Distances

Dans tout projet d'analyse de données, il est nécessaire de pouvoir quantifier la ressemblance (ou la dissemblance) entre deux observations. Pour cela, on utilise la notion de **distance** (ou **similarité**) entre les observations. Le choix de cette distance influence directement les résultats des algorithmes d'apprentissage, de regroupement et de visualisations.

Notion de distance

Une **distance** est une fonction mathématique mesurant à quel point deux objets sont éloigné l'un de l'autre dans un espace donnée. Plus la distance est grande, plus les observations somt éloigné.

Définition de mesure de distance

Une fonction $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une distance sur un ensemble $\mathcal{X}$ si, pour tout $x, y, z \in \mathcal{X}$, les conditions suivantes sont vérifiées :

1. non-négativité : $d(x, y) \geq 0$;
2. séparation : $d(x, y) = 0 \Leftrightarrow x = y$;
3. symétrie : $d(x, y) = d(y, x)$;
4. inégalité triangulaire : $d(x, y) \leq d(x, z) + d(y, z)$.

La distance euclidienne

Lorsque les observations sont représentées par des vecteurs numériques dans $\mathbb{R}^p$ de même ordre de grandeur, la **distance euclidienne** est souvent un bon choix.
Soit $x, y \in \mathbb{R}^p$, la distance euclidienne est données par :

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

La distance L_q (ou de Minkowski)

Soit $x, y \in \mathbb{R}^p$, la distance L_q est donnée, pour $q > 0$, par :

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

Cas particuliers :

— Pour $q = 1$, on obtient la distance de Manhattan :

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

— Pour $q = 2$, on obtient la distance euclidienne.

Exemple

Considérons le jeu de données suivant :

TABLE 1 – Taille et poids moyens au Canada (Source : Statistique Canada, Enquête sur la santé dans les collectivités canadiennes (2008)).

Nom	Taille	Poids
Alice	162.1	66.8
Bob	175.8	81.6

La distance euclidienne entre Alice et Bob est

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

La distance de Manhattan entre Alice et Bob est

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

La distance L_q n'est pas invariante aux changements d'échelle. Par exemple, si on multiplie toutes les composantes d'un vecteur par un facteur λ, la distance entre deux vecteurs change du facteur λ.

En practique, on préfère travailler avec des variables standardisées. Ainsi, en notant, μ_i, la moyenne, et σ_i, l'écart-type de la variable i, la distance euclidienne avec des variables standar-

disées est donnée par :

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

Propriété

La distance euclidienne avec des variables standardisées est invariante par changement d'échelle.

Preuve

Soit $\lambda \neq 0$ et soit X une variable aléatoire. On a $\mathbb{E}(\lambda X) = \lambda \mathbb{E}(X)$ et $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X)$. Donc

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

Notion de similarité

À l'opposé de la notion de distance, une **mesure de similarité** quantifie à quel point deux observations sont proches dans un espace donné. Ainsi, plus la similarité est grande, plus les observations sont proches.

Définition de mesure de similarité

Une fonction $s : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une mesure de similarité sur un ensemble $\mathcal{X}$ si, pour tout $x, y \in \mathcal{X}$, les conditions suivantes sont vérifiées :

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

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

Cette transformation garantit que plus la distance est grande, plus la similarité est faible. Toutefois, l'inverse n'est pas toujours possible car une mesure de similarité ne respecte pas forcément l'inégalité triangulaire. On peut aussi définir la dissemblance entre deux objets :

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