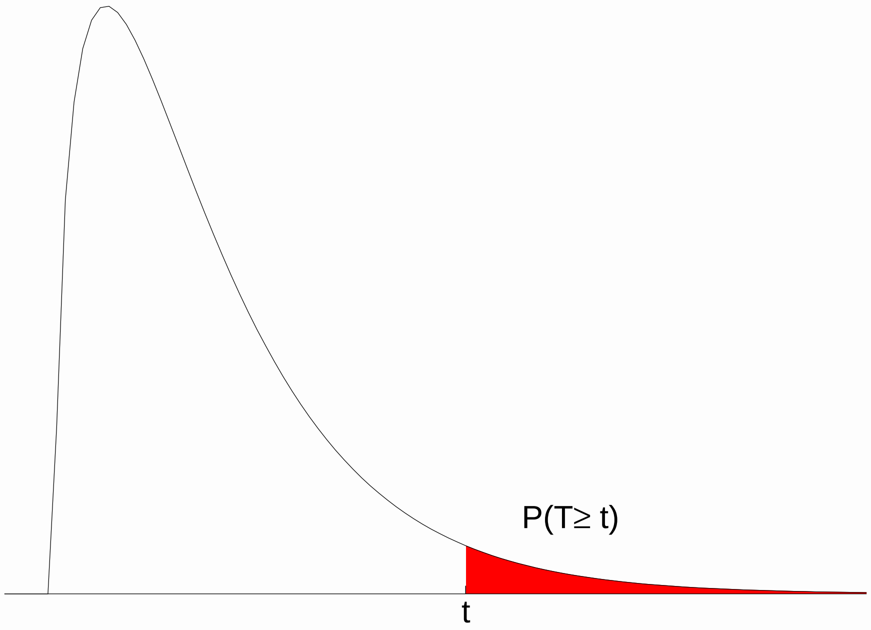
Experiment by Georgy Ananov for MDM 2023

Statistical hypothesis testing

Procedure:

1. decide a **null hypothesis** H_0 to test
   - describes the state where there isn't any clustering
   - e.g., H_0: All sets of n locations in certain region are equally likely.
2. decide a **test statistic** T
   - may be a validity index
3. What is the probability to obtain at least as good test statistic values as in data (where $T = t$) if H_0 was true?

Statistical hypothesis testing



Assume that large T value good

Idea: If $P(T \geq t)$ very small $\Rightarrow$ unlikely that the observed clustering had occurred by chance

- $P(T \geq t)$ is the **p-value** that can be used as a significance measure

Statistical hypothesis testing

Problem: How to evaluate p-value? (T's distribution seldom known!)

- often by Monte Carlo experiments (randomization tests):
  - generate random data sets fulfilling H_0, cluster them and evaluate T
  - p-value $\approx$ proportion of random sets that obtained $T \geq t$ (if large T good)
- – computationally demanding (a lot of simulations!)
- – many alternatives for H_0s and Ts

Other evaluation: What the clustering reveals?

- Look at cluster sizes (e.g., C_1: $n-2$ data points and C_2: 2 points – likely outliers!)
- How do the clusters differ? (selected and external features)
  - e.g., rats clustered by body measurements (weight, tail and body length, organ weights)
  - 2 clusters: big and small rats
  - vs. 3 clusters: C_1: young or sick rats, C_2: pregnant or nursing females, C_3: other adults
- Are all clusters clear? (e.g., C_1 and C_3 intermingled, C_2 separate)

Summary

- Remember validation, but be cautious!
    - even random data can produce clusterings, but they seldom pass validation
    - problem: indices biased or do not reflect the underlying clustering
    - try always more than one validation technique
- Objective, distance measure, clustering method and validation should match!

Sources and further reading

- Halkidi et al. (2001): On clustering validation techniques, Journal of Intelligent Information Systems 17: 107–145. `https://www.researchgate.net/publication/2500099_On_Clustering_Validation_Techniques`
- Jain and Dubes (1988): Algorithms for clustering data, Ch 4.
- Gan, Ma, Wu (2007): Data clustering - theory, algorithms, and applications, Ch 17, `https://www.researchgate.net/publication/220694937_Data_Clustering_Theory_Algorithms_and_Applications`

Sources and further reading

- Vargha, Bergman, Takacs: Performing Cluster Analysis Within a Person-Oriented Context: Some Methods for Evaluating the Quality of Cluster Solutions. Journal of Person-Oriented Research, 2: 78-86, 2016.