Artificial Intelligence I 2022/2023

Week 11 Tutorial and Additional Exercises

Evaluation of Clustering Algorithms

School of Computer Science

May 8, 2023

In this tutorial. . .

In this tutorial we will be covering

- Supervised and unsupervised clustering validation criteria.
- Silhouette coefficient.
- Classification-oriented validation criteria
- Similarity-oriented validation criteria.
- Cophenetic correlation coefficient.

Silhouette Coefficient

- Let $\mathbf{x}^{(i)}$ be an example in cluster C, and define
  1. a_i to be the average distance of $\mathbf{x}^{(i)}$ to all other examples in C, i.e.,

$$a_i := \frac{\sum_{\mathbf{x}\in C, \mathbf{x}\neq \mathbf{x}^{(i)}} d(\mathbf{x}^{(i)}, \mathbf{x})}{(\text{no. of examples in cluster } C) - 1}.$$

  2. b_i to be the minimum of the average distance of $\mathbf{x}^{(i)}$ to examples in other clusters, i.e.

$$b_i := \min_{\substack{k=1,\dots,K \\ C_k \neq C}} \frac{\sum_{\mathbf{x}\in C_k} d(\mathbf{x}^{(i)}, \mathbf{x})}{\text{no. of examples in } C_k}.$$

- The SC for $\mathbf{x}^{(i)}$ is defined as

$$s_i := \frac{b_i - a_i}{\max(a_i, b_i)}.$$

Silhouette Coefficient (continued)

- The SC of a cluster C is defined as

$$s_C := \frac{\sum_{\{i:\mathbf{x}^{(i)}\in C\}} s_i}{\text{no. of examples in cluster } C}.$$

- The SC of a clustering structure $\mathbf{C}$ with N examples is defined as

$$s_{\mathbf{C}} := \frac{\sum_{i=1}^{N} s_i}{N}.$$

Exercise 1

- Consider a dataset with 4 examples, clustered by an algorithm as
$$C_1 = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\}, \qquad C_2 = \{\mathbf{x}^{(3)}, \mathbf{x}^{(4)}\}.$$
- The distance matrix for these examples is the following

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$
$\mathbf{x}^{(1)}$	0	0.10	0.65	0.55
$\mathbf{x}^{(2)}$	0.10	0	0.70	0.60
$\mathbf{x}^{(3)}$	0.65	0.70	0	0.90
$\mathbf{x}^{(4)}$	0.55	0.60	0.90	0

- Compute the SC for each point, for each cluster, and for the overall clustering structure $\mathbf{C} = \{C_1, C_2\}$.
- Comment on the suitability of examples assigned to C_1.

Exercise 1: Solution

- We first find the a_i's and the b_i's

$$a_1 = 0.1, a_2 = 0.1, a_3 = 0.9, a_4 = 0.9.$$

$$b_1 = 0.6, b_2 = 0.65, b_3 = 0.675, b_4 = 0.575.$$

- We then find the SC of each example

$$s_1 = 0.8333, s_2 = 0.8461, s_3 = -0.25, s_4 = -0.3611.$$

- We then find the SC of each cluster

$$s_{C_1} = 0.8397, s_{C_2} = -0.3055.$$

- We then find the SC of the clustering structure

$$s_{\mathbf{C}} = 0.2670.$$

Classification-oriented validation criteria

- Consider a set of L different classes, clustered into K clusters.
- *Precision* of cluster i with respect to class j

$$precision(i,j) := \frac{\text{no. of examples of class } j \text{ in cluster } i}{\text{no. of examples in cluster } i}.$$

- *Recall* of cluster i with respect to class j

$$recall(i,j) := \frac{\text{no. of examples of class } j \text{ in cluster } i}{\text{no. of examples in class } j}.$$

- *F-measure* of cluster i with respect to class j

$$F(i,j) := \frac{2 \cdot precision(i,j) \cdot recall(i,j)}{precision(i,j) + recall(i,j)}.$$

Classification-oriented validation criteria (continued)

- The *entropy* of cluster i is defined as

$$e_i := -\sum_{j=1}^{L} precision(i,j) \cdot \log_2(precision(i,j)).$$

- The *total entropy* of the set of clusters is defined as

$$e := \sum_{i=1}^{K} \frac{\text{no. of examples in cluster } i}{\text{total no. of examples}} e_i .$$

- We want a low entropy.

Classification-oriented validity measures (continued)

- The *purity* of cluster i is defined as

$$p_i := \max_j \mathit{precision}(i, j).$$

- The *overall purity* of the set of clusters is defined as

$$p := \sum_{i=1}^{K} \frac{\text{no. of examples in cluster } i}{\text{total no. of examples}} p_i.$$

- We want a high purity.

Exercise 2

- Consider the set with 10 examples and 3 classes, clustered into 3 clusters (classes and clusters are not the same)

Example	Class	Cluster	Example	Class	Cluster
$\mathbf{x}^{(1)}$	1	1	$\mathbf{x}^{(6)}$	3	1
$\mathbf{x}^{(2)}$	3	2	$\mathbf{x}^{(7)}$	2	2
$\mathbf{x}^{(3)}$	2	3	$\mathbf{x}^{(8)}$	2	2
$\mathbf{x}^{(4)}$	1	1	$\mathbf{x}^{(9)}$	1	3
$\mathbf{x}^{(5)}$	3	2	$\mathbf{x}^{(10)}$	2	1

- Write down the confusion matrix.
- Compute the following
  1. $precision(1, 3)$.
  2. $recall(1, 3)$.
  3. $F(1, 3)$.
  4. e_2.
  5. p_2.

Exercise 2: Solution

- The confusion matrix is the following

	Cluster 1	Cluster 2	Cluster 3	Total
Class 1	2	0	1	3
Class 2	1	2	1	4
Class 3	1	2	0	3
Total	4	4	2	10

- We also compute the following
  1. $precision(1,3) = 1/4$.
  2. $recall(1,3) = 1/3$.
  3. $F(1,3) = 2/7$.
  4. $e_2 = 1$.
  5. $p_2 = 1/2$.

Similarity-oriented validation criteria

- Consider a set of N examples of different classes, clustered into clusters.
- The *ideal cluster similarity matrix* is an $N \times N$ matrix whose ij-th element equals 1 if examples i and j are in the same cluster, and 0 otherwise.
- The *ideal class similarity matrix* is an $N \times N$ matrix whose ij-th element equals 1 is examples i and j are in the same class, and 0 otherwise.
- We can compute the correlation between these two matrices.
- We can also use binary similarity-based measures.

Binary similarity-based measures

- Consider a set of N examples of different classes, clustered into clusters and define the following
  1. f_{00} := no. of pairs having different class and different cluster.
  2. f_{01} := no. of pairs having different class and same cluster.
  3. f_{10} := no. of pairs having same class and different cluster.
  4. f_{11} := no. of pairs having same class and same cluster.
- The *Rand statistic* is defined as

$$\textit{Rand statistic} = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}.$$

- The *Jaccard coefficient* is defined as

$$\textit{Jaccard coefficient} = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}.$$

Exercise 3

- Reconsider the first five examples of the previous set with 3 classes, clustered into 3 clusters

Example	Class	Cluster
$\mathbf{x}^{(1)}$	1	1
$\mathbf{x}^{(2)}$	3	2
$\mathbf{x}^{(3)}$	2	3
$\mathbf{x}^{(4)}$	1	1
$\mathbf{x}^{(5)}$	3	2

- Write down the ideal cluster similarity matrix and the ideal class similarity matrix.
- Compute the Rand statistic and the Jaccard coefficient.

Exercise 3: Solution

- The ideal cluster similarity matrix is the following

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$\mathbf{x}^{(1)}$	1	0	0	1	0
$\mathbf{x}^{(2)}$	0	1	0	0	1
$\mathbf{x}^{(3)}$	0	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1	0
$\mathbf{x}^{(5)}$	0	1	0	0	1

- The ideal class similarity matrix is the following

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$\mathbf{x}^{(1)}$	1	0	0	1	0
$\mathbf{x}^{(2)}$	0	1	0	0	1
$\mathbf{x}^{(3)}$	0	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1	0
$\mathbf{x}^{(5)}$	0	1	0	0	1

Exercise 3: Solution (continued)

- We first compute the following
  1. $f_{00} = 8$.
  2. $f_{01} = 0$.
  3. $f_{10} = 0$.
  4. $f_{11} = 2$.
- Therefore, *Rand statistic* $= 1$.
- Also, *Jaccard coefficient* $= 1$.

Cophenetic correlation coefficient

- Consider a set of N examples, clustered by an agglomerative clustering algorithm.
- The *cophenetic distance* of examples i and j is the distance at which an agglomerative algorithm puts these examples in the same cluster.
- The *cophenetic distance matrix* is an $N \times N$ matrix whose ij-th element equals the cophenetic distance between examples i and j.
- Note that the cophenetic distance matrix is different from the distance matrix we studied last week.

Cophenetic correlation coefficient (continued)

- Let $P_{i,j}$ denote the ij-th element of the cophenetic distance matrix P and $D_{i,j}$ denote the ij-th element of a distance matrix D, for some choice of distance function.
- Let d denote the average of the non-zero elements of D and p denote the average of the non-zero elements of P.
- The *cophenetic correlation coefficient* is defined as

$$CPCC := \frac{\sum_{\substack{i,j=1 \\ i<j}}^{N} (D_{i,j} - d)(P_{i,j} - p)}{\sqrt{\sum_{\substack{i,j=1 \\ i<j}}^{N} (D_{i,j} - d)^2 \sum_{\substack{i,j=1 \\ i<j}}^{N} (P_{i,j} - p)^2}}.$$

Exercise 4

- Consider a set with 5 examples and the following distance matrix (for some choice of distance function).

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$\mathbf{x}^{(1)}$	0	0.90	0.59	0.45	0.65
$\mathbf{x}^{(2)}$	0.90	0	0.36	0.53	1.02
$\mathbf{x}^{(3)}$	0.59	0.36	0	0.56	1.15
$\mathbf{x}^{(4)}$	0.45	0.53	0.56	0	1.24
$\mathbf{x}^{(5)}$	0.65	1.02	1.15	1.24	0

- Use single-linkage agglomerative clustering to cluster this set, and write down the resulting dendrogram.
- Using the dendrogram, write down the cophenetic distance matrix.
- Finally, compute the CPCC using the distance matrix and the cophenetic distance matrix.

Exercise 4: Solution

- We start with each example in its own cluster and calculate the distance matrix for these clusters.

	$\{\mathbf{x}^{(1)}\}$	$\{\mathbf{x}^{(2)}\}$	$\{\mathbf{x}^{(3)}\}$	$\{\mathbf{x}^{(4)}\}$	$\{\mathbf{x}^{(5)}\}$
$\{\mathbf{x}^{(1)}\}$	0	0.90	0.59	0.45	0.65
$\{\mathbf{x}^{(2)}\}$	0.90	0	0.36	0.53	1.02
$\{\mathbf{x}^{(3)}\}$	0.59	0.36	0	0.56	1.15
$\{\mathbf{x}^{(4)}\}$	0.45	0.53	0.56	0	1.24
$\{\mathbf{x}^{(5)}\}$	0.65	1.02	1.15	1.24	0

- The closest clusters are $\{\mathbf{x}^{(2)}\}$ and $\{\mathbf{x}^{(3)}\}$.
- The new clusters are

$$\{\mathbf{x}^{(1)}\}, \{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}, \{\mathbf{x}^{(4)}\}, \{\mathbf{x}^{(5)}\}.$$

Exercise 4: Solution (continued)

- We then recalculate the distance matrix for the new clusters.

	$\{\mathbf{x}^{(1)}\}$	$\{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$	$\{\mathbf{x}^{(4)}\}$	$\{\mathbf{x}^{(5)}\}$
$\{\mathbf{x}^{(1)}\}$	0	0.59	0.45	0.65
$\{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$	0.59	0	0.53	1.02
$\{\mathbf{x}^{(4)}\}$	0.45	0.53	0	1.24
$\{\mathbf{x}^{(5)}\}$	0.65	1.02	1.24	0

- The closest clusters are $\{\mathbf{x}^{(1)}\}$ and $\{\mathbf{x}^{(4)}\}$.
- The new clusters are

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(4)}\}, \{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}, \{\mathbf{x}^{(5)}\}.$$

Exercise 4: Solution (continued)

- We then recalculate the distance matrix for the new clusters.

	$\{\mathbf{x}^{(1)}, \mathbf{x}^{(4)}\}$	$\{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$	$\{\mathbf{x}^{(5)}\}$
$\{\mathbf{x}^{(1)}, \mathbf{x}^{(4)}\}$	0	0.53	0.65
$\{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$	0.53	0	1.02
$\{\mathbf{x}^{(5)}\}$	0.65	1.02	0

- The closest clusters are $\{\mathbf{x}^{(1)}, \mathbf{x}^{(4)}\}$ and $\{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\}$.
- The new clusters are

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}\}, \{\mathbf{x}^{(5)}\}.$$

Exercise 4: Solution (continued)

- We then recalculate the distance matrix for the new clusters.

	$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}\}$	$\{\mathbf{x}^{(5)}\}$
$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}\}$	0	0.65
$\{\mathbf{x}^{(5)}\}$	0.65	0

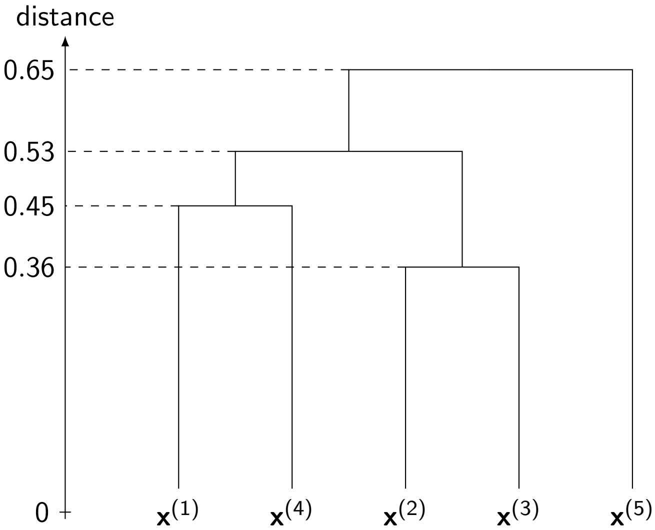
- The closest clusters are $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}\}$ and $\{\mathbf{x}^{(5)}\}$.
- The new clusters are

$$\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}\}.$$

- Finally, we construct the dendrogram.

Exercise 4: Solution (continued)

The dendrogram is the following:



Exercise 4: Solution (continued)

- The cophenetic distance matrix is the following

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$\mathbf{x}^{(1)}$	0	0.53	0.53	0.45	0.65
$\mathbf{x}^{(2)}$	0.53	0	0.36	0.53	0.65
$\mathbf{x}^{(3)}$	0.53	0.36	0	0.53	0.65
$\mathbf{x}^{(4)}$	0.45	0.53	0.53	0	0.65
$\mathbf{x}^{(5)}$	0.65	0.65	0.65	0.65	0

- We also compute $d = 0.745$ and $p = 0.553$.
- We finally compute $CPCC = 0.7978$.

Optional Material

Optional Exercise 1

- Recall the formal definition of a *distance metric*.

Definition 1 (Distance metric)

A function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *distance metric*, if and only if, for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$, the following hold:

1. $f(\mathbf{x}, \mathbf{y}) = 0$, if and only if, $\mathbf{x} = \mathbf{y}$;
2. $f(\mathbf{x}, \mathbf{y}) = f(\mathbf{y}, \mathbf{x})$; and
3. $f(\mathbf{x}, \mathbf{z}) \leq f(\mathbf{x}, \mathbf{y}) + f(\mathbf{y}, \mathbf{z})$.

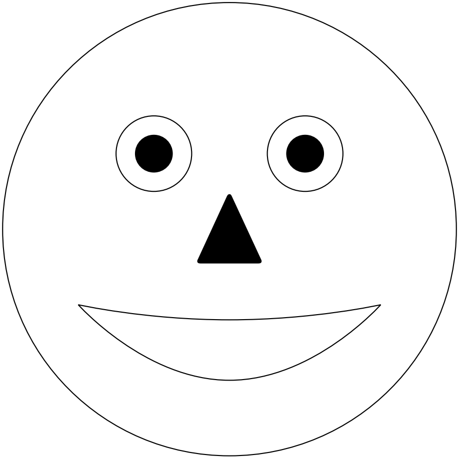
- Show that cophenetic distance is a distance metric.
- Hint: Argue in words instead of formulas.

Optional Exercise 1: Solution

- Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ be arbitrary vectors and denote the cophenetic distance by $P(\cdot, \cdot)$. We have
  1. If $\mathbf{x} = \mathbf{y}$, then they are in the same cluster from the start, and therefore $P(\mathbf{x}, \mathbf{y}) = 0$.
     If $P(\mathbf{x}, \mathbf{y}) = 0$, then the clusters of $\mathbf{x}$ and $\mathbf{y}$ are merged at distance 0, so $\mathbf{x}$ and $\mathbf{y}$ are in the same cluster from the start. But each vector starts in its own cluster, so this is only possible if $\mathbf{x} = \mathbf{y}$.
  2. Assume that $P(\mathbf{x}, \mathbf{y}) = a$. This means that the clusters of $\mathbf{x}$ and $\mathbf{y}$ are merged at distance a. Clearly then, the clusters of $\mathbf{y}$ and $\mathbf{x}$ are also merged at distance a, and therefore $P(\mathbf{y}, \mathbf{x}) = a = P(\mathbf{x}, \mathbf{y})$.
  3. By definition, we have $P(\mathbf{x}, \mathbf{z}) \leq \max\{P(\mathbf{x}, \mathbf{y}), P(\mathbf{y}, \mathbf{z})\}$. Since the cophenetic distance is non-negative by definition, we also have $\max\{P(\mathbf{x}, \mathbf{y}), P(\mathbf{y}, \mathbf{z})\} \leq P(\mathbf{x}, \mathbf{y}) + P(\mathbf{y}, \mathbf{z})$. Thus, $P(\mathbf{x}, \mathbf{z}) \leq P(\mathbf{x}, \mathbf{y}) + P(\mathbf{y}, \mathbf{z})$.
- Therefore, cophenetic distance is a distance metric.

Any questions?

Until the next time...



Thank you for your attention!