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

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

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 distances between each pair of 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 & Inertia

- 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 *inertia* of C is defined as

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

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.

Within Cluster Sum of Squares (WCSS)

- The centroid and inertia considered above are only defined for one cluster.
- If $\mathcal{C} = \{C_1, \ldots, C_k\}$ 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).

Advanced Material

(OPTIONAL) Advanced Exercise 1

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

Definition 1 (Distance metric)

A function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 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.

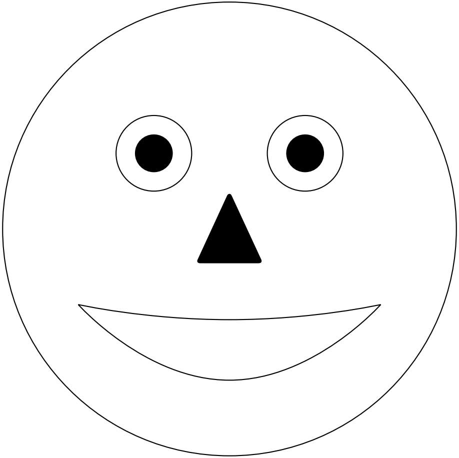
(OPTIONAL) Advanced Exercise 2

- Let $\mathcal{C} = \{C_1, \ldots, C_k\}$ 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. $WCSS(\mathcal{C}) := \sum_{i=1}^{k} \sum_{\mathbf{x} \in C_i} L^2(\mathbf{x}, \mathbf{c}_i)^2$; 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 $L^2(\mathbf{x}, \mathbf{y})^2 = (\mathbf{x} - \mathbf{y})^T(\mathbf{x} - \mathbf{y})$.

Any questions?



Thank you for your attention!