TD d'analyse de données. 3

1. Pour un tableau N de contingence comportant m_1 lignes et m_2 colonnes, formé de valeurs entières : $\mathrm{N} = [n^{(1)}\ n^{(2)}\ n^{(m_2)}] = {}^t[n_1\ n_2\ \ n_{m1}]$, on note $n_{i\,\circ}$ la somme des éléments de la i-ème ligne n_i, et $n_{\circ\,j}$ la somme des éléments de la j-ème colonne $n^{(j)}$ du tableau. On note n la somme de tous les éléments de N, et D_1 désigne la matrice diagonale d'éléments $n_{1\,\circ}, n_{2\,\circ}, \ldots, n_{m1\,\circ}$.
Indiquez la forme du profil $D_1^{-1}\,\mathrm{N}$, et du centre de gravité $g = \frac{1}{n}\,{}^t(D_1^{-1}\,\mathrm{N})\,D_1\,{}^t1_{m1}$
On définit la distance d du Khi deux entre deux lignes i et i' de N par :

$$d^2(i,i') = \sum_{j=1}^{m_2} \frac{n}{n_{\circ j}} \left(\frac{n_{ij}}{n_{i\,\circ}} - \frac{n_{i'j}}{n_{i'\,\circ}}\right)^2$$

i) Vérifiez que d est bien une distance sur N^{m_2}

ii) Vérifiez que dans l'écriture de d^2 on peut remplacer les valeurs n_c par les fréquences $f_c = \frac{n_c}{n}$

iii) Qu'obtient-on si on échange les rôles des lignes et des colonnes de N ?

2. Si d est une application de $E \times E$ dans R qui vérifie $d(x, y) = d(y, x) \geq 0$, $\forall (x, y) \in E^2$, d est appelée *dissimilarité* si $\forall x \in E$, $d(x, x) = 0$, et *similarité* si $\forall (x, y) \in E^2$, $d(x, x) \geq d(x, y)$. Une similarité peut-elle être une distance ? Donnez des exemples de similarités et de dissimilarités (autres que des distances).

3. Indiquez la forme du produit (est-ce un produit scalaire ?) en dimension 2, associé à la matrice $M = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Quelle est la condition pour que la 'distance' entre deux points soit nulle ?

4. Comment peut s'écrire l'inertie d'un nuage de M points ? Comment définir une inertie intra classe et une inertie inter classes ? Quelles sont les relations qui relient ces valeurs ?

5. On souhaite comparer le fonctionnement d'une classification hiérarchique et celui d'une méthode *k*-means. Observez le comportement de ces deux algorithmes en dimension 1 – on pourra prendre des valeurs entières, comme les n premiers entiers, ou leurs carrés. Considérez des exemples en dimension 2, en proposant des critères d'agrégation des classes, pour la distance euclidienne dans le plan.

6. Soit un nuage de 10 points : (0, 0); (2,0); (4, 0); (8, 0); (0, 1); (0, 6); (8, 2); (3, 3); (6, 5); (2, 8)
En utilisant la distance euclidienne du plan, écrivez une Classification Ascendante Hiérarchique jusqu'à obtenir 3 classes exactement. Ecrivez pour l'ensemble des étapes (agrégation de classes) la valeur de l'inertie intra classe totale. Comparez avec une approche par moyennes mobiles pour 3 classes, en prenant des centres initiaux dans des classes distinctes de la CAH.

7. Soit $A = (a_{ij})$ une matrice d'appartenance, ie une matrice binaire, pour laquelle $a_{ij} = 1$ signifie que le i-ème individu (ligne i) appartient à la classe no j. Comment peuvent être écrits les effectifs des différentes classes ? Quelle est la condition pour qu'un individu appartienne à au plus une classe (respectivement exactement une classe) ? Peut-on représenter cette situation par un graphe ? Peut-on utiliser une matrice d'appartenance non déterministe, dont les éléments sont des nombres compris entre 0 et 1 ?

Analyse de données : TD3

Exo 2

$d : E \times E \to \mathbb{R}^+$

i) $d(x,y) = d(y,x)$

ii) $d(x,y) = 0 \Rightarrow x = y$ → dissim

ii') $d(x,y) \leq d(x,x)$ → sim

ii'') $d(x,x) = 0$ → dissim

iii) $d(x,z) \leq d(x,y) + d(y,z)$ → nécessaire pour une distance

Si on prend à la fois sim et dissim on a que $d(x,x) \geq 0$ et $d(x,x)$ majoré → fction nulle

distance = cas particulier de dissim

Exemple de sim

- $\begin{cases} d(x,x) = 1 \\ d(x,y) = 0 \text{ si } x \neq y \end{cases}$
- $d(x,y) = \min(|x|, |y|)$

- $C_{x,y} = \dfrac{cov(X,Y)}{\sqrt{V(X)V(Y)}}$, $cov(X,Y) = E([X - E(X)] \cdot [Y - E(Y)])$

  ↳ $-1 \leq C_{x,y} \leq 1$ → $d(x,y) = |C_{x,y}|$

  $C_{x,x} = 1$ car $cov(x,x) = V(x)$

Exemple dissim

- $d(x,y) = (x-y)^2$ → inégalité plus vraie

- Faire une composition avec une distance ne donne pas une distance mais une similarité en général

Exo 3

produit scalaire associé à M :
$\langle u, u' \rangle = u^t M u'$

$M = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$, $u = \begin{bmatrix} x \\ y \end{bmatrix}$, $u' = \begin{bmatrix} x' \\ y' \end{bmatrix}$

$\langle u, u' \rangle = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} x' \\ y' \end{bmatrix} = xx' + xy + yx' + yy'$

$\langle u, u \rangle = x^2 + 2xy + y^2$
$= (x+y)^2$

$\langle u, u \rangle = 0 \iff x+y=0$
$\iff x = -y \not\iff x = y = 0$
Ce n'est donc pas un produit scalaire (car $\langle u, u \rangle = 0 \not\Rightarrow u = 0$ en général)

Donc $\|u\| = \sqrt{\langle u, u \rangle}$ ne définit pas une norme.
↳ et donc pas réellement une distance

$d(u, u') = \|u - u'\| = 0$
si $\langle u - u', u - u' \rangle = 0$ $\quad u - u' = \begin{bmatrix} x - x' \\ y - y' \end{bmatrix}$

$(x - x') + (y - y') = 0 \iff x + y = x' + y'$

vérifie que $d(x,x) = 0$ et $d(x,y) = d(y,x)$ → ok ✓
↳ dissim

Et les inégalités triangulaires ? Apparemment ça vérifie

Exo 5 : Classifications hiérarchique

Pour une classif hiérarchique :

$d(x,y)$ ← * définition d'une fonction d'écart (dissimilarité en général)

$\delta(c,c')$ ← * définition d'une fonction d'écart entre classes

* Pour n points on définit n classes $C_k = \{x_k\}$, $k = n$

nbre de classe final ← , ou un nbre $K > 1$ pour être plus pertinent

* while $(k > 1)$ on fusionne les classes les + proches ($\delta(c,c')$ minimal), donc $k--$

Dans le cas où on représente des barycentres

$n = 5$ $\quad E = \{x_1, x_2, x_3, x_4, x_5\}$

$C_3 = \{x_3, x_4\} \rightarrow 00110$

↳ Au début la matrice C des classes représentées par un barycentre est I_n (quand on part n classes de 1 pt au début) puis ça évolue

$C = \begin{bmatrix} \text{classe 1} \\ \text{classe 2} \\ \vdots \end{bmatrix}$ → 1 ligne = 1 classe

Fusionner 2 classes c_i et c_j $\Rightarrow$ $s_i + s_j$ avec s la matrice qui représente sa classe

Moyenne des éléments de C

$s = \text{ones}(n,1)$ dans matlab ← $s = [1\ 1\ 0 \ldots 0 \ldots 0] = C_i$

$nC = s_k \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ → car s = présence des 1 dans la classe, va être transposé en sa colonne et $\times$ pour multiplier par un vecteur colonne pour le faire.

nbre d'éléments de la classe

$\bar{c} = \frac{1}{nC}(v_1 + v_2 + \ldots + v_{nc})$ → SOMME

centre des classes

$E = \{v_1, v_2, \ldots, v_n\}$, $v_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$

$\bar{c} = \begin{bmatrix} \frac{1}{nc} \cdot s \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ \frac{1}{nc} \cdot s \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \end{bmatrix}$

up(n) → crée I_n sous mathlab

on peut s'arrêter sur autre chose que le nombres de classes :

→ inertie intra classe

$$c : \sum_{v,v' \in C} d(v,v')^2 = i_C$$

distance entre tous les points d'une même classe

La condition : $\left(\frac{\sum_C i_C}{i_E}\right)$ somme des inerties de classes / inertie totalité élément (distance entre tous les points)

puis $\gtrless 1$

$E[0;1]$ et croissant $\leftarrow$ avec que des classes taille 1 $= 0$ / Avec 1 classe $= 1$

• Les k-means

→ nbres de classes fixés K

→ choix de K points → $C = \{x\}$ classe à 1 élem

$F = \{v_1, v_2, \dots, v_K\}$ et $C_k = \{v_k\}$

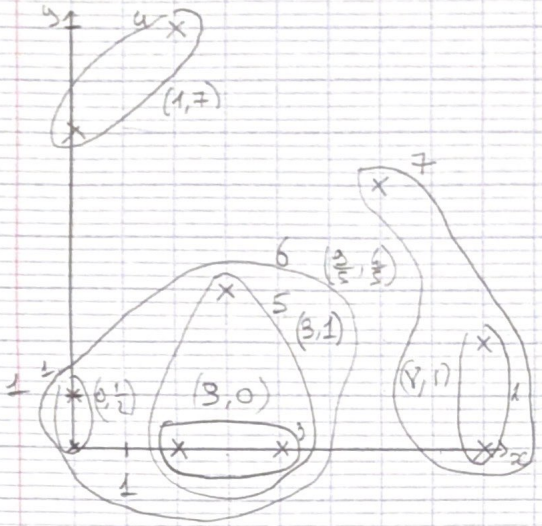
→ tant que $E \setminus F \neq \emptyset$

- choisir $v \in E \setminus F$
- associer v à une des K classes (minimisation des distances)
- $F = F \cup \{v\}$

On peut obtenir un meilleur résultat en prenant le résultat d'une class hiérarchique de condition d'arrêt K. On utilise les moyennes des classes obtenues. (on "remplace" des pts par des moyennes) Mieux que de prendre des points au pif.

Analyse de données : TD 3

Exo 6



On a 10 classes de 1 point.
On va regrouper par dmin

(0,0) et (0,1) → {(0,0),(0,1)}
→ $\bar{m} = (0, \frac{1}{2})$

(8,0) et (8,2) → {(8,0),(8,2)}
→ $\bar{m} = (8,1)$

(2,0) et (4,0) → {(4,0),(4,0)}
→ $\bar{m} = (3,0)$

(0,6) et (7,8) → {(0,6),(7,8)}
→ $\bar{m} = (1,7)$

5: $\bar{m} = \frac{(7,0)+(4,0)+(3,3)}{3} = (3,1)$

6: $\bar{m} = \frac{(2,0)+(4,0)+(3,3)+(0,0)+(0,1)}{5} = (\frac{9}{5}, \frac{4}{5})$

↳ Reste 4 classe à cette étape

7: $\frac{(8,0)+(8,2)+(6,5)}{3} = (\frac{22}{3}, \frac{7}{3})$