DIFIQ
Niveau 2
Statistiques
Les Métriques

Jean-François Berger-Lefébure

21 Novembre 2024

Contents

1 Métriques

1.1 Dissimilarité, distance et distance ultramétrique

1.1.1 Notion de métrique

Une **métrique** est une fonction d qui mesure la distance entre deux individus dans un espace donné.
Définition: La métrique d est définie sur l'ensemble des couples d'individus $\mathcal{O} \times \mathcal{O}$, c'est-à-dire qu'elle associe une distance entre chaque paire d'individus.

—

1.2 Propriétés fondamentales d'une métrique

Pour trois individus quelconques (i_1, i_2, i_3) dans l'espace, la métrique doit vérifier les conditions suivantes:

Propriété 1: Non-négativité

$$d(i_1, i_2) \geq 0$$

Lecture: La distance entre deux points est toujours positive ou nulle.
Interprétation: Il est impossible d'avoir une distance négative entre deux individus. Si les points sont identiques, la distance est nulle.

—

Propriété 2: Symétrie

$$d(i_1, i_2) = d(i_2, i_1)$$

Lecture: La distance entre deux points ne dépend pas de l'ordre dans lequel ils sont considérés.
Interprétation: La distance de i_1 à i_2 est la même que celle de i_2 à i_1.

—

Propriété 3: Identité des indiscernables

$$d(i_1, i_2) = 0 \Rightarrow i_1 = i_2$$

Lecture: Deux individus sont identiques si et seulement si leur distance est nulle.
Interprétation: Une distance nulle signifie que les deux individus occupent exactement la même position dans l'espace.

—

Propriété 4: Inégalité triangulaire

$$d(i_1, i_2) \leq d(i_1, i_3) + d(i_3, i_2)$$

Lecture: La distance directe entre deux points est toujours inférieure ou égale à la distance passant par un troisième point.
Interprétation: Imaginez un triangle formé par trois points. La somme des deux côtés est toujours plus grande ou égale à la longueur du troisième côté.

—

Propriété 5: Inégalité ultratriangulaire

$$d(i_1, i_2) \leq \max\{d(i_1, i_3), d(i_3, i_2)\}$$

Lecture: La distance entre deux points est toujours inférieure ou égale à la plus grande des distances obtenues en passant par un troisième point.
Interprétation: Cette condition est plus forte que l'inégalité triangulaire classique et impose une structure rigide aux distances, typique des arbres hiérarchiques.

—

1.2.1 Différenciation des concepts

- **Dissimilarité:** Vérifie uniquement les propriétés (1), (2) et (3).
- **Distance:** Vérifie les propriétés (1), (2), (3) et (4).
- **Distance ultramétrique:** Vérifie les propriétés (1), (2), (3) et (5).

—

1.2.2 Applications pratiques

- **Dissimilarité:** Utilisée pour mesurer une différence qualitative entre des objets (par exemple, similitude entre textes ou ADN).
- **Distance:** Appropriée pour des mesures géométriques standard (par exemple, distance euclidienne).
- **Distance ultramétrique:** Utilisée pour des structures en arbre, comme en classification hiérarchique (CAH). Elle modélise des relations parent-enfant strictes.

—

1.2.3 Questions pour réviser

Question 1: Quelles propriétés sont vérifiées par une dissimilarité ?
Réponse: Non-négativité, symétrie et identité des indiscernables.
Question 2: Quelles propriétés définissent une distance mais pas une dissimilarité ?
Réponse: La distance respecte en plus l'inégalité triangulaire.
Question 3: Quelle propriété distingue une distance ultramétrique d'une distance classique ?
Réponse: L'inégalité ultratriangulaire remplace l'inégalité triangulaire.
Question 4: Pourquoi une distance ultramétrique est-elle utile en classification hiérarchique ?
Réponse: Elle impose une structure stricte compatible avec les regroupements en arbre.

—

1.3 Conclusion

Ce chapitre met en évidence les différences entre une dissimilarité, une distance et une distance ultramétrique. Chaque concept s'appuie sur des propriétés spécifiques qui reflètent des contraintes géométriques et logiques. La distance ultramétrique, avec sa structure rigide, est particulièrement adaptée aux algorithmes de classification hiérarchique. —

2 Cas des variables quantitatives

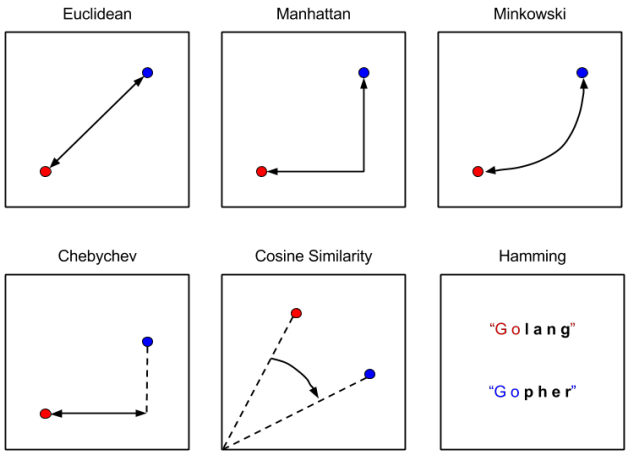
2.1 Introduction

La distance est une fonction qui respecte des propriétés mathématiques formelles, comme la non-négativité, la symétrie et la propriété du triangle. Des exemples de distances incluent:

- **Distance Euclidienne**: Utilisée pour mesurer la distance directe entre deux points dans un espace multidimensionnel.
- **Distance Ultramétrique**: Utilisée dans des méthodes spécifiques comme la Classification Ascendante Hiérarchique (CAH), où elle respecte des propriétés plus strictes dans l'agrégation des clusters.

La dissimilarité est une mesure plus générale de la différence entre deux objets ou individus. Elle ne respecte pas nécessairement toutes les propriétés formelles d'une distance, et peut être plus flexible dans sa définition et son utilisation.

2.2 Les distances



2.2.1 Distance de Minkowski

La distance de Minkowski est une généralisation des distances classiques utilisées pour mesurer la différence entre deux individus dans un espace multidimensionnel.
Formule:

$$d(i_1, i_2) = \left[\sum_{j=1}^{p} |x_{i_1 j} - x_{i_2 j}|^q \right]^{\frac{1}{q}}$$

Lecture: La distance entre les individus i_1 et i_2 est la racine d'ordre q de la somme des différences absolues élevées à la puissance q sur p dimensions (ou variables).
Interprétation: Cette formule permet d'ajuster la mesure de distance avec le paramètre q. Pour différents choix de q, on obtient des distances spécifiques comme Manhattan ($q = 1$) ou Euclidienne ($q = 2$).

—

2.2.2 Distance de Manhattan

La distance de Manhattan est un cas particulier de Minkowski avec $q = 1$. Elle mesure la somme des différences absolues entre les coordonnées.
Formule:

$$d(i_1, i_2) = \sum_{j=1}^{p} |x_{i_1 j} - x_{i_2 j}|$$

Lecture: La distance entre i_1 et i_2 est la somme des écarts absolus sur toutes les dimensions (p).
Interprétation: C'est comme mesurer la distance dans une ville où l'on doit se déplacer uniquement le long des rues en ligne droite (quadrillage).

Exemple: Soient deux points $(1, 2)$ et $(4, 6)$.

$$d = |1 - 4| + |2 - 6| = 3 + 4 = 7$$

—

2.2.3 Distance de Chebychev

La distance de Chebychev mesure la plus grande différence entre les coordonnées des individus.
Formule:

$$d(i_1, i_2) = \max_{j \in \{1,\dots,p\}} |x_{i_1 j} - x_{i_2 j}|$$

Lecture: La distance entre i_1 et i_2 est déterminée par la plus grande différence absolue parmi toutes les dimensions (p).
Interprétation: Elle reflète le mouvement sur une grille où l'on peut se déplacer dans toutes les directions, mais où la distance est contrôlée par l'étape la plus longue.
Exemple: Soient deux points $(1, 2)$ et $(4, 6)$.

$$d = \max\{|1 - 4|, |2 - 6|\} = \max\{3, 4\} = 4$$

—

2.2.4 Distance Euclidienne

La distance euclidienne est un cas particulier de la distance de Minkowski avec $q = 2$, et c'est la mesure la plus couramment utilisée dans de nombreux algorithmes de clustering et d'optimisation.
Formule:

$$d(i_1, i_2) = \left[\sum_{j=1}^{p} (x_{i_1 j} - x_{i_2 j})^2 \right]^{\frac{1}{2}}$$

Lecture: La distance entre les individus i_1 et i_2 est la racine carrée de la somme des carrés des différences entre leurs coordonnées sur chaque dimension (p).
Interprétation: C'est la distance "à vol d'oiseau" entre deux points dans un espace euclidien, qui est la plus naturelle dans un espace cartésien. Elle est couramment utilisée lorsque les dimensions sont indépendantes et lorsque les données sont bien distribuées.
Exemple: Soient deux points $(1, 2)$ et $(4, 6)$.

$$d = \sqrt{(1 - 4)^2 + (2 - 6)^2} = \sqrt{9 + 16} = \sqrt{25} = 5$$

Spécificité de la distance Euclidienne: La distance euclidienne est la distance la plus utilisée en pratique, en particulier dans les algorithmes de clustering (comme le K-means) et d'optimisation. Elle est considérée comme la norme "naturelle" pour mesurer la similarité entre des points dans un espace multidimensionnel et est souvent le point de départ pour la compréhension des autres types de distances.

—

2.3 Comparaison des distances

Distance	**Paramètre q**	**Formule**	**Utilisation principale**
Minkowski	$q > 0$	$\left[\sum_{j=1}^{p} \lvert x_{i_1 j} - x_{i_2 j}\rvert^q\right]^{1/q}$	Générale, ajuste q pour divers besoins.
Euclidienne	$q = 2$	$\left[\sum_{j=1}^{p} (x_{i_1 j} - x_{i_2 j})^2\right]^{1/2}$	Mesure classique dans un espace cartésien.
Manhattan (Minkowski, $q = 1$)	$q = 1$	$\sum_{j=1}^{p} \lvert x_{i_1 j} - x_{i_2 j}\rvert$	Déplacements sur grille, données discrètes.
Chebychev	$q \to \infty$	$\max_j \lvert x_{i_1 j} - x_{i_2 j}\rvert$	Prend l'écart maximal.

—

2.4 Problématique du choix de la Distance

Il n'y a pas de **bonne ou mauvaise réponse** pour le choix de la distance entre des options comme **Manhattan**, **Euclidienne** ou **Chebyshev**.
Le choix dépend du **contexte métier** et des objectifs spécifiques:

- **Distance de Chebyshev** : Idéale si l'on veut pénaliser les écarts importants sur une seule variable.
- **Distance de Manhattan** : Utile pour minimiser l'impact des écarts moyens sur les variables.
- **Distance Euclidienne** : La plus couramment utilisée par défaut, mais pas nécessairement la meilleure dans tous les cas.

En résumé, le choix de la distance doit être fait en fonction des objectifs spécifiques de l'analyse. Si aucune préférence particulière n'émerge, la **distance Euclidienne** est généralement adoptée par défaut.

—

2.5 Questions pour réviser

Question 1: Quelle distance est obtenue avec Minkowski pour $q = 1$?
Réponse: La distance de Manhattan.
Question 2: Quelle distance mesure uniquement l'écart maximal entre deux individus ?
Réponse: La distance de Chebychev.
Question 3: Quelle est la différence entre Minkowski avec $q = 2$ et Manhattan ?
Réponse: Minkowski avec $q = 2$ est la distance euclidienne, qui mesure la distance en ligne droite, tandis que Manhattan mesure la distance en suivant un quadrillage.
Question 4: Pourquoi la distance de Chebychev est-elle utile dans certains contextes ?
Réponse: Elle est adaptée lorsque l'on veut contrôler une distance maximale sur toutes les dimensions, comme dans des problèmes de tolérance ou d'alignement strict.

—

2.6 Conclusion

Ces différentes mesures de distance offrent des approches flexibles pour évaluer la similarité ou la dissimilarité entre individus en fonction du contexte d'analyse. Minkowski est la formule la plus générale, permettant de dériver Manhattan et Chebychev comme cas particuliers adaptés à des situations spécifiques.

—

3 Pour les Variables Qualitatives

Au lieu des distances, pour les variables qualitatives, on utilise des mesures de similarité ou dissimilarité qui évaluent si les catégories sont identiques ou non.
Exemples :

- **Dissimilarité binaire** : Pour deux individus avec une variable qualitative (comme la couleur des yeux), on pourrait dire qu'ils sont identiques si leurs valeurs sont égales, sinon ils sont différents.
- **Indice de Jaccard** : Utilisé pour comparer des ensembles ou des catégories, souvent pour des données binaires ou catégorielles.

Les distances fonctionnent bien pour les variables quantitatives car elles sont numériques et mesurent des écarts mesurables.
Pour les variables qualitatives, les distances classiques ne sont pas appropriées. Au lieu de cela, on utilise des mesures adaptées aux données catégorielles, comme des indices de similarité ou des matrices de dissimilarité.