

Artificial Intelligence I 2023/2024

Week 5 Tutorial and Additional Exercises

Clustering

School of Computer Science

February 13, 2024

In this tutorial...

In this tutorial we will be covering

- Unsupervised Learning.
- Distance metrics.
- Clustering.
- Advanced theoretical exercises.

Supervised and unsupervised learning

- In *supervised learning*, each available instance has a label.
- An example of supervised learning is classification.
- In *unsupervised learning*, the instances do not have labels.
- In this tutorial, we will study *clustering*, which is an unsupervised learning algorithm.

Distance metrics revisited

- Recall that a *distance metric* is a way to quantify the similarity or dissimilarity between instances.
- In this week, we will study the Chebyshev distance.
- Given two vectors with m numerical variables

$$\mathbf{x}^{(1)} = (x_1^{(1)}, \dots, x_m^{(1)}) \quad \text{and} \quad \mathbf{x}^{(2)} = (x_1^{(2)}, \dots, x_m^{(2)})$$

their *Chebyshev distance* is defined as

$$L^\infty(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \max_j |x_j^{(1)} - x_j^{(2)}|.$$

- This is a limiting case of the Minkowski distance, when taking $p \rightarrow \infty$.

Exercise 1

- Consider the following vectors with 3 numerical variables.

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0 \\ 3 \\ -1 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} -2 \\ 3 \\ -1 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

- Compute the Chebyshev distances between each pair of vectors.

Exercise 1: Solution

- The Chebyshev distances between feature vectors can be summarized in the following distance matrix:

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

- Compare this matrix with the Euclidean and Manhattan distance matrices for the same vectors from last week.
- Based on the properties of distance function reviewed in lecture, reason why the following properties hold for this matrix?
 - Symmetric (along the main diagonal);
 - Hollow (has diagonal elements of zero).

- *Clustering* is one of the most popular unsupervised learning algorithms.
- Given unlabeled instances, clustering aims at grouping together similar ones, producing clusters.
- It uses distance metrics to find similar instances and to assign instances to clusters.
- Its goal is to ensure high intra-cluster similarity and low inter-cluster similarity.
- We will next recall some basic definitions and formulas.

- Given a cluster C that consists of n vectors $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, the *centroid* of C is another vector defined as

$$\text{centroid}(C) := \frac{1}{n} \sum_{i=1}^n \mathbf{x}^{(i)}.$$

- The *inertia* of C is defined as

$$\text{inertia}(C) := \sum_{i=1}^n L^2(\mathbf{x}^{(i)}, \text{centroid}(C))^2$$

where $L^2(\cdot)^2$ is the squared Euclidean distance.

- The inertia measures how compact a cluster is.

Exercise 2

- Consider the following data set with 5 vectors and 3 variables:

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

- Treat all vectors as one cluster, C , and compute the centroid and inertia of C .

Exercise 2: Solution

- The centroid of C is

$$\text{centroid}(C) = \begin{bmatrix} 0.2 \\ 1.0 \\ 0.6 \end{bmatrix}.$$

- To find the inertia of C , we first calculate the squared Euclidean distances of each vector from the centroid.

Vector	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$L^2(\cdot)^2$	1	18	10.6	21.8	12.6

- The inertia of C is therefore

$$\text{inertia}(C) = 64.$$

Within Cluster Sum of Squares (WCSS)

- The centroid and inertia considered above are only defined for one cluster.
- If $\mathcal{C} = \{C_1, \dots, C_k\}$ is a set of several clusters, the *WCSS* of \mathcal{C} is defined as

$$WCSS(\mathcal{C}) := \sum_{j=1}^k inertia(C_j).$$

- In a clustering algorithm, we find a set of clusters that has as small WCSS as possible.

Exercise 3

- Reconsider this data set with 5 vectors and 3 variables:

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

- Assume a set of two different clusters, $\mathcal{C} = \{C_1, C_2\}$, where

$$C_1 = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}\} \text{ and } C_2 = \{\mathbf{x}^{(4)}, \mathbf{x}^{(5)}\}.$$

- Compute the centroids and inertia of C_1 and C_2 . Then, compute the WCSS of \mathcal{C} . Also compute the squared Euclidean distance between the two centroids (**do not use normalisation**).

Exercise 3: Solution

- The centroids of C_1 and C_2 are

$$\text{centroid}(C_1) = \begin{bmatrix} 0.667 \\ 3.000 \\ 1.333 \end{bmatrix} \text{ and } \text{centroid}(C_2) = \begin{bmatrix} -0.500 \\ -2.000 \\ -0.500 \end{bmatrix}.$$

- The squared Euclidean distance between the centroids is

$$L^2(\text{centroid}(C_1), \text{centroid}(C_2))^2 = 29.722.$$

- To find the inertia of C_1 and C_2 , we first calculate the squared Euclidean distances of each vector from the centroid of its cluster.

Exercise 3: Solution (continued)

- For C_1 , the squared Euclidean distances are

Vector	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$
$L^2(\cdot)^2$	5.89	5.56	3.89

- For C_2 , the squared Euclidean distances are

Vector	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$L^2(\cdot)^2$	6.5	6.5

- The inertia of C_1 and C_2 are therefore

$$inertia(C_1) = 15.33 \text{ and } inertia(C_2) = 13.$$

- Therefore, the WCSS of \mathcal{C} is

$$WCSS(\mathcal{C}) = 28.33.$$

Exercise 3: Solution (continued)

Some comments about this exercise:

- The number of clusters and the assignment of the vectors among them were fixed and were chosen arbitrarily.
- Different numbers of clusters and different assignments will result in different measures of WCSS.
- For a deeper understanding, consider these advanced tasks:
 - ① Redistribute the 5 vectors, in 2 clusters, in any way you wish, and find the WCSS. Is it lower or higher?
 - ② How many different assignments of 5 vectors can be made among 2 clusters, so that each cluster has at least 1 vector?
 - ③ What if we have 10 vectors and 2 clusters?
- The above points hint that the number of all possible cluster assignments grows very quickly as the sample size increases, and computing the WCSS of each assignment becomes prohibitively hard.

Advanced Material

(OPTIONAL) Advanced Exercise 1

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

Definition 1 (Distance metric)

A function $f : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 Chebyshev distance is a distance metric.

(OPTIONAL) Advanced Exercise 1: Solution

- Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ be arbitrary vectors. We have
 - ① If $\mathbf{x} = \mathbf{y}$, then $L^\infty(\mathbf{x}, \mathbf{y}) = \max_j |x_j - y_j| = 0$.
If $L^\infty(\mathbf{x}, \mathbf{y}) = 0$, then
 $\max_j |x_j - y_j| = 0 \Rightarrow x_1 = y_1, \dots, x_d = y_d \Rightarrow \mathbf{x} = \mathbf{y}$.
 - ② $L^\infty(\mathbf{x}, \mathbf{y}) = \max_j |x_j - y_j| = \max_j |y_j - x_j| = L^\infty(\mathbf{y}, \mathbf{x})$.
 - ③ $L^\infty(\mathbf{x}, \mathbf{z}) = \max_j |x_j - z_j| = \max_j |x_j - y_j + y_j - z_j| \leq$
 $\max_j (|x_j - y_j| + |y_j - z_j|) \leq \max_j |x_j - y_j| + \max_j |y_j - z_j| =$
 $L^\infty(\mathbf{x}, \mathbf{y}) + L^\infty(\mathbf{y}, \mathbf{z})$.
- Therefore, Chebyshev distance is a distance metric.

(OPTIONAL) Advanced Exercise 2

- Let $\mathcal{C} = \{C_1, \dots, C_k\}$ be a set of clusters and, let
 - ① n_i be the number of points in C_i , for all $i = 1, \dots, k$;
 - ② \mathbf{c}_i be the centroid of cluster C_i , for all $i = 1, \dots, k$; and
 - ③ \mathbf{c} is the centroid of all points as a single cluster.
- Also define the following:
 - ① $TSS := \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} L^2(\mathbf{x}, \mathbf{c})^2$;
 - ② $WCSS(\mathcal{C}) := \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} L^2(\mathbf{x}, \mathbf{c}_i)^2$; and
 - ③ $BCSS(\mathcal{C}) := \sum_{i=1}^k n_i L^2(\mathbf{c}_i, \mathbf{c})^2$.
- Prove the following identity:

$$TSS = WCSS(\mathcal{C}) + BCSS(\mathcal{C}).$$

- Hint: Use the fact that, for all vectors \mathbf{x}, \mathbf{y} , we have $L^2(\mathbf{x}, \mathbf{y})^2 = (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})$.

(OPTIONAL) Advanced Exercise 2: Solution I

We proceed as follows:

$$\begin{aligned}TSS &= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} L^2(\mathbf{x}, \mathbf{c})^2 \\&= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c})^T (\mathbf{x} - \mathbf{c}) \\&= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i + \mathbf{c}_i - \mathbf{c})^T (\mathbf{x} - \mathbf{c}_i + \mathbf{c}_i - \mathbf{c}) \\&= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} ((\mathbf{x} - \mathbf{c}_i)^T (\mathbf{x} - \mathbf{c}_i) + (\mathbf{c}_i - \mathbf{c})^T (\mathbf{c}_i - \mathbf{c}) \\&\quad + 2(\mathbf{c}_i - \mathbf{c})^T (\mathbf{x} - \mathbf{c}_i))\end{aligned}$$

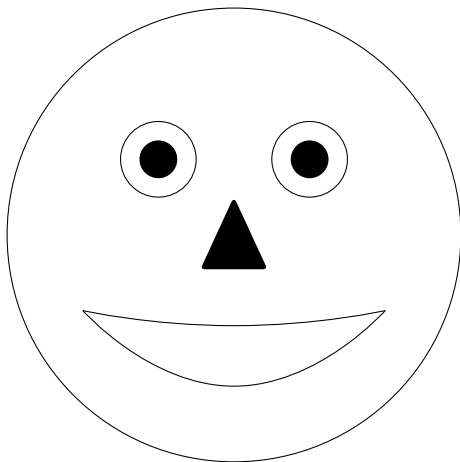
(OPTIONAL) Advanced Exercise 2: Solution II

$$\begin{aligned} &= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)^T (\mathbf{x} - \mathbf{c}_i) + \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} (\mathbf{c}_i - \mathbf{c})^T (\mathbf{c}_i - \mathbf{c}) \\ &+ 2 \sum_{i=1}^k (\mathbf{c}_i - \mathbf{c})^T \underbrace{\sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)}_0 \\ &= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)^T (\mathbf{x} - \mathbf{c}_i) + \sum_{i=1}^k n_i (\mathbf{c}_i - \mathbf{c})^T (\mathbf{c}_i - \mathbf{c}) \\ &= \sum_{i=1}^k \sum_{\mathbf{x} \in C_i} L^2(\mathbf{x}, \mathbf{c}_i)^2 + \sum_{i=1}^k n_i L^2(\mathbf{c}_i, \mathbf{c})^2 \\ &= WCSS(\mathcal{C}) + BCSS(\mathcal{C}). \end{aligned}$$



Any questions?

Until the next time...



Thank you for your attention!