Lecture 3: Metrics and distances

Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2022-2023

Reminders on previous session

In this lecture, we will study **distances***: how can we define how "far" or "close" two individuals are ?

Question

Can anyone tell me what a **distance** is ?

Distances are FUNDAMENTAL to Machine Learning. A good understanding of distances is the first tool of any Data Scientist !

Definition

Distance

A distance* is a **numerical measurement** of how far apart objects or points are.

To be called a distance, a function $d : \mathcal{X}^2 \to \mathbb{R}$ must satisfy the following rules:

- $\forall x, y \in \mathcal{X}^2, d(x, y) = 0 \rightleftharpoons x = y$
- $\forall x, y \in \mathcal{X}^2, d(x, y) > 0$
- $\forall x, y \in \mathcal{X}^2, d(x, y) = d(y, x)$
- $\forall x, y, z \in \mathcal{X}^3, d(x, y) + d(y, z) \geq d(x, z)$ (**Triangular inequality**)

Definition

Less formally, a distance must respect:

- The distance of an object to **itself** is **always zero** and a **distance of zero** implies equality between objects.
- The distance between two **different** objects is always **strictly superior** to zero.
- A distance is **symetric**
- A distance must respect **the triangular inequality**

Exercise

Exercise

- Prove that $d : \mathbb{R}^2 \rightarrow \mathbb{R} : (x, y) = |x - y|$ is a distance.
- Generalize to the case of vectors in $\mathbb{R}^n$ and show that: $d : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} : (x, y) = \sum_{i=1}^{n} |x_i - y_i|$ is a distance (Manhattan distance).

Minkowski distance

Minkowski distance

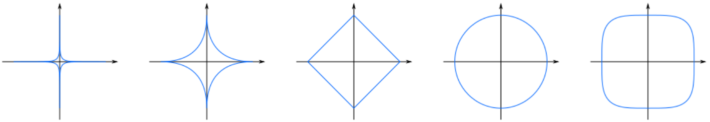
The **Minkowski*** distance of order p $(p \leq 1)$ between two vectors $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ and $y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ is defined as:

$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$

The Minkowski distance is equal to:

- $p = 1$: Manhattan distance
- $p = 2$: Euclidean distance
- $p \to \infty$: Chebychev distance

Minkowski distance



$p = 2^{-2} = 0.25$ $p = 2^{-1} = 0.5$ $p = 2^{0} = 1$ $p = 2^{1} = 2$ $p = 2^{2} = 4$

Exercise

Exercise

For x = [1, 2, 4, 5] and y = [0, 4, 3, 6], compute the Minkowski distance for :

- $p = 1$
- $p = 2$
- $p = 3$

Distance and dissimilarity

Because of a language abuse between computer scientists and mathematicians, **what is usually called a distance is not always a distance in the mathematical sense of the term**.

- The separation property is relaxed, the only requirement is that $\forall x \in \mathcal{X}\ d(x, x) = 0$ (the distance of an object to itself is always zero).
- The triangular inequality is not always verified

But every library will refer to them as distance ! To avoid confusion, we will call them **similarity/dissimilarity**.

Cosine similarity

Cosine similarity: cosine of the angle between products (dot product of the vectors divided by the product of their length, because $x \cdot y = ||x|| \times ||y|| \times cos(x, y)$)

Cosine dissimilarity

The **Cosine*** dissimilarity between two vectors $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ and $y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$ is defined as:

$d(x, y) = 1 - \frac{x \cdot y}{||x|| \times ||y||} = 1 - \frac{\sum_{i=1}^n x_i \times y_i}{\sum_{i=1}^n x_i^2 \times \sum_{i=1}^n y_i^2}$

Question

Why can't the cosine distance be considered a distance in the mathematical sense of the term ?

Exercise

Exercise

For $x = [1, 2, 4, 5]$ and $y = [0, 4, 3, 6]$, compute the cosine distance.

To go further...

If you want to explore more distances (it will help you for the lab sessions), checkout the *pairwise_distances* module from *sklearn*
`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise_distances.html`

Questions

Questions ?