# Artificial Intelligence I 2022/2023

## Week 9 Tutorial and Additional Exercises

Unsupervised Learning

School of Computer Science

April 28, 2023

# In this tutorial...

In this tutorial we will be covering

- Unsupervised Learning.
- Distance metrics.
- Normalisation.
- Clustering.
- Optional 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 cover basics of unsupervised learning algorithms.

# 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

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

  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 $p \to \infty$.

# 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 distance matrix for these vectors.
- Hint: You need to compute 6 distances on total.

# Exercise 1: Solution

- The Chebyshev distance matrix is the following:

| | $\mathbf{x}^{(1)}$ | $\mathbf{x}^{(2)}$ | $\mathbf{x}^{(3)}$ | $\mathbf{x}^{(4)}$ |
|---|---|---|---|---|
| $\mathbf{x}^{(1)}$ | 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.

# Normalisation revisited

- Recall that normalisation is a technique to reduce the effect of variables with large ranges, when calculating distances.
- This week we will study the *z-score standardisation*.
- Given a set of $n$ vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, with $m$ numerical variables, for all $j = 1, \ldots, m$, we write

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} x_j^{(i)} \text{ and } \sigma_j^2 = \frac{1}{n} \sum_{i=1}^{n} (x_j^{(i)} - \mu_j)^2.$$

- Then, the $j$-th variable of the $i$-th vector is normalised as

$$\mathit{normalise}(x_j^{(i)}) = \frac{x_j^{(i)} - \mu_j}{\sigma_j}.$$

- We calculate the above formula for all $i = 1, \ldots, n$ and for all $j = 1, \ldots, m$ and normalise all variables in all vectors.

# Exercise 2

- Consider the following vectors with 3 numerical variables.

$$\mathbf{x}^{(1)} = \begin{bmatrix} -2 \\ 3 \\ 300 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 2 \\ 1 \\ -100 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 0 \\ 2 \\ 100 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 1 \\ 2 \\ -200 \end{bmatrix}.$$

- Normalise all variables in all vectors, using the z-score standardisation.
- Hint: First compute $\mu_j$ and $\sigma_j^2$ for all $j = 1, 2, 3$. Then use the z-score standardisation formula.

# Exercise 2: Solution

- We first find that

$$\mu_1 = 0.25, \qquad \mu_2 = 2, \qquad \mu_3 = 25$$

and

$$\sigma_1^2 = 2.1875, \qquad \sigma_2^2 = 0.5, \qquad \sigma_3^2 = 36875.$$

- We then normalise all variables in all vectors as follows:

$$\tilde{\mathbf{x}}^{(1)} = \begin{bmatrix} -1.5 \\ 1.4 \\ 1.4 \end{bmatrix}, \tilde{\mathbf{x}}^{(2)} = \begin{bmatrix} 1.2 \\ -1.4 \\ -0.7 \end{bmatrix}, \tilde{\mathbf{x}}^{(3)} = \begin{bmatrix} -0.2 \\ 0 \\ 0.4 \end{bmatrix}, \tilde{\mathbf{x}}^{(4)} = \begin{bmatrix} 0.5 \\ 0 \\ -1.2 \end{bmatrix}.$$

- Compare the Euclidean distances between the normalised vectors and those between the original vectors.

# Clustering

- *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 distances 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.

# Centroid & Variability

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

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

- The *variability* of $C$ is defined as

$$variability(C) := \sum_{i=1}^{n} Dist(\mathbf{x}^{(i)}, centroid(C))$$

  where $Dist(\cdot)$ is some distance metric, e.g. the squared Euclidean distance, $L^2(\cdot)^2$.

- The variability measures how compact a cluster is.

# Exercise 3

- 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 variability of $C$, using the squared Euclidean distance for the variability.

# Exercise 3: Solution

- The centroid of $C$ is

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

- To find the variability 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 variability of $C$ is therefore

$$variability(C) = 64.$$

# Dissimilarity

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

$$\mathit{dissimilarity}(\mathbf{C}) := \sum_{j=1}^{k} \mathit{variability}(C_j).$$

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

# Exercise 4

- 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, $\mathbf{C} = \{C_1, C_2\}$, 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 variability of $C_1$ and $C_2$, using the squared Euclidean distance for the variability. Then, compute the dissimilarity of $\mathbf{C}$. Also compute the squared Euclidean distance between the two centroids (do not use normalisation).

# Exercise 4: Solution

- The centroids of $C_1$ and $C_2$ are

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

- The squared Euclidean distance between the centroids is

$$L^2(centroid(C_1), centroid(C_2))^2 = 29.722.$$

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

# Exercise 4: 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 variabilities of $C_1$ and $C_2$ are therefore

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

- Therefore, the dissimilarity of $\mathbf{C}$ is

$$\textit{dissimilarity}(\mathbf{C}) = 28.33.$$

# Exercise 4: 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 dissimilarity measures.
- For a deeper understanding, consider these optional tasks:
  1. Redistribute the 5 vectors, in 2 clusters, in any way you wish, and find the dissimilarity. 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 dissimilarities of each assignment becomes prohibitively hard.

# 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 Chebyshev distance is a distance metric.

# Optional Exercise 1: Solution

- Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{X}$ be arbitrary vectors, and $p \geq 1$. 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
     $\max_j |x_j - y_j| = 0 \Rightarrow x_1 = y_1, \ldots, x_d = y_d \Rightarrow \mathbf{x} = \mathbf{y}$.
  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 Exercise 2

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

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

- 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 Exercise 2: Solution I

We proceed as follows:

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

# Optional Exercise 2: Solution II

$$= \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)^T (\mathbf{x} - \mathbf{c}_i) + \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (\mathbf{c}_i - \mathbf{c})^T (\mathbf{c}_i - \mathbf{c})$$

$$+ 2 \sum_{i=1}^{k} (\mathbf{c}_i - \mathbf{c})^T \underbrace{\sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)}_{0}$$

$$= \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \mathbf{c}_i)^T (\mathbf{x} - \mathbf{c}_i) + \sum_{i=1}^{k} n_i (\mathbf{c}_i - \mathbf{c})^T (\mathbf{c}_i - \mathbf{c})$$

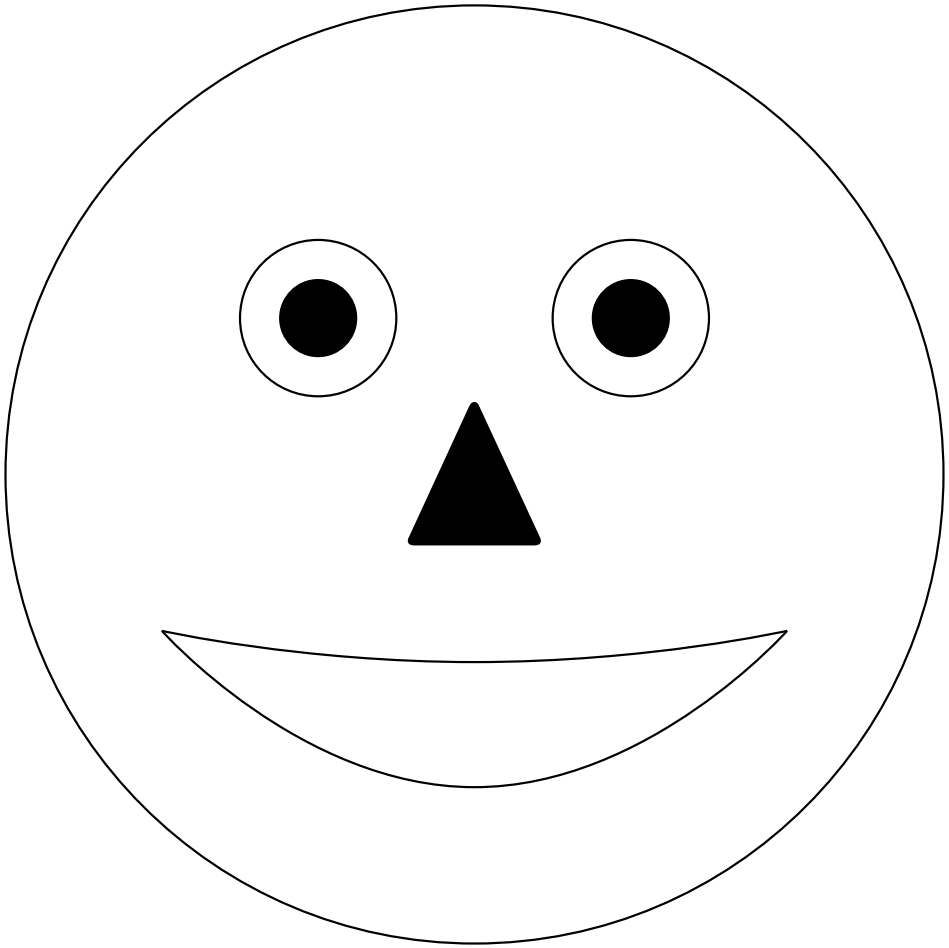
$$= \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} L^2(\mathbf{x}, \mathbf{c}_i)^2 + \sum_{i=1}^{k} n_i L^2(\mathbf{c}_i, \mathbf{c})^2$$

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

□

Any questions?

# Until the next time...



Thank you for your attention!