Distances

In any data analysis project, it is necessary to be able to quantify the similarity (or dissimilarity) between two observations. To do this, we use the concept of **distance** (or **similarity**) between observations. The choice of this distance directly influences the results of learning, clustering, and visualization algorithms.

1 Concept of distance

A **distance** is a mathematical function that measures how far two objects are from each other in a given space. The greater the distance, the further apart the observations are.

Definition of distance measurement

A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a distance on a set $\mathcal{X}$ if, for all $x, y, z \in \mathcal{X}$, the following conditions are satisfied:

1. non-negativity: $d(x, y) \geq 0$;
2. separation: $d(x, y) = 0 \Leftrightarrow x = y$;
3. symmetry: $d(x, y) = d(y, x)$;
4. triangle inequality: $d(x, y) \leq d(x, z) + d(y, z)$.

Euclidean distance

When observations are represented by numerical vectors in $\mathbb{R}^p$ of the same order of magnitude, **Euclidean distance** is often a good choice.
Let $x, y \in \mathbb{R}^p$, the Euclidean distance is given by:

$$d(x, y) = \|x - y\|_2 = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{1/2}.$$

The L_q (or Minkowski) distance

Let $x, y \in \mathbb{R}^p$, the distance L_q is given, for $q > 0$, by:

$$d(x, y) = \|x - y\|_q = \left(\sum_{i=1}^{p} |x_i - y_i|^q \right)^{1/q}.$$

Special cases:

- For $q = 1$, we obtain the Manhattan distance:

$$d(x, y) = \|x - y\|_1 = \sum_{i=1}^{p} |x_i - y_i|.$$

- For $q = 2$, we obtain the Euclidean distance.

Example

Consider the following dataset:

Table 1: Average height and weight in Canada (Source: Statistics Canada, Canadian Community Health Survey (2008)).

Name	Height	Weight
Alice	162.1	66.8
Bob	175.8	81.6

The Euclidean distance between Alice and Bob is

$$d(\text{Alice}, \text{Bob}) = \sqrt{(162.1 - 175.8)^2 + (66.8 - 81.6)^2} = 20.16.$$

The Manhattan distance between Alice and Bob is

$$d(\text{Alice}, \text{Bob}) = |162.1 - 175.8| + |66.8 - 81.6| = 28.5.$$

The distance L_q is not invariant to changes in scale. For example, if we multiply all the components of a vector by a factor of , the distance between two vectors changes by a factor of .

In practice, we prefer to work with standardized variables. Thus, denoting μ_i as the mean, and σ_i as the standard deviation of variable i, the Euclidean distance with standardized variables

is given by:

$$d(x,y) = \sum_{i=1}^{p} \left\{ \frac{x_i - \mu_i}{\sigma_i} - \frac{y_i - \mu_i}{\sigma_i} \right\}^2 = \sum_{i=1}^{p} \left(\frac{x_i - y_i}{\sigma_i} \right)^2.$$

Property

The Euclidean distance with standardized variables is invariant under changes in scale.

Proof

Let $\lambda \neq 0$ and let X be a random variable. We have $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ and $\text{Var}(\lambda X) = \lambda^2 \text{Var}(X)$. So

$$d(\lambda x, \lambda y) = \sum_{i=1}^{p} \left\{ \frac{\lambda x_i - \lambda \mu_i}{\lambda \sigma_i} - \frac{\lambda y_i - \lambda \mu_i}{\lambda \sigma_i} \right\}^2 = \sum_{i=1}^{p} \left(\frac{x_i - y_i}{\sigma_i} \right)^2 = d(x,y).$$

2 Notion of Similarity

In contrast to the notion of distance, a **similarity measure** quantifies how close two observations are in a given space. Thus, the greater the similarity, the closer the observations are.

Similarity Measure Definition

A function $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a similarity measure on a set $\mathcal{X}$ if, for all $x, y \in \mathcal{X}$, the following conditions hold:

1. $s(x,y) \geq 0$;
2. $s(x,y) = s(y,x)$;
3. $s(x,x) = 1 \geq s(x,y)$.

A distance can be transformed into a similarity by setting

$$s(x,y) = \frac{1}{1 + d(x,y)}.$$

This transformation ensures that the greater the distance, the lower the similarity. However, the reverse is not always possible because a similarity measure does not necessarily respect the triangle inequality. We can also define the dissimilarity between two objects:

$$d^\star(x,y) = 1 - s(x,y).$$

3 Case of qualitative variables

When working with qualitative variables, the usual numerical distances (such as L_p distances) are generally meaningless. For example, if a variable takes its values from the set

$$\mathcal{X} = \{\text{Red}, \text{Green}, \text{Blue}\},$$

then it makes no sense to calculate the difference between Blue and Red, because these categories carry no intrinsic numerical structure and have no notion of order or distance.

Bad practice would be to arbitrarily assign numerical values to the variables (e.g., Red = 1, Green = 2, Blue = 3), which would introduce an artificial order between them. This could significantly bias the analyses.

3.1 One-of-K Encoding

When we want to use a model based on a notion of distance between observations (i.e., most models), we must use a suitable encoding. One-of-K encoding (*one-hot encoding*) consists of encoding a categorical variable with K categories as a binary vector of dimension K, in which only one entry is 1, the others are 0. Thus, for the example of $\mathcal{X} = \{\text{Red, Green, Blue}\}$, we will obtain the following encoding: "Red" gives $(1, 0, 0)$, "Green" gives $(0, 1, 0)$, and "Blue" gives $(0, 0, 1)$.

This encoding method has the advantage of not introducing an artificial order between the modalities. However, if the variable has many modalities, the representation space will be large, which can hinder the effectiveness of some analysis methods.

3.2 Define an appropriate distance

Once the modalities have been encoded, we can define a distance between two observations of qualitative variables.

Discrete Distance (or Hamming Distance)

Let $\mathcal{X}$ be a discrete set and let x and y be two observations of X. The **discrete distance** is given by

$$d(x, y) = \begin{cases} 0, & \text{if } x = y, \\ 1, & \text{if } x \neq y \end{cases}.$$

For vectors of qualitative variables (e.g., the comparison of several individuals described by several characteristics), the discrete distance is the **sum of disagreements** between the components:

$$d(x, y) = \sum_{i=1}^{p} \mathbb{1}(x_i \neq y_i),$$

where p is the number of variables.

Example

Consider the characteristics of three people.

Table 2: Characteristics of three people.

Name	Color	Eyes	Hair
Alice	Red	Blue	Blonde
Bob	Green	Blue	Red
Chris	Red	Green	Blonde

The distance between two people is calculated as the number of different characteristics. Thus,

$$d(\text{Alice}, \text{Bob}) = 1 + 0 + 1 = 2,$$

$$d(\text{Alice}, \text{Chris}) = 0 + 1 + 0 = 1,$$

$$d(\text{Bob}, \text{Chris}) = 1 + 1 + 1 = 3.$$

Rather than counting the differences, we can also count the agreements and normalize them:

$$s(x, y) = \frac{1}{p} \sum_{i=1}^{p} \mathbb{1}(x_i = y_i),$$

which gives a similarity measure between 0 (no agreement) and 1 (identical).

Example

Continuing the previous example (see Table 2), we find the following similarities:

$$s(\text{Alice}, \text{Bob}) = \frac{0 + 1 + 0}{3} = \frac{1}{3},$$

$$s(\text{Alice}, \text{Chris}) = \frac{1+0+1}{3} = \frac{2}{3},$$

$$s(\text{Bob}, \text{Chris}) = \frac{0+0+0}{3} = 0.$$

3.3 Jaccard Distance

When the number of binary variables is large (e.g., in the case where a 1 out of K encoding has been performed on p qualitative variables), the discrete distance is not necessarily very suitable because the number of agreements is likely to be small compared to the total number of variables ($K \times p in the previous example), which will result in small distances in all cases. One solution is to focus only on attributes that equal 1, because generally, a binary variable equal to 0 does not provide any particular information (at least, less than a binary variable equal to 1). The Jaccard index (Intersection over Union, IoU) was introduced to take this into account.

Definition: Jaccard Index

Consider two observations x and y of K binary variables. All variables can take the values 0 and 1.
Let's define the following quantities:

- M_{11}, the number of 1 variables for x and y;
- M_{10}, the number of 1 variables for x and 0 for y;
- M_{01}, the number of 0 variables for x and 1 for y;
- M_{00}, the number of 0 variables for x and y.

Since each binary variable is necessarily counted in either M_{11}, M_{10}, M_{01}, or M_{00}, their sum is therefore equal to K.
The **Jaccard index** is defined as

$$J(x, y) = \frac{M_{11}}{M_{10} + M_{01} + M_{11}} = \frac{M_{11}}{K - M_{00}}.$$

Paying attention to the case where the two observations consist only of 0 (we take $J(x, y) = 1$ in this case), the Jaccard index is a measure of similarity.

Property: Jaccard Distance

For two observations x and y of K binary variables, the **Jaccard distance** is given by

$$d(x, y) = 1 - J(x, y).$$

Proof

To show that the Jaccard distance is indeed a distance, we must demonstrate the four properties of distances. First, note that the Jaccard distance can be rewritten as

$$d(x, y) = \frac{M_{10} + M_{01}}{M_{01} + M_{10} + M_{11}}.$$

1. All terms in the numerator and denominator are positive, so $d(x, y) \geq 0$.
2. Let us show that $d(x, y) = 0 \Leftrightarrow x = y$.

Suppose that $d(x, y) = 0$. Then $M_{01} + M_{10} = 0$. Therefore, there are no variables that equal 0 for x and 1 for y, and vice versa. Since $M_{01} + M_{10} + M_{11} > 0$, we have $x = y$. Now, suppose that $x = y$. Then $M_{01} = M_{10} = 0$. Therefore $d(x, y) = 0$.

3. We have that $d(x, y) = d(y, x)$ because the Jaccard index is symmetric.
4. The proof of the triangle inequality will be done in an exercise (see TD)).

Example

Consider a questionnaire with 5 closed questions. Suppose the answer "Yes" is coded as 1 and the answer "No" is coded as 0.

Table 3: Characteristics of two people.

Name	Q1	Q2	Q3	Q4	Q5
Alice	1	0	1	0	0
Bob	1	0	0	1	0

For the distance between Alice and Bob, we have $M_{11} = 1$, $M_{10} = 1$, $M_{01} = 1$, and $M_{00} = 2$. Therefore, the Jaccard similarity is given by $J(\text{Alice}, \text{Bob}) = \frac{1}{3}$. Thus, the Jaccard distance is $d(\text{Alice}, \text{Bob}) = 1 - J(x, y) = \frac{2}{3}$.