EXAMEN

ANALYSE DE DONNÉES

DATE : 09 JUIN 2021 — DURÉE : 1H30
SECTION : ING2 GÉNIE INFORMATIQUE — NOMBRE DE PAGES : 2
ENSEIGNANTS : I.KAMMOUN, W.BARHOUMI, S.ZOUAOUI — DOCUMENTS NON AUTORISÉS

Exercice 1 : Clustering (10 pts)

On veut utiliser l'algorithme du k-means et la distance euclidienne pour regrouper les 8 points suivants en 3 clusters : $A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $A_4(5,8)$, $A_5(7,5)$, $A_6(6,4)$, $A_7(1,2)$, $A_8(4,9)$ de même poids. La matrice de distance basée sur la distance Euclidienne est fournie ci-dessous.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{4}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

1. (a) On considère comme centre de classes à l'initialisation les points A_1, A_4 et A_7. Déroulez une première itération de l'algorithme de k-means pour ces données et donnez :
   i. Les nouveaux clusters ;
   ii. Les centres de chaque cluster ;
   (b) Donnez au moins deux conditions d'arrêt possibles de ce processus.
   (c) Comment varie l'inertie totale, l'inertie inter-classes et l'inertie intra-classes dans cette méthode de classification.
2. (a) Déroulez une première itération de l'algorithme de classification ascendante hiérarchique (CAH) pour ces données en adoptant le critère de Ward pour l'agrégation et donnez :
   i. La nouvelle agrégation
   ii. La nouvelle matrice de distance suite à l'agrégation
   (b) Comment varie l'inertie totale, l'inertie inter-classes et l'inertie intra-classes dans cette méthode de classification.
3. Décrivez les avantages et les inconvénients de ces deux méthodes de classification.
4. Décrire brièvement la méthode de la classification non supervisée mixte.

Exercice 2 : Modèle Linéaire (5 pts)

Un certain composant électronique est fabriqué une fois par mois par l'entreprise Micro-Systèmes. La quantité fabriquée varie avec la demande du marché. Dans le but de planifier la production et d'établir certaines normes sur le nombre d'hommes minutes exigés pour la production de différents lots de ce composant électronique, le responsable de la production a relevé l'information suivante pour 15 cédules de production. Le nombre d'hommes-minutes est identifié par Y et la quantité fabriquée par X.

Yi	150	192	264	371	300	358	192	134	242	238	226	302	340	182	169
Xi	35	42	64	88	70	85	40	30	55	60	51	72	80	44	39

1. Quelle serait la première étape à franchir avant d'aborder tout calcul préliminaire ?
2. Le responsable de la production envisage d'utiliser le modèle linéaire simple comme modèle prévisionnel. Spécifiez ce modèle et identifiez chacune des composantes du modèle dans le contexte de ce problème ainsi que les hypothèses du modèle.
3. Estimez les coefficients du modèle en adoptant la méthode des Moindres Carrés Ordinaires (MCO).
   On donne : $\overline{x} = 57$; $\overline{y} = 244$; $\sigma_x^2 = 332.4$; $\sigma_y^2 = 5427.9$; $Cov(X,Y) = 1335.1$
4. D'après l'équation de régression, si le nombre d'unités à fabriquer augmente de 10, quelle sera l'augmentation correspondante du nombre moyen d'hommes- minutes requis ?
5. Déterminez et interprétez le coefficient du détermination R^2 de ce modèle
   On donne : La somme des carrés expliqués par le modèle SCE = 80439.5

Exercice 3 : QCM (5 pts) ***(Cette feuille n'est pas à rendre veuillez indiquer le numéro de la question et les codes des affirmations correctes sur votre feuille d'examen)***

1. Indiquez les affirmations correctes concernant l'Analyse Factorielle des Correspondances :

   A. Plus la distance euclidienne entre deux lignes du tableau de contingence est faible, plus les deux modalités lignes sont associées ;
   B. Chaque cellule ℓ_{ij} du tableau de profils-lignes présente une probabilité conditionnelle de la modalité i de la variable ligne sachant la modalité j de la variable colonne ;
   C. Contrairement à la distance euclidienne, avec la métrique du $\chi 2$, la distance entre deux lignes ne dépend pas des poids respectifs des colonnes ;
   D. On peut réaliser l'étude sur deux variables qualitatives dont le nombre de modalités est différents ;
   E. L'objectif de l'analyse de correspondance simple (AFC) est d'étudier la relation entre une variable qualitative et une variable quantitative ;

2. Indiquez les affirmations correctes concernant l'Analyse en Composantes Principales :

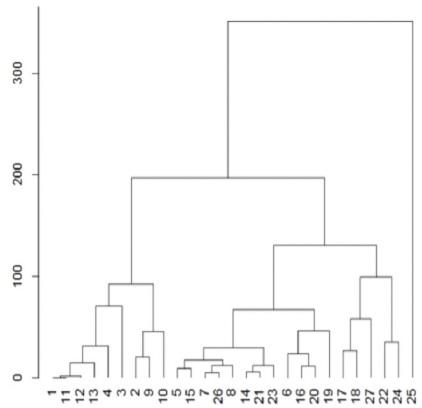
   A. La somme des qualités de représentation d'un individu donné sur tous les axes factoriels est égale à 100% ;
   B. La somme des qualités de représentation de tous les individus sur un axe factoriel donné est égale à 100% ;
   C. Le rang de la matrice de corrélation d'une ACP normée sur p variables vaut p ;
   D. La somme des contributions des individus à l'inertie expliquée par un axe factoriel est égale à 1 ;
   E. L'objectif de l'ACP est de réduire la dimension de l'espace en éliminant certaines variables initiales ;

3. Indiquez les affirmations correctes concernant la classification Floue FCM

   A. Le nombre minimal de variables (ou caractères) nécessaires pour effectuer la classification est 2 ;
   B. Pour chaque individu, la somme des degrés d'appartenance aux différentes classes est égale à 1 ;
   C. Pour chaque classe, la somme des degrés d'appartenance des différents individus à cette classe est égale à 1 ;
   D. Les coordonnées du centre d'une classe sont calculées à partir des individus, formant cette classe, pondérés par leurs degrés d'appartenance à cette classe ;
   E. D'une itération à une autre le critère de variabilité intra-classes diminue ;
   F. Un des critères d'arrêt de l'algorithme FCM est que le critère de variabilité intra-classes soit presque nul ;
   G. Plus la distance entre un individu et le centre d'une classe est grande plus son degrés d'appartenance à cette classe est faible ;

4. On applique une CAH sur 27 individus avec comme critère d'agrégation la distance de Ward. On donne l'arbre de classification avec en ordonnée la perte de l'inertie interclasses engendrée. En coupant l'arbre de classification, on a une inertie inter-classes d'environ 680. Combien a-t-on retenue de classes suite à cette coupure ?

   A. 1 classe ;
   B. 2 classes ;
   C. 3 classes ;
   D. 4 classes ;
   E. 5 classes ;
   F. On ne peut pas savoir ;



Correction Exam 2021

Première Itération.

$C_1 = \{A_1\} \rightsquigarrow G_1 = A_1 = (2;10)$

$C_2 = \{A_4, A_3, A_5, A_6, A_8\} \rightsquigarrow G_2 = \left(\frac{5+8+7+6+4}{5}, \frac{8+4+5+6+3}{5}\right) = (6;6)$

$C_3 = \{A_7, A_2\} \rightsquigarrow G_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1,5;3,5)$

1/ Condition d'arrêt K-means.

$|I_w^{(k+1)} - I_w^{(k)}| < \varepsilon$ L'inertie intra-classe, cste.

La partition reste inchangée.

1/ $I(G) = I_w + I_B$

$I(G)$: cste dans les différentes itérations

$I_w \downarrow$ et $I_B \uparrow$.

2/ a/ $S(A,B) = \frac{1}{2} m \cdot d^2(A,B)$

à la 1ère itération les 2 pts qui seront regroupés sont A_3 et A_5

$S(x_i,x_j)$	A_1	$\{A_3,A_5\}$	A_2	A_4	A_6	A_7	A_8
A_1	0		$\frac{25}{2}$	$\frac{13}{2}$	$\frac{59}{2}$	$\frac{65}{2}$	$\frac{5}{2}$
$\{A_3,A_5\}$		0	.	.	.	.	.
A_2			0	$\frac{18}{2}$	$\frac{97}{2}$	$\frac{19}{2}$	$\frac{20}{2}$
A_4				0	$\frac{17}{2}$	$\frac{5}{2}$	$\frac{2}{2}$
A_6					0	$\frac{29}{2}$	$\frac{29}{2}$
A_7						0	$\frac{58}{2}$
A_8							0

$\{A_3,A_5\} \rightsquigarrow G_{35} = (7,5;4,5)$

$S(A_1, \{A_3,A_5\}) = \frac{1 \times 2}{3} \underbrace{d^2(A_1, G_{35})}_{(2-7,5)^2 + (10-4,5)^2}$

$S(A_2, G_{35}) = \frac{2}{3}\left((2-7,5)^2 + (5-4,5)^2\right)$

$S(A_4, G_{35}) = \frac{2}{3}\left((5-7,5)^2 + (8-4,5)^2\right)$

$S(A_6, G_{35}) = \frac{2}{3}\left((6-7,5)^2 + (4-4,5)^2\right)$

$S(A_7, G_{35}) = \frac{2}{3}\left((1-7,5)^2 + (2-4,5)^2\right)$

$S(A_8, G_{35}) = \frac{2}{3}\left((4-7,5)^2 + (3-4,5)^2\right)$

b/ $I(G) = I_w + I_B$ $I_w \uparrow$ et $I_B \downarrow$

cte dans chaque itération

3/ K-means | CAH

Simple, Rapide	La méthode nous permet de trouver le nb de classes optimal.	Les +
La partition finale dépend de l'initialisation	Plus Complexe que K-means (temps de Calcul)	Les -

4/ Nous pouvons utiliser la méthode de minkowski et le nombre d'individus est très important. et le nombre de classes inconnu.
On applique tout d'abord la méthode de K-means avec un nb de classes assez élevé. ensuite On applique la CAH afin d'avoir le nb de classes optimal puis revenir à K-means avec ce nb de classes trouvé puis que la complexité de cet algorithme est moins importante.

Ex 2: 1/ On peut tracer le nuage de points entre X et Y afin de voir si le nuage est bien allongé et peut être ajusté par une droite. On calculera ensuite le coefficient de corrélation linéaire

2/ $Y = \alpha X + \beta + E$

Le n° H-M variable à expliquer en fonction de X. qtés fabriquées. Hyp: E indép de X

$E_i \sim N(0, \sigma^2)$

E_i indép.

3/ MCO $\rightarrow \hat{\alpha} = \frac{cov(X,Y)}{\sigma_X} = \frac{1335,1}{332,14} \simeq 4$

$\hat{\beta} = \bar{Y} - \hat{\alpha}\bar{X} = 244 - 57 \times 4 = 16.$

4/ si le nb

α = sensibilité $= \frac{\partial Y}{\partial X}$ si X ↗ de 1 unité en moyenne

$= 4$

Y ↗ de 4 unités en moyenne.

5/ $R^2 = \frac{SCE}{SCT} = \frac{80439,5}{15 \times s_y^2} = 0,98$

98% des variations de Y sont expliquées (inhabituelles).
par les variations des quantités fabriquées.
X explique bien la variable Y.