Licence d'Informatique 3ème année
AD Développeur (C5-160512-INFO)
CC du 26 octobre 2020 (1h30)

Carl FRÉLICOT – Dpt Info / Lab MIA

- Autorisés : calculatrice (pas une application sur smartphone), un formulaire **manuscrit** sans **aucun exemple numérique**. Tout **autre document est interdit**
- Utilisez les cadres réservés pour inscrire vos réponses.
  Vous devez **écrire la formule** utilisée avant de donner le détail du moindre calcul ou **bien expliquer**.
- Une réponse **non justifiée** sera considérée comme **fausse**.

Nom, prénom : Note : 20 /20

Exercice 1

Soient les données X ci-contre, à partir desquelles on a calculé les tableaux de données centrées X' et de covariance V incomplets.

$X =$

4	-3
5	0
3	-3

$X' =$

0	
1	
-1	

$V =$

2/3	1
1	2

$^t\overline{x} =$

1) Complétez les cases vides.

pour $\overline{x}$: (4+5+3)/3 = 4 et (-3+0-3)/3 = -2

$$X'_2 = X_2 - \begin{pmatrix} -2 \\ -2 \\ -2 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

2) Calculez la distance *cosinus* entre les deux variables.

$$< X_1, X_2 >_D = (4 \;\; 5 \;\; 3) \frac{1}{3} \begin{pmatrix} -3 \\ 0 \\ -3 \end{pmatrix} = -7$$

$$||X_1||^2_D = (4 \;\; 5 \;\; 3) \frac{1}{3} \begin{pmatrix} 4 \\ 5 \\ 3 \end{pmatrix} = 50/3$$

$$||X_2||^2_D = (-3 \;\; 0 \;\; 3) \frac{1}{3} \begin{pmatrix} -3 \\ 0 \\ 3 \end{pmatrix} = 6$$

$$d_{cos}(X_1, X_2) = 1 - \frac{<X_1, X_2>_D}{||X_1||_D \; ||X_2||_D} = 1 - \frac{-7}{\sqrt{50/3}\;\sqrt{6}} = 1 + \frac{7}{10} = 1.7$$

3) Donnez le tableau de données centrées-réduites X'', puis calculez sa matrice de covariance par produit matriciel.

Dans V, on a les D-normes des variables qui valent $\sqrt{3/2}$ et $\sqrt{2}$,

alors à partir de X' on obtient $X'' = \begin{pmatrix} 0 & -\sqrt{2}/2 \\ \sqrt{3/2} & \sqrt{2} \\ -\sqrt{3/2} & -\sqrt{2}/2 \end{pmatrix}$ de matrice de covairance

$$^tX'' \; D \; X'' = \begin{pmatrix} 0 & \sqrt{3/2} & -\sqrt{3/2} \\ -\sqrt{2}/2 & \sqrt{2} & -\sqrt{2}/2 \end{pmatrix} \frac{1}{3} \begin{pmatrix} 0 & -\sqrt{2}/2 \\ \sqrt{3/2} & \sqrt{2} \\ -\sqrt{3/2} & \sqrt{2}/2 \end{pmatrix} = \begin{pmatrix} 1 & \sqrt{3}/2 \\ \sqrt{3}/2 & 1 \end{pmatrix}$$

qui est bien évidemment la matrice de corrélations R

4) Un analyste décide de réaliser une ACP **normée** sur ces données. La variance de la 2ème composante principale vaut 0.134. Peut-on déduire celle de la 1ère ? Si oui, combien vaut-elle ?

l'ACP est normée, donc la somme des variances vaut $trace(R) = p = 2$.
et alors $\lambda_1 = 2 - \lambda_2 = 2 - 0.134 = 1.866$

Auriez-vous fait le même choix que l'analyste ?

non, les variances des variables étant homogènes (2/3 et 2), il ne faut pas faire une ACP normée

5) On donne $U = \frac{\sqrt{2}}{2}$ | 1 | -1 | / | 1 | 1 | et $C =$ | -0.5 | | / | 1.866 | | / | -1.366 | |. Complétez le tableau des composantes principales par produit matriciel.

l'ACP est normée, il faut projeter le nuage X''

$$C'_2 = X''u_2 = \begin{pmatrix} 0 & -\sqrt{2}/2 \\ \sqrt{3/2} & \sqrt{2} \\ -\sqrt{3/2} & -\sqrt{2}/2 \end{pmatrix} \frac{\sqrt{2}}{2} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$
$$= \frac{\sqrt{2}}{2} \begin{pmatrix} -\sqrt{2}/2 \\ -\sqrt{3/2} + \sqrt{2} \\ \sqrt{3/2} - \sqrt{2}/2 \end{pmatrix}$$
$$= \begin{pmatrix} -0.5 \\ 0.134 \\ 0.366 \end{pmatrix}$$

6) Pourra-t-on expliquer la 1ère composante à l'aide du 3ème individu ?

$< x''_3, u_1 >_I = -1.366$

$$||x''_3||_I^2 = \begin{pmatrix} -\sqrt{3/2} & -\sqrt{2}/2 \end{pmatrix} \begin{pmatrix} -\sqrt{3/2} \\ -\sqrt{2}/2 \end{pmatrix} = \frac{3}{2} + \frac{1}{2} = 2$$

$||u_1||_I = 1$

$cos(x''_3, u_1) = \frac{<x''_3, u_1>_I}{||x''_3||_I \ ||u_1||_I} = \frac{-1.366}{\sqrt{2}} = -0.966$

le $cos^2(x''_3, u_1) = (-0.966)^2 = 0.933$ est proche de 1, alors oui
on pourra s'appuyer sur cet individu pour expliquer cette composante

7) Pourra-t-on expliquer la 2ème composante à l'aide de la 1ère variable ?

$$C''_2 = C'_2/\sqrt{\lambda_2} = C'_2/\sqrt{0.134} = \begin{pmatrix} -0.5 \\ 0.134 \\ 0.366 \end{pmatrix} / \sqrt{0.134} = \begin{pmatrix} -1.366 \\ 0.366 \\ 1 \end{pmatrix}$$

X''_1 est centrée-réduite, donc $||X''_1||_D = 1$; de même $||C''_2||_D = 1$

$$< X''_1, C''_2 >_D = \begin{pmatrix} 0 & \sqrt{3/2} & -\sqrt{3/2} \end{pmatrix} \frac{1}{3} \begin{pmatrix} -1.366 \\ 0.366 \\ 1 \end{pmatrix} = -0.259$$

$cos(X''_1, C''_2) = \frac{<X''_1, C''_2>_D}{||X''_1||_D \ ||C''_2||_D} =< X''_1, C''_2 >_D = -0.259$

le $cos^2(X''_1, C''_2) = (-0.259)^2 = 0.067$ est proche de 0, alors non
on ne pourra pas s'appuyer sur cet individu pour expliquer cette composante

Exercice 2

Soit le tableau de données X ci-contre sur lequel a exécuté trois itérations des *C-Means* avec $C = 3$ et la distance euclidienne usuelle.

$X =$

2	4	24	6	x
4	16	18	2	y
0	0	20	20	z
4	6	22	8	t
12	10	16	2	u

1) Faites-les calculs permettant de compléter les tableaux ci-dessous.

$Y^{(0)} = [1, 2, 2, 2, 3] \rightarrow V^{(1)} = \begin{bmatrix} \overline{x}_1 = (2, 4, 24, 6) \\ \overline{x}_2 = (\quad , \quad , \quad , \quad) \\ \overline{x}_3 = (12, 10, 16, 2) \end{bmatrix} \rightarrow$

d^2	x	y	z	t	u
$\overline{x}_1$	0		232	16	216
$\overline{x}_2$	43.56	144.9	160.9	11.56	174.2
$\overline{x}_3$	216	104	584	152	
$Y^{(1)}$	1	3	2	2	

$\rightarrow V^{(2)} = \begin{bmatrix} \overline{x}_1 = (\quad , \quad , \quad , \quad) \\ \overline{x}_2 = (2, 3, 21, 14) \\ \overline{x}_3 = (8, 13, 17, 2) \end{bmatrix} \rightarrow$

d^2	x	y	z	t	u
$\overline{x}_1$	0	200	232	16	216
$\overline{x}_2$	74	326		50	318
$\overline{x}_3$	182	26	566	126	26
$Y^{(2)}$	1	3		1	3

$\rightarrow V^{(3)} = \begin{bmatrix} \overline{x}_1 = (3, 5, 23, 7) \\ \overline{x}_2 = (\quad , \quad , \quad , \quad) \\ \overline{x}_3 = (8, 13, 17, 2) \end{bmatrix} \rightarrow$

d^2	x	y	z	t	u
$\overline{x}_1$		172	212	4	180
$\overline{x}_2$	232	600		200	584
$\overline{x}_3$	182	26	566	126	
$Y^{(3)}$					

itération 1 :
$\overline{x}_2 = \frac{y+z+t}{3} = (8/3, 22/3, 60/3, 30/3) = (2.67, 7.33, 20, 10)$

$d^2(\overline{x}_1, y) = ||\overline{x}_1 - y||^2 = (2-4, 4-16, 24-18, 6-2) \begin{pmatrix} -2 \\ -12 \\ 6 \\ 4 \end{pmatrix} = 200$

$d^2(\overline{x}_3, u) = 0$ car $\overline{x}_3 = u$ et alors $Y^{(1)}(u) = 3$

itération 2 :
$\overline{x}_1 = x = (2, 4, 24, 6)$ car le cluster 1 est un singleton $\{x\}$
$d^2(\overline{x}_2, z) = d^2(\overline{x}_2, t) = 50$ car le cluster 2 ne contient que z et t, équidistants du barycentre
$Y^{(2)}(z) = argmin(232, 50, 566) = 2$

itération 3 :
$\overline{x}_2 = z = (0, 0, 20, 20)$
$d^2(\overline{x}_1, x) = d^2(\overline{x}_1, t) = 4$ car le cluster 2 ne contient que x et t, équidistants du barycentre
$d^2(\overline{x}_2, z) = 0$ car le cluster 1 est un singleton $\{z\}$
$d^2(\overline{x}_3, u) = d^2(\overline{x}_3, y) = 26$ car le cluster 3 ne contient que y et u, équidistants du barycentre
$Y^{(3)} = (argmin(4, 232, 182), argmin(172, 600, 26), argmin(212, 0, 566), argmin(4, 200, 126), argmin(180, 584, 26)$
$= (1, 3, 2, 1, 3)$

2) Calculez la distance de *Chebychev* entre y et z.

$|y - z| = |(4-0, 16-0, 18-20, 2-20)| = (4, 16, 2, 18)$

$d_\infty(y, z) = max(4, 16, 2, 18) = 18$

3) L'inertie *inter-clusters* de la partition finale vaut 86.24. Une autre exécution de l'algorithme renvoit une partition pour laquelle cette inertie vaut 68.37. Vaut-il mieux retenir la première ou la seconde ?

plus l'inertie *inter-clusters* est grande, plus les *clusters* sont séparés,

il faut donc mieux retenir la première (partition finale)

Exercice 3 : que cherche-t-on à faire ?

```
centroids = cmeans(data.T,10)
```

le tableau de données est transposé, on calcule les barycentres de 10 groupes de variables ;

on cherche à réduire la dimensionnalité, passer de p (certainement grand) à 10

Exercice 4 : écrivez une fonction python (sans boucle) qui, à partir de deux tableaux de données de mêmes dimensions, retourne le tableau des cosinus entre les variables du premier et les variables du deuxième et un indicateur validant (ou non) la qualité de représentation des premières par les secondes. Vous pouvez appeler la fonction *standardize* écrite en TP.

```
import numpy as np
def cosvar(data1, data2, seuil = 0.5):
    n, p = data1.shape
# centrer-réduire les données
   data1s = standardize(data1,scale=True)[0]
   data2s = standardize(data2,scale=True)[0]
# métrique dans l'espace des variables
   Dn = np.eye(n) / n
# cosinus
   cosvar = data1s.T.dot(Dn.dot(data2s))
# booléen : cos2 élevé ou non
   qualite = cosvar **2 > seuil
   return cosvar, qualite
```