Distances

Introduction

On a besoin d'une notion de distance et de similarité. On a besoin d'une définition de ce que sont des observations similaires ou des observations différentes. On veut quantifier la similarité ou la distance entre deux observations. Le choix de la métrique est une étape essentielle dans le processus d'analyse de données.

Notion de distance

Définition de distance

Une mesure de distance d doit satisfaire les propriétés suivantes pour tout $x, y, z \in A$, A étant un ensemble quelconque :

1. $d(x, y) \geq 0$;
2. $d(x, x) = 0$;
3. $d(x, y) = d(y, x)$;
4. $d(x, y) \leq d(x, z) + d(y, z)$.

La distance euclidienne :

Si les observations sont constituées de p nombres réels de même ordre de grandeur, alors la distance euclidienne entre deux éléments de $\mathbb{R}^p$ est une mesure raisonnable.

Soit $x, y \in \mathbb{R}^n$, la distance euclidienne est données par :

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

Que fait-on dans d'autres cas ? Plusieurs mesures ont été développées pour leur application particulière. Voici les plus classiques.

Une distance est peu plus générale que la distance euclidienne est la distance L_p. Soit $x, y \in \mathbb{R}^n$, la distance L_p est donnée, pour $p > 0$, par :

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

La distance euclidienne correspond à la distance L_p avec $p = 2$. Cas particulier avec $p = 1$, distance de Manhattan :

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

Properties

La distance L_p n'est pas invariante à un changement d'échelle, *i.e.*, $L_p(x, y) \neq L_p(\lambda x, \lambda y)$.

Standardisation

Notion de similarité

Définition d'un indice de similarité

Un indice de similarité s doit satisfaire les propriétés suivantes pour tout $x, y \in A$, A étant un ensemble quelconque :

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

Une distance peut se transformer en similarité en posant $s(x, y) = \frac{1}{1+d(x,y)}$. L'inverse n'est pas vrai, dû à l'inégalité triangulaire. On peut aussi définir la dissemblance entre deux objects : $d^\star(x, y) = 1 - s(x, y)$.