Exercices SY09
Centrage et distance

1 Centrage

Exercice 1 Soit x un vecteur de taille n. On note $\overline{x}$ la moyenne empirique de l'échantillon x,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

et on dit que x est centré lorsque $\overline{x} = 0$.

1. Montrer que x est un vecteur centré si et seulement si $\mathbb{1}_n^T x = 0$ avec $\mathbb{1}_n^T = [1, \dots, 1]$.

Corrigé

$$\mathbb{1}_n^T x = 0 \Longleftrightarrow \begin{pmatrix}1 \dots 1\end{pmatrix}\begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix} = 0 \Longleftrightarrow \sum_{i=1}^{n} x_i = 0 \Longleftrightarrow \overline{x} = 0$$

2. On pose $x^\dagger = x - \overline{x}\mathbb{1}_n$. Montrer que $x^\dagger$ est centré.

Corrigé

D'après la question précédente

$$\begin{aligned}\mathbb{1}_n^T x^\dagger &= \mathbb{1}_n^T (x - \overline{x}\mathbb{1}_n)\\ &= \mathbb{1}_n^T x - \overline{x}\mathbb{1}_n^T\mathbb{1}_n\\ &= n\overline{x} - n\overline{x}\\ &= 0.\end{aligned}$$

Donc $x^\dagger$ est centré.

On appellera $x^\dagger$ le centrage de x.

3. Montrer que le centrage est une opération linéaire :
   (a) $(x + y)^\dagger = x^\dagger + y^\dagger$
   (b) $(\lambda x)^\dagger = \lambda x^\dagger$

Corrigé

$$\begin{aligned}(x+y)^\dagger &= x+y-\overline{x+y}\mathbb{1}_n \\ &= x+y-\overline{x}\mathbb{1}_n-\overline{x}\mathbb{1}_n \\ &= x^\dagger+y^\dagger\end{aligned}$$

4. Montrer que $x^\dagger = 0$ si et seulement si $x = \lambda\mathbb{1}_n$ avec $\lambda \in \mathbb{R}$.

Corrigé

$$\begin{aligned}x^\dagger = 0 &\iff x = \overline{x}\mathbb{1}_n \\ &\iff x = \lambda\mathbb{1}_n\end{aligned}$$

Autrement dit, le centrage d'un vecteur n'est jamais nul sauf si le vecteur est constant.

5. Montrer que x et y ont même centrage si et seulement si il existe $\lambda \in \mathbb{R}$ tel que $y = x + \lambda\mathbb{1}_n$.

Corrigé

En utilisant la linéarité, on a successivement

$$\begin{aligned}x^\dagger = y^\dagger &\iff (y-x)^\dagger = 0 \\ &\iff y-x = \lambda\mathbb{1}_n \\ &\iff y = x+\lambda\mathbb{1}_n\end{aligned}$$

6. Montrer que

$$\begin{pmatrix} x \\ y \end{pmatrix}^\dagger = \begin{pmatrix} x^\dagger \\ y^\dagger \end{pmatrix} \quad \text{si et seulement si} \quad \overline{x} = \overline{y}.$$

Corrigé

On suppose x de taille n et y de taille m et on pose $z = \begin{pmatrix} x \\ y \end{pmatrix}$. Par la suite, on a successivement

$$\begin{aligned}\begin{pmatrix} x \\ y \end{pmatrix}^\dagger = \begin{pmatrix} x^\dagger \\ y^\dagger \end{pmatrix} &\iff z-\overline{z}\mathbb{1}_{n+m} = \begin{pmatrix} x^\dagger \\ y^\dagger \end{pmatrix} \\ &\iff \begin{pmatrix} x-\overline{z}\mathbb{1}_n \\ y-\overline{z}\mathbb{1}_m \end{pmatrix} = \begin{pmatrix} x-\overline{x}\mathbb{1}_n \\ y-\overline{y}\mathbb{1}_m \end{pmatrix} \\ &\iff \begin{pmatrix} \overline{z}\mathbb{1}_n \\ \overline{z}\mathbb{1}_m \end{pmatrix} = \begin{pmatrix} \overline{x}\mathbb{1}_n \\ \overline{y}\mathbb{1}_m \end{pmatrix} \\ &\iff \overline{x} = \overline{y}\end{aligned}$$

2 Distance

Exercice 2 Préciser la proximité que représente (ou pas) chacune des matrices suivantes.

1. $\begin{pmatrix} 0 & 1 & 2 \\ 3 & 0 & -1 \\ 1 & 2 & 0 \end{pmatrix}$
2. $\begin{pmatrix} 1 & 3 & 2 \\ 3 & 1 & 2 \\ 2 & 2 & 1 \end{pmatrix}$
3. $\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$
4. $\begin{pmatrix} 0 & 2 & 3 \\ 2 & 0 & 1 \\ 4 & 1 & 0 \end{pmatrix}$
5. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
6. $\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$
7. $\begin{pmatrix} 10 & 1 & 3 \\ 6 & 9 & 2 \\ 5 & 0 & 8 \end{pmatrix}$
8. $\begin{pmatrix} 10 & 4 & -1 \\ 4 & 10 & 5 \\ -1 & 5 & 10 \end{pmatrix}$

Corrigé

1. N'est pas une proximité car une entrée est négative.
2. N'est ni une similarité ni une dissimilarité car la diagonale n'est pas dominante ni égale à zéro.
3. Matrice symétrique à diagonale nulle, c'est une dissimilarité. C'est en fait aussi une distance, une ultramétrique et une distance euclidienne.
4. Matrice à diagonale nulle mais pas symétrique.
5. Matrice de similarité car positive, symétrique à diagonale constante et dominante.
6. Matrice positive, symétrique, diagonale constante et dominante, c'est une similarité.
7. Diagonale non constante, ce n'est pas une similarité.
8. N'est pas une proximité car une entrée est négative.

Exercice 3 Soit Ω est ensemble muni d'une ultramétrique d. Montrer que tout triangle est soit équilatéral soit isocèle avec une petite base.

Corrigé

Soit x, y, z trois points de Ω. On pose $a = d(x, y)$, $b = d(y, z)$ et $c = d(x, z)$. Sans perte de généralité on peut supposer que $a \leq b \leq c$. En appliquant l'inégalité ultramétrique à deux reprises :

$$\begin{aligned} d(x,z) \leq \max(d(x,y), d(y,z)) \quad &\Longleftrightarrow \quad c \leq \max(a,b) = b \\ d(y,z) \leq \max(d(y,x), d(x,z)) \quad &\Longleftrightarrow \quad b \leq \max(a,c) = c \end{aligned}$$

On a donc $b = c$. Si $a < b$, le triangle est isocèle avec une petite base. Si $a = b$, le triangle est équilatéral.

Exercice 4 Montrer que la distance qui vaut tout le temps 1 sauf pour deux éléments identiques où elle vaut 0 est une distance ultramétrique.

Corrigé

La distance d est donc définie par

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

(Symétrie) : On a bien évidemment $d(x,y) = d(y,x)$ dès que $x = y$. Si $x \neq y$ on a alors $d(x,y) = 1 = d(y,x)$. La distance d est donc symétrique.
(Séparation) : Par définition, on a

$$d(x,y) = 0 \quad \Longleftrightarrow \quad x = y,$$

ce qui est exactement la propriété de séparation.
(Inégalité ultramétrique) : Soit x, y et z trois éléments de Ω. Dès que les éléments ne sont pas tous distints, l'inégalité ultramétrique est trivialement vérifiée. Dans le cas contraire, si x, y et z sont distints, on a $d(x,y) = d(y,z) = d(x,z) = 1$ d'après la définition. L'inégalité ultramétrique est encore vérifiée. En conclusion la distance est ultramétrique.

Exercice 5 On suppose que $X_1, \ldots, X_n$ sont n caractéristiques binaires d'une population Ω. On note x_i la i-ième caractéristique de l'individu x et on considère la similarité entre deux individus x et y suivante,

$$s(x,y) = \frac{a+d}{a+d+b+c},$$

où

$$\begin{aligned} a &= \text{card}\{i, x_i = 1, y_i = 1\} \\ d &= \text{card}\{i, x_i = 0, y_i = 0\} \\ b &= \text{card}\{i, x_i = 0, y_i = 1\} \\ c &= \text{card}\{i, x_i = 1, y_i = 0\} \end{aligned}$$

On pose $d = 1 - s$.

1. Montrer que

$$d(x,y) = \frac{b+c}{n}$$

Corrigé

Il suffit de remarquer que $a + b + c + d = n$.

2. Montrer que d vérifie les propriétés de séparation et de symétrie.

Corrigé

L'expression $b + c$ est symétrique en x et y d'où la symétrie de d. Pour la séparation,

$$\begin{aligned} d(x,y) = 0 \quad &\Longleftrightarrow \quad b = c = 0 \\ &\Longleftrightarrow \quad x = y \end{aligned}$$

3. On note

$$\begin{aligned}
A &= \operatorname{card}\{i, x_i = 0, y_i = 0, z_i = 0\}\\
B &= \operatorname{card}\{i, x_i = 0, y_i = 0, z_i = 1\}\\
C &= \operatorname{card}\{i, x_i = 0, y_i = 1, z_i = 0\}\\
D &= \operatorname{card}\{i, x_i = 0, y_i = 1, z_i = 1\}\\
E &= \operatorname{card}\{i, x_i = 1, y_i = 0, z_i = 0\}\\
F &= \operatorname{card}\{i, x_i = 1, y_i = 0, z_i = 1\}\\
G &= \operatorname{card}\{i, x_i = 1, y_i = 1, z_i = 0\}\\
H &= \operatorname{card}\{i, x_i = 1, y_i = 1, z_i = 1\}
\end{aligned}$$

(a) Exprimer $d(x, y)$, $d(y, z)$ et $d(x, z)$ en fonction de A, B, C, D, E, F, G et H.

Corrigé

On a

$$\begin{aligned}
d(x, y) &= \frac{C + D + E + F}{n}\\
d(x, z) &= \frac{B + D + E + G}{n}\\
d(y, z) &= \frac{B + C + F + G}{n}
\end{aligned}$$

On en déduit

$$d(x, y) + d(y, z) - d(x, z) = 2\frac{D + E}{n} \geq 0,$$

d'où l'inégalité triangulaire.

(b) En déduire que d est une distance

Corrigé

La proximité d vérifie les propriétés de symétrie, de séparation et l'inégalité triangulaire, c'est donc une distance.