Dissimilarités entre jeux de données
-
Preuves

15 février 2017

1 Preuve de la proposition 4

Proposition *Soit $E_1...E_n$ une suite d'ensembles finis et A leur produit cartésien $\prod_{i=1}^{n} E_i$. Soit $d_1...d_n$ une suite de fonctions de dissimilarité respectivement normalisées sur $E_1...E_n$. Alors, la* ***dissimilarité normalisée par la borne supérieure*** *sur A selon $d_1...d_n$ est une* ***fonction de dissimilarité normalisée*** *sur A.*

Preuve *Soit $\delta \in \mathbb{R}^*$. Les E_i étant des ensembles finis, et les d_i étant des dissimilarités normalisées, $\max\limits_{(x,y)\in A^2} d_i(x_i, y_i)$ existe $\forall i$. Soit alors $(X, Y, Z) \in A^3$ tels que*

$$\exists i \in [1..n], d_i(X_i, Y_i) = d_i(X_i, Z_i) + \delta * \max_{(x,y)\in A^2}(d_i(x_i, y_i))$$
$$\forall j \in [1..n],\ j \neq i,\ d_j(X_j, Y_j) = d_j(X_j, Z_j)$$

Ce qui implique $d_A^{ubr}(X, Y) = d_A^{ubr}(X, Z) + \delta$. Ainsi, $\exists \Delta = \delta$, et d_A^{ubr} est bien une ***fonction de dissimilarité normalisée*** *sur A.* ■

2 Preuve de la proposition 7

Proposition *Avec σ une fonction de mapping optimale, $d^{\sigma}_{F(\omega)}$ est une fonction de dissimilarité normalisée sur $F(\omega)$.*

Preuve *On démontre successivement quatre propriétés : Positivité, Indiscernabilité des identiques, Symétrie et Normalisation (au sens de la Définition 2).*

1. *Soient $x, x' \in \omega$ et σ une fonction de mapping optimale.*
   Montrons que $d^{\sigma}_{F(\omega)}(x, x') \geq 0$.
   $\forall f \in F$, δ_f est une fonction de dissimilarité, donc
   $$\forall i, j \in [1, n] * [1, n'],\ \delta_f(x_i, x'_j) \geq 0$$
   $$\text{et par somme, } d^{\sigma}_{F(\omega)}(x, x') \geq 0$$
   □

2. *Soit $x \in \omega$ ayant n attributs et σ une fonction de mapping optimale.*
   Montrons $d^{\sigma}_{F(\omega)}(x, x) = 0$.

$$d^{\sigma}_{F(\omega)}(x,x) = \frac{1}{n}(\sum_{\sigma_1(i)=j}^{i,j} \delta_F(x_i,x_j) + \sum_{\sigma_1(i)=\emptyset}^{i} \delta_F(x_i,\emptyset) + \sum_{\sigma_2(j)=\emptyset}^{j} \delta_F(\emptyset,x_j))$$

$\forall f \in F$, δ_f est une fonction de dissimilarité, donc

$$\forall i,j \in [1,n],\ \delta_f(x_i,x_j) \geq 0$$

$$\textit{et par somme, } d^{\sigma}_{F(\omega)}(x,x') \geq 0$$

De plus, comme δ_f est une fonction de dissimilarité,

$$\forall i \in [1,n],\ \delta_f(x_i,x_i) = 0$$

Donc, pour le mapping (Id,Id), avec Id la fonction Identité, on a :

$$d^{(Id,Id)}_{F(\omega)}(x,x) = \frac{1}{n}(\sum_{i\in[1,n]} \delta_F(x_i,x_i)) = 0$$

Or, σ est optimale, donc :

$$d^{\sigma}_{F(\omega)}(x,x) = \min_{\sigma'\in Mappings(x,x)} d^{\sigma'}_{F(\omega)}(x,x)$$

$$0 \leq d^{\sigma}_{F(\omega)}(x,x) \leq d^{(Id,Id)}_{F(\omega)}(x,x)$$

$$0 \leq d^{\sigma}_{F(\omega)}(x,x) \leq 0$$

$$d^{\sigma}_{F(\omega)}(x,x) = 0$$

□

3. *Soient $x,x' \in \omega$ possédant respectivement n et n' attributs et σ une fonction de mapping optimale. Montrons que $d^{\sigma}_{F(\omega)}(x,x') = d^{\sigma}_{F(\omega)}(x',x)$.*

   Notons $\sigma(x,x') = (\sigma_1,\sigma_2)$ et $\sigma(x',x) = (\sigma'_1,\sigma'_2)$
   Moyennant substitution, supposons que $d^{\sigma}_{F(\omega)}(x,x') \leq d^{\sigma}_{F(\omega)}(x',x)$

   $\forall f \in F$, δ_f est une fonction de dissimilarité, donc

$$\forall i,j \in [1,n]*[1,n'],\quad \begin{cases} \delta_f(x_i,x'_j)=\delta_f(x'_j,x_i) \\ \delta_f(x_i,\emptyset) = \delta_f(\emptyset,x_i) \\ \delta_f(\emptyset,x'_j) = \delta_f(x'_j,\emptyset) \end{cases}$$

   Il existe donc un mapping $(\sigma'_1,\sigma'_2) = (\sigma_2,\sigma_1)$ tel que

$$\frac{1}{max(n,n')}(\sum_{\sigma_1(i)=j}^{i,j} \delta_F(x_i,x'_j) + \sum_{\sigma_1(i)=\emptyset}^{i} \delta_F(x_i,\emptyset) + \sum_{\sigma_2(j)=\emptyset}^{j} \delta_F(\emptyset,x'_j)) \qquad (A)$$

$$=$$

$$\frac{1}{max(n,n')}(\sum_{\sigma_2(j)=i}^{i,j} \delta_F(x'_j,x_i) + \sum_{\sigma_2(i)=\emptyset}^{i} \delta_F(\emptyset,x_i) + \sum_{\sigma_1(j)=\emptyset}^{j} \delta_F(x'_j,\emptyset)) \qquad (B)$$

   Or, σ est une fonction de mapping optimale, donc

$$d^{\sigma}_{F(\omega)}(x',x) \leq A$$

   Et comme $A = B = d^{\sigma}_{F(\omega)}(x,x')$

$$d^{\sigma}_{F(\omega)}(x,x') \leq d^{\sigma}_{F(\omega)}(x',x) \leq d^{\sigma}_{F(\omega)}(x,x')$$

$$d^{\sigma}_{F(\omega)}(x,x') = d^{\sigma}_{F(\omega)}(x',x)$$

□

4. *Soit $x \in \omega$. L'ensemble $F(x_i)$ des valeurs de méta-attributs décrivant le feature i de x est alors une description atomique de x_i. De plus, ω est fini, ce qui entraine $F(\omega)$ fini et donc $d^{\sigma}_{F(\omega)}$ bornée sur $F(\omega)$ atomique. Selon la Définition 2.1, $d^{\sigma}_{F(\omega)}$ est donc bien normalisée sur $F(\omega)$.* $\square$

$d^{\sigma}_{F(\omega)}$ est donc bien une fonction de dissimilarité normalisée sur $F(\omega)$. $\blacksquare$

3 Preuve de la proposition 10

Proposition *δ_{KS} est une fonction de dissimilarité normalisée.*

Preuve *On doit montrer quatre propriétés : Positivité, Indiscernabilité des identiques, Symétrie et Normalisation (au sens de la Définition 2). La positivité, l'indiscernabilité des identiques, et la symétrie sont assurées par les propriétés de la statistique de Kolmogorov-Smirnov, qui, pour les échantillons x_i et y_j, est la borne supérieure de la différence absolue entre les fonctions de répartitions empirique de x_i et y_j.*

— *Une différence absolue est positive, donc $\delta_{KS}(x_i, y_j) \geq 0$.*
— *La différence absolue entre deux fonctions de répartitions identiques est toujours nulle, donc $\delta_{KS}(x_i, x_i) = 0$.*
— *Une différence absolue est symétrique, donc $\delta_{KS}(x_i, y_j) = \delta_{KS}(y_j, x_i)$.*

La distribution d'un attribut est un bon exemple d'ensemble atomique (on ne peut le diviser sans perte d'information). De plus, ω étant fini, δ_{KS} est nécessairement bornée sur ω. δ_{KS} est donc bien normalisée au sens de la Définition 2. $\blacksquare$