Similarités, Distances et Proximité

Gilles HUNAULT

Angers, 1997

Contenu

1. Introduction

La notion de *distance* est une notion "transversale" puisqu'elle peut s'appliquer à n'importe quel "objet" : les distances de la vie quotidienne (comme la distance entre deux villes) sont là pour le montrer. En science, on trouve des distances partout : entre éléments d'un même ensemble (non forcément numérique), entre ensembles, entre fonctions, entre suites, entre classes ou centres de gravité de classes, voire même entre distances elles-mêmes.

Les distances jouent un rôle important en mathématiques car elles permettent de construire certaines topologies via la notion de limite, continuité. En biologie, c'est en particulier pour les classifications, les typologies est les comparaisons et les alignements que les distances sont utiles. En informatique théorique, elles permettent aussi de comparer des éléments, de les classer.

Nous commencerons par rappeler les définitions classiques des métriques et par donner des exemples simples et généraux avant de passer aux distances sur tableaux en 0/1 et sur tableaux de comptages et aux tableaux de variables quantitatives. Ensuite, des modèles généraux de construction de distances seront fournis avec, comme suite logique, la présentation des liens entre distances et relations binaires.

On notera que si au départ on dispose d'indices ou de coefficients numériques d pour quantifier les distances, on peut se ramener très facilement à un cas booléen (0 ou 1) en posant $d_{bool}(x,y) = 0$ si $d(x,y) = 0$ et 1 ailleurs.

2. Définitions

Soit E un ensemble quelconque, c'est à dire non forcément numérique. Soit f une fonction de $E \times E$ dans $\mathbb{R}^+$. f est donc positive ou nulle. Pour tous x, y, z et t dans E, on définit les termes suivants

f est	*d-nulle*	ssi	$f(x,x) = 0$
f est	*séparée*	ssi	$f(x,y) = 0 \Rightarrow y = x$
f est	*symétrique*	ssi	$f(x,y) = f(y,x)$
f est	*équidiagonale*	ssi	$f(x,x) = f(y,y)$
f est	*sous-diagonale*	ssi	$f(x,y) \leq f(x,x) = f(y,y)$
f est	*triangulaire*	ssi	$f(x,y) \leq f(x,t) + f(t,y)$
f est	*sous-maximale*	ssi	$f(x,y) \leq max(\ \{f(x,z), f(y,z)\}\ )$

Une **similarité**	est une fonction symétrique et sous-diagonale.
Une **dissimilarité**	est une fonction symétrique et d-nulle.
Une **pseudo-métrique**	est un indice de dissimilarité triangulaire.
Une **métrique**	est une pseudo-métrique séparée.
Une **ultramétrique**	est une pseudo-métrique sous-maximale.

On utilise classiquement les termes **écart** et **indice d'écart** comme synonymes de pseudo-métrique, le mot **distance** comme synonyme de métrique. Dans certains ouvrages, on trouve les termes "indice de (dis)similarité" ou "coefficient de (dis)similarité" plutôt que "(dis)similarité".

Il est à noter qu'une ultramétrique est une distance particulière.

3. Exemples

3.1 Exemples généraux

La fonction nulle ν est définie par $\nu(x,x) = 0$ pour tout x. Elle est d-nulle, symétrique, sous-diagonale, triangulaire et sous-maximale. Elle n'est pas séparée.

L'application égalitaire est la fonction δ telle que $\delta(x,x) = 0$ et $\delta(x,y) = 1$ pour $y \neq x$. Elle est d-nulle, séparéee, symétrique, triangulaire. Elle n'est pas sous-diagonale.

Soit E un ensemble numérique. L'application $(x,y) \to |x-y|$ est une distance nommée *distance usuelle* sur E.

L'application $(x,y) \to \left|\frac{1}{x} - \frac{1}{y}\right|$ est une distance sur $\mathbb{R}^*$.

L'application $(x,y) \to |arctangente(x-y)|$ est une distance sur $\mathbb{R}^*$.

Si f et g sont deux fonctions intégrables,

$$(f,g) \to \int_0^1 \ |f-g| \text{ est une distance.}$$

$$(f,g) \to \int_0^1 \ (|f-g| + |f'-g'|) \text{ est une distance.}$$

$$(f,g) \to \sup_{[0,1]} |f-g| \text{ est une distance.}$$

Citons aussi les distances p-adiques qui utilisent la décomposition en base p, les distances d'éditions entre chaines de caractères, les distances entre prédicats, les distances entre suites numériques...

3.2 Cas où les éléments de E sont des vecteurs de 0/1

Ce cas est très important en biologie, en informatique, en statistiques car il correspond aux descriptions d'états en présence/absence, aux tables de vérité en vrai/faux, aux événements en réussite/échec... De nombreuses fonctions portent alors le nom d'indice. Afin de les comparer entre eux et pour de nombreux couples, certains indices sont normalisés à 1, ce qui signifie que $0 \leq f(x, y) \leq 1$ pour tous x et y.

Soit I un ensemble fini de valeurs de cardinal n. Si $x = (x_i)_{i \in I}$ et $y = (y_i)_{i \in I}$ sont deux vecteurs de 0 et de 1, on définit classiquement les nombres et indices suivants où toutes les sommes sont prises pour i de 1 à n (uu est souvent jugé comme "déterminant").

uu	le nombre de fois où	$x = 1$ et $y = 1$	soit $\sum x_i y_i$
zz	le nombre de fois où	$x = 0$ et $y = 0$	soit $\sum (1 - x_i)(1 - y_i)$
uz	le nombre de fois où	$x = 1$ et $y = 0$	soit $\sum x_i (1 - y_i)$
zu	le nombre de fois où	$x = 0$ et $y = 1$	soit $\sum (1 - x_i) y_i$
ux	le nombre de fois où	$x = 1$	soit $uu + uz = \sum x_i$
uy	le nombre de fois où	$y = 1$	soit $uu + zu = \sum y_i$
zx	le nombre de fois où	$x = 0$	soit $zu + zz = \sum (1 - x_i)$
zy	le nombre de fois où	$y = 0$	soit $uz + zz = \sum (1 - y_i)$
n	le nombre	$ux + zx = uy + zy$	$= uu + zz + uz + zu.$

Notation	Nom	Année	Formule
I_1	Jaccard	1908	$uu/(uu + uz + zu)$
I_2	Sokal-Michener	1958	$(uu + zz)/p$
I_3	Sokal-Sneath I		$I_2/(2 - I_2)$
I_4	Czekanowski	1913	$2I_1/(1 + I_1)$
I_5	Russel-Rao	1940	uu/p
I_6	Kulczynski	1927	$uu(1/ux + 1/uy)/2$
I_7	Sokal-Sneath II	1963	$I_1/(2 - I_1)$
I_8	Ochiai	1957	$\sqrt{I_4 I_6}$

L'indice I_2 de Sokal-Michener est aussi nommé indice de Rogers-Tanimoto (1960).
L'indice I_4 de Czekanowski est aussi nommé indice de Dice (1945) ou encore indice de Sorensen (1948).

3.3 Cas où les éléments de E sont des comptages

Soit à nouveau I un ensemble fini de valeurs de cardinal n et soient maintenant $x = (x_i)_{i \in I}$ et $y = (y_i)_{i \in I}$ sont deux vecteurs d'entiers positifs ou nuls. On note d_i la différence $x_i - y_i$, s_i la somme $x_i + y_i$, r l'inverse de n, p_i le minimum de x_i et y_i et g_i le maximum de x_i et y_i. On utilise classiquement les indices suivants (toutes les sommes sont prises pour i de 1 à n).

Notation	Nom	Année	Formule
I_9	Czekanowski	1932	$r \sum \lvert di \rvert$
I_{10}	Euclidienne moyenne		$\sqrt{r \sum di^2}$
I_{11}	Lance-Williams	1966	$r \sum(\lvert di \rvert / si)$
I_{12}	Clark	1952	$\sqrt{r \sum(\lvert di \rvert / si)^2}$
I_{13}	Odum	1950	$(\sum \lvert di \rvert)/(\sum si)$
I_{14}	Kulczynski		$(\sum m_i)\ /\ (\sum s_i)$
I_{15}	Marczewski-Steinhaus		$(\sum[d_i \rvert)\ /\ (\sum g_i)$

I_9 est aussi nommé indice de Haltenorth (1937), indice de Cain-Harrison (1958).

I_{13} est aussi nommé indice de Bray-Curtis (1957),

On remarquera que $I_{14} = (1 - I_{13})/2$ et que $I_{15} = 2I_{13}/(1 + I_{13})$.

3.4 Cas des tableaux de variables quantitatives

Il y a historiquement de nombreuses formules de distances (entre lignes) et de coefficients de liaison ou "corrélations" (entre colonnes). Nous rappelons ici les plus importantes. L'application de formules en général prévues pour les colonnes du tableau des données (les "variables" de la statistique classique) aux lignes du même tableau peut alors n'avoir aucun sens. Les distances traditionnelles entre lignes sont nommées

d_e	Distance euclidienne usuelle ("plate")
d_{ep}	Distance euclidienne pondérée
d_χ	Distance du χ^2
d_m	Distance de l'écart moyen
d_M	Distance de l'écart maximum

et les coefficients entre colonnes (sauf ρ_D)

ρ	Coefficient de corrélation linéaire usuel [de Pearson(-Bravais)]
ρ_K	Coefficient de corrélation des rangs de Spearman
ρ_S	Coefficient de corrélation des rangs de Kendall
ρ_D	Coefficient de corrélation des divergences [de Clark]

Rappelons maintenant les formules. Soit $X[i,j]$ un tableau de données où les n individus i sont en ligne et les p variables j sont en colonne. On note :

m_j	la moyenne de la colonne j,	soit $m_j = \Sigma_i\ X[i,j]/n$,
σ_j	l'écart-type de la colone j	soit $\sigma_j^2 = \Sigma_i\ (X[i,j]_m(j))^2/n$,
$s_{i.}$	la somme de la ligne i	soit $s_{i.} = \Sigma_j\ X[i,j]$,
$s_{.j}$	la somme de la colonne j	soit $s_{.j} = \Sigma_i\ X[i,j]$,
$r_j(a,b)$	le sens de j entre a et b,	
$r(i,j)$	le rang de $X[i,j]$ pour j.	

On rappelle que

$$\begin{aligned} r_j(a,b) &= 1 \text{ si } X[a,j] > X[b,j],\\ r_j(a,b) &= 0 \text{ si } X[a,j] = X[b,j],\\ r_j(a,b) &= \text{-1 si } X[a,j] < X[b,j]. \end{aligned}$$

Si a et b sont deux indices de lignes du tableau

$$
\begin{aligned}
d_e(a,b)^2 &= \textstyle\sum_j \; (X[a,j]-X[b,j])^2 \\
d_{ep}(a,b)^2 &= \textstyle\sum_j \; w_j\,(X[a,j]-X[b,j])^2 \text{ (souvent } w_j = 1/\sigma_j^2) \\
d_\chi(a,b)^2 &= \textstyle\sum_j \; (X[a,j]/s_{a.}-X[b,j]/s_{b.})^2\,/s_{.j} \\
d_m(a,b) &= \textstyle\sum_j \; |X[a,j]-X[b,j]|\,/p \\
d_M(a,b) &= Max_j \; \{\; X[a,j]-X[b,j] \;\} \\
\rho_D(a,b) &= (1/p) \; \textstyle\sum_j \; \dfrac{X[a,j]-X[b,j]}{X[a,j]+X[b,j]}
\end{aligned}
$$

Si c et d sont deux indices de colonnes du tableau

$$
\begin{aligned}
\rho(c,d) &= \textstyle\sum_i \; (\,X[i,c]-m_c\,)\,(\,X[i,d]-m_d\,)\,/\sigma_c\sigma_d \\
\rho_S(c,d) &= 1-6\textstyle\sum_i (\,r(i,c)-r(i,d)\,)^2\,/n(n^2-1) \\
\rho_K(c,d) &= 2\sum_{a<b} r_c(a,b)r_d(a,b)/n(n-1)
\end{aligned}
$$

4. Modèles de fonctions de distances sur E

Si s est une similarité alors d telle que $d(x,y) = s_M - s(x,y)$ est une dissimilarité où $s_M = s(x,x) = s(y,y)$ désigne la valeurs de s sur la diagonale.

4.1 Cas où E est sans dissimilarité

Soit E un ensemble fini, et F l'ensemble des applications de E dans R. Soit f de E dans F. Soit p une fonction de poids sur E : p(x) noté p_x est positif ou nul et la somme des p_x vaut 1. Alors pour tout q réel supérieur ou égal à 1, l'application d définie par

$$d(A,B) = \sqrt[q]{\sum_{x \in E} p_x |f_A(x) - f_B(x)|^q}$$

est une distance, nommé distance de $Minkowski$ discrète à l'ordre q relative au poids p.

Cas particulier : q=1, $f(A) = \chi_A$, fonction caractéristique de A. Cette distance est nommée distance de la différence symétrique.

Autre cas particulier : q=2, $f(x) = 1/\sqrt{p(A)}$ si $x \in A$. Cette distance est nommée distance d'Ochiai.

Autre cas particulier : distance du χ^2 entre distributions de probabilité.

4.2 Cas où E est avec dissimilarité

Les distances associées sont alors plus compliquées que les distances précédentes. Nous nous contentons de citer deux distances : la distance de *Hausdorff* et la distance du maximum. On trouvera sur le web avec Google des articles présentant ces distances.

4.3 Distances sur $\mathbb{R}^n$

4.4 Distances induites

Si A est muni d'une distance d et si f est une fonction de I dans A alors l'application f_d telle que $f_d(x, y) = d(\ f(x), f(y))$ transporte les propriétés de d. Si f est injective, f_d est aussi une distance. Ainsi $(x, y) \to |x^2 - y^2|$ n'est pas une distance mais $(x, y) \to |x^3 - y^3|$ en est une. Si d est une distance $(x, y) \to min(\ \{1, d(x, y)\}\ )$ et $(x, y) \to \dfrac{d(x, y)}{1 + d(x, y)}$ sont aussi des distances. Si les d_i sont les distances, si les α_i sont des réels psositifs ou nuls, alors $m = max(d_i)$, $c = \sum \alpha_i d_i$ et $r = \sqrt{\sum \alpha_i d_i^2}$ sont aussi des distances.

5. Distances et relations binaires

5.1 Distance induite par une relation binaire

Soit R une relation binaire totale sur E et δ_{R} la fonction définie sur $E \times E$ par $\delta_{\mathsf{R}}(x, y) = 0$ si $\mathsf{R}(x, y)$ est vrai et $\delta_{\mathsf{R}}(x, y) = 1$ si $\mathsf{R}(x, y)$ est faux. Si R est réflexive alors à δ_{R} est équidiagonale (donc d-nulle). R est symétrique équivaut à δ_{R} est symétrique. De plus si R est transitive alors δ_{R} est sous-maximale donc triangulaire. Une relation d'équivalence RE induit donc une distance δ_{RE} nommée *distance discrète induite par* RE.

5.2 Relation induite par une fonction numérique

Si f est une fonction définie sur $E \times E$ à valeurs dans R et ε un réel positif ou nul, on dit que x et y sont ε-proches au sens f ssi $|d(x, y)| \leq \varepsilon$. Soit $\mathsf{R}_{f,\varepsilon}$ la relation binaire définie par $\mathsf{R}_{f,\varepsilon}(x, y)$ est vraie ssi x et y sont ε-proches au sens de f et faux sinon.

Si f est une distance d, $\mathsf{R}_{f,\varepsilon}$ est alors réflexive, symétrique et transitive, c'est la relation d'équivalence nommée *relation de proximité* induite par d et pour toute fonction f, $\mathsf{R}_{f,0}$ est la relation classique d'égalité.