Artificial Intelligence I 2023/2024

Week 8 Tutorial and Additional Exercises

Hierarchical Clustering and Evaluation of Clustering Algorithms

School of Computer Science

4th of March 2024

In this tutorial...

In this tutorial we will be covering

- Hierarchical Clustering.
- Cutting the Dendrogram.
- Supervised and unsupervised clustering validation criteria.
- Silhouette coefficient.
- Classification-oriented validation criteria
- Similarity-oriented validation criteria.

Inter-Cluster Dissimilarity Metrics

- Distance metrics can be generalised for clusters to define inter-cluster dissimilarity measures. Let C_1 and C_2 be clusters containing n_1 and n_2 examples respectively. Some examples of distance metrics between C_1 and C_2 are:
  1. *Single linkage* is defined as
  $$d_{SL}(C_1, C_2) := \min_{\mathbf{x}^{(1)} \in C_1, \mathbf{x}^{(2)} \in C_2} Dist(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}).$$
  2. *Complete linkage* is defined as
  $$d_{CL}(C_1, C_2) := \max_{\mathbf{x}^{(1)} \in C_1, \mathbf{x}^{(2)} \in C_2} Dist(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}).$$
  3. *Group Average linkage* is defined as
  $$d_{GL}(C_1, C_2) := \frac{1}{n_1 n_2} \sum_{\mathbf{x}^{(1)} \in C_1} \sum_{\mathbf{x}^{(2)} \in C_2} Dist(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}).$$
- $Dist(\cdot)$ can be any distance function between vectors.

Hierarchical Clustering

Recall the formal algorithm of *Hierarchical Clustering*.

Algorithm 1: Hierarchical clustering.

Input: Distance matrix corresponding to the data set $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$
Output: Dendrogram.

repeat
  Find two clusters C_1, C_2 with the smallest inter-cluster dissimilarity. That is,

$$\arg\min_{C_1, C_2} d_A(C_1, C_2)$$

  where $A \in \{SL, CL, GL\}$ denotes single linkage (SL), complete linkage (CL) or group linkage (GL);
  Merge together C_1, C_2 into a single cluster;
  Note the clusters merged and their corresponding linkage $d_A(\cdot, \cdot)$ in a dendrogram.
until *Only one cluster remains.*;
return *Dendrogram.*

Exercise 1

- Consider the set with 6 examples and the following distance matrix (for some choice of distance function).

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$
$\mathbf{x}^{(1)}$	0	0.20	0.15	0.76	0.54	0.31
$\mathbf{x}^{(2)}$	0.20	0	0.89	0.18	0.66	0.27
$\mathbf{x}^{(3)}$	0.15	0.89	0	0.82	0.73	0.56
$\mathbf{x}^{(4)}$	0.76	0.18	0.82	0	0.42	0.39
$\mathbf{x}^{(5)}$	0.54	0.66	0.73	0.42	0	0.51
$\mathbf{x}^{(6)}$	0.31	0.27	0.56	0.39	0.51	0

- Use Hierarchical clustering in algorithm 1 to merge all examples into a single cluster.
- Use **single linkage** as the inter-cluster dissimilarity metric.
- Sketch the dendrogram you found along the way clearly depicting the height at which two clusters fuse.

Exercise 2

- Reconsider the set with 6 examples and the following distance matrix (for some choice of distance function).

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$
$\mathbf{x}^{(1)}$	0	0.20	0.15	0.76	0.54	0.31
$\mathbf{x}^{(2)}$	0.20	0	0.89	0.18	0.66	0.27
$\mathbf{x}^{(3)}$	0.15	0.89	0	0.82	0.73	0.56
$\mathbf{x}^{(4)}$	0.76	0.18	0.82	0	0.42	0.39
$\mathbf{x}^{(5)}$	0.54	0.66	0.73	0.42	0	0.51
$\mathbf{x}^{(6)}$	0.31	0.27	0.56	0.39	0.51	0

- Use Hierarchical clustering in algorithm 1 to merge all examples into a single cluster.
- Use **complete linkage** as the inter-cluster dissimilarity metric.
- Sketch the dendrogram you found along the way.

Cutting the Dendrogram

- In Hierarchical Clustering, we can impose a threshold on the inter-cluster distance or on the number of clusters.
- When this threshold is surpassed, the algorithm terminates without forming any further clusters, and returns the clusters formed so far.
- Different thresholds can result in different clusters.

Exercise 3

- Reconsider the set with 6 examples and the following distance matrix (for some choice of distance function).

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$	$\mathbf{x}^{(6)}$
$\mathbf{x}^{(1)}$	0	0.20	0.15	0.76	0.54	0.31
$\mathbf{x}^{(2)}$	0.20	0	0.89	0.18	0.66	0.27
$\mathbf{x}^{(3)}$	0.15	0.89	0	0.82	0.73	0.56
$\mathbf{x}^{(4)}$	0.76	0.18	0.82	0	0.42	0.39
$\mathbf{x}^{(5)}$	0.54	0.66	0.73	0.42	0	0.51
$\mathbf{x}^{(6)}$	0.31	0.27	0.56	0.39	0.51	0

- Use Hierarchical clustering but impose a threshold of 0.35 on the inter-cluster distance. Write the final clusters.
- Use Hierarchical clustering but impose a threshold of 3 on the number of clusters. Write the final clusters.
- Use **single linkage** as the inter-cluster dissimilarity metric.

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.
  $$b_i := \min_{\substack{k=1,\ldots,K \\ C_k\neq C}} \frac{\sum_{\mathbf{x}\in C_k} d(\mathbf{x}^{(i)}, \mathbf{x})}{\text{no. of examples in } C_k}.$$
- 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 $\mathcal{C}$ with N examples is defined as

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

Exercise 4

- 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 $\mathcal{C} = \{C_1, C_2\}$.
- Comment on the suitability of examples assigned to C_1.

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

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

  where $-x \log_2 x := 0$, when $x = 0$.

- 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 5

- 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. $\mathit{precision}(1, 3)$.
  2. $\mathit{recall}(1, 3)$.
  3. $F(1, 3)$.
  4. e_2.
  5. p_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 6

- 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

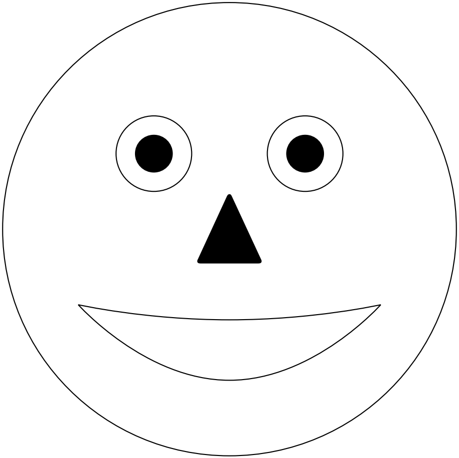
- Write down the ideal cluster similarity matrix and the ideal class similarity matrix.
- Compute the Rand statistic and the Jaccard coefficient.

Advanced Material

(OPTIONAL) Advanced Exercise 1

- Let C_1, C_2 and C_3 be clusters. Prove the following:
  1. $d_{SL}(C_1, C_2 \cup C_3) = \min\{d_{SL}(C_1, C_2), d_{SL}(C_1, C_3)\}$.
  2. $d_{CL}(C_1, C_2 \cup C_3) = \max\{d_{CL}(C_1, C_2), d_{CL}(C_1, C_3)\}$.
- Recall that
  1. $C \cup C' = \{\mathbf{x} : \mathbf{x} \in C \vee \mathbf{x} \in C'\}$.
  2. $d_{SL}(C, C') = \min_{\mathbf{x} \in C, \mathbf{x}' \in C'} Dist(\mathbf{x}, \mathbf{x}')$.
  3. $d_{CL}(C, C') = \max_{\mathbf{x} \in C, \mathbf{x}' \in C'} Dist(\mathbf{x}, \mathbf{x}')$.
  4. $Dist(\cdot, \cdot)$ is some distance function for vectors.
- Hint: $\min_{\mathbf{x} \in C \vee \mathbf{x} \in C'} f(\mathbf{x}) = \min\{\min_{\mathbf{x} \in C} f(\mathbf{x}), \min_{\mathbf{x} \in C'} f(\mathbf{x})\}$, for any sets C, C', and real-valued function $f(\cdot)$. The same holds if we replace all the min's with max's.

Until the next time...



Thank you for your attention!