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 :

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

Cas de variables qualitatives

Lorsque l'on travaille avec des variables qualitatives, les distances numériques habituelles (comme les distances L_p) n'ont généralement pas de sens. Par exemple, si une variable prend ses valeurs dans l'ensemble

$$\mathcal{X} = \{\text{Rouge}, \text{Vert}, \text{Bleu}\},$$

alors il n'y a pas de sens à calculer la différence entre Bleu et Rouge, car ces modalités ne portent aucune structure numérique intrinsèque et n'ont aucune notion d'ordre ou d'écart.

Une mauvaise pratique consisterait à attribuer arbitrairement des valeurs numériques aux momdalités (e.g. Rouge = 1, Vert = 2, Bleu = 3) ce qui introduirait un ordre artificiel entre elles. Cela risquerait de biaiser fortement les analyses.

Encodage 1 parmi K

Lorsque l'on veut utiliser un modèle se basant sur une notion de distance entre les observations (i.e. la plupart des modèles), on doit utiliser un encodage adapté. L'**encodage 1 parmi K** (*one-hot encoding*) consiste à encoder une variable qualitative à K modalités sous la forme d'un vecteur binaire de dimension K, dans lequel une seule entrée est à 1, les autres à 0. Ainsi, pour l'exemple de $\mathcal{X} = \{\text{Rouge, Vert, Bleu}\}$, on obtiendra l'encodage suivant : "Rouge" donne $(1, 0, 0)$, "Vert" donne $(0, 1, 0)$ et "Bleu" donne $(0, 0, 1)$.

Cette méthode d'encodage a l'avantage de ne pas introduire d'ordre artificiel entre les modalités. Si la variable a beaucoup de modalités, l'espace de représentation de grande dimension, ce qui peut nuire à l'efficacité de certaines méthodes d'analyse.

Définir une distance adaptée

Une fois les modalités encodées, on peut définir une distance entre deux observations de variables qualitatives.

La distance discrète (ou distance de Hamming)

Soit $\mathcal{X}$ un ensemble discret et soient x et y deux observations de X, la **distance discrète** est donnée par

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

Pour des vecteurs de variables qualitatives (e.g. la comparaison de plusieurs individus décrits par plusieurs caractéristiques), la distance discrète est la **somme des désaccords** entre les

composantes :

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

où p est le nombre de variables.

Exemple

Prenons les caractéristiques de trois personnes.

TABLE 2 – Caractéristiques de trois personnes.

Nom	Couleur	Yeux	Cheveux
Alice	Rouge	Bleu	Blond
Bob	Vert	Bleu	Roux
Chris	Rouge	Vert	Blond

On calcule la distance entre deux personnes comme le nombre de caractéristiques différentes. Ainsi,

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

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

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

Plutôt que de compter les différences, on peut aussi compter les accords et les normaliser :

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

ce qui donne une mesure de similarité comprise entre 0 (aucun accord) et 1 (identique).

Exemple

En reprenant l'exemple précédent (cf. Table 2), on trouve les similarités suivantes :

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

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

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

Distance de Jaccard

La distance de Jaccard est une mesure très utilisée dans le cas d'observations binaires. Elle prend mieux en compte ces cas par rapport à la distance discrète.

Définition

Considérons deux observations X et Y de K variables binaires. Chaques variables peut prendre les valeurs 0 et 1.
Définissons les quantités suivantes :

- M_{11}, le nombre de variables à 1 pour X et Y ;
- M_{10}, le nombre de variables à 1 pour X et 0 pour Y ;
- M_{01}, le nombre de variables à 0 pour X et 1 pour Y ;
- M_{00}, le nombre de variables à 0 pour X et Y.

Chaque variable binaire est forcément comptée, soit dans M_{11}, soit dans M_{10}, soit dans M_{01}, soit dans M_{00}. Donc $K = M_{11} + M_{10} + M_{01} + M_{00}$.
La **distance de Jaccard** est définie comme

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

(Est-ce qu'elle est définie comme ça, ou bien c'est 1 - ça ?)

Exemple

Prenons un questionnaire de 5 questions fermées. Supposons que la réponse "Oui" est encodée par 1 et la réponse "Non" est encodée par 0.

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

Pour la distance entre Alice et Bob, on a $M_{11} = 1$, $M_{10} = 1$, $M_{01} = 1$ et $M_{00} = 2$. Donc $d(\text{Alice}, \text{Bob}) = \frac{1}{3}$.