Licence d'Informatique 3, AD Programmeur (C5–160512-INFO) TD 3 – *Clustering* (Apprentissage Non Supervisé)

Carl FRÉLICOT – Dpt Info / Lab MIA

0. Tableaux de Données

On a testé deux logiciels (`Blue` et `Yellow`) de *Machine Learning* et leur association `Green` sur 10 apprenants à qui on a demandé de graduer sur une échelle de 0 à 20 leur ressenti en termes d'effet potentiel sur leur compréhension du *Machine Learning* :

- sans effet (`sans`),
- amélioration (`amel.`) et
- amélioration significative (`amel.+`).

Les résultats sont donnés dans le tableau ci-contre.

	logiciel	sans	amel.	amel.+
1	Blue	2.0	4.0	10.0
2	Blue	2.0	16.0	2.0
3	Yellow	11.0	7.0	5.0
4	Green	16.0	11.0	19.0
5	Yellow	19.0	4.0	4.0
6	Green	3.0	13.0	17.0
7	Yellow	12.0	6.0	11.0
8	Green	10.0	14.0	20.0
9	Blue	1.0	14.0	2.0
10	Blue	7.0	15.0	9.0
	$\overline{x}$	8.3	10.4	9.9
	s	6.0	4.45	6.49

1. Distances

1-1) Calculez les distances de Manhattan et de Chebychev entre les apprenants 6 et 7.
(HW) De même entre les deux premiers, ainsi que leur distance euclidienne.
1-2) Calculez la distance Cosinus entre les variables `amel.` et `amel.+`.
(HW) Calculez leur distance corrélation.

Pour la suite, considérons le tableau T de données limité aux cinq derniers individus renommés x, y, z, t et u.

2. Algorithme *C-means*

Comme il y a 3 modalités d'utilisation des logiciels, on a exécuté trois itérations des `C-means` sur T, avec $c = 3$ et la distance euclidienne usuelle.

2-1) Faites les calculs (ou les déductions !) permettant de compléter les tableaux ci-dessous.

$$Y^{(0)} = [1,2,3,1,2] \;\rightarrow\; V^{(1)} = \begin{bmatrix} \overline{x}_1 = (2\;, \quad , 9{,}25) \\ \overline{x}_2 = (9.5, 10.5, 10), \\ \overline{x}_3 = (10, 14, 20) \;=3 \end{bmatrix} \rightarrow$$

$d^2(\overline{x}_j, x_i)$	x	y	z	t	u
$\overline{x}_1$	57.5	158.5	174.5	57.5	27.5
$\overline{x}_2$	97.5	27.5	112.5	148.5	0,75
$\overline{x}_3$	59	149	0	405	131
$Y^{(1)}$	1	2	3	1	1

$$\rightarrow\; V^{(2)} = \begin{bmatrix} \overline{x}_1 = (11/3, 28/3, \;1\;), \\ \overline{x}_2 = (12, 6, 11), \\ \overline{x}_3 = (3\;, \quad , \quad) \end{bmatrix} \rightarrow$$

$d^2(\overline{x}_j, x_i)$	x	y	z	t	u
$\overline{x}_1$	60.22	136.22	153.89	60.89	12.22
$\overline{x}_2$	166		149	266	110
$\overline{x}_3$	59	149		405	131
$Y^{(2)}$	3	2	3	1	1

$$\rightarrow\; V^{(3)} = \begin{bmatrix} \overline{x}_1 = (4, 14.5, 5.5), \\ \overline{x}_2 = (\quad , \quad , \quad), \\ \overline{x}_3 = (6.5, 13.5, 18.5) \end{bmatrix} \rightarrow$$

$d^2(\overline{x}_j, x_i)$	x	y	z	t	u
$\overline{x}_1$	135.5	166.5	246.5	21.5	
$\overline{x}_2$	166	0	149	266	110
$\overline{x}_3$	14.75	142.75	14.75	302.75	92.75
$Y^{(3)}$	3	2	3	1	1

2-2) Eût-il été judicieux d'itérer davantage ?

3. Autour d'une Partition

On donne ci-contre la matrice de covariance et le centre des données du tableau T, et on s'intéresse à la partition finale $Y^{(3)}$ dont les centres sont dans $V^{(3)}$.

$$V = \begin{bmatrix} 17.04 & -7.84 & 11.72 \\ -7.84 & 10.64 & -0.32 \\ 11.72 & -0.32 & 39.76 \end{bmatrix} \qquad \overline{x} = (6.6, 12.4, 11.8)$$

3-1) On rappelle que le critère optimisé par l'algorithme *C-means* est l'inertie intra-*clusters* $\mathcal{D}(U,V) = \frac{1}{n}\sum_{i=1}^{c}\sum_{x_k \in C_i} d_2^2(x_k, \overline{x}_i)$. Calculez $\mathcal{D}(Y^{(3)}, V^{(3)})$.
3-2) Calculez, à partir de $V^{(3)}$, le tableau G' des *centroids* centrés.
3-3) Posez le calcul de la matrice de covariance de ces *centroids*, mais n'en calculez que les termes diagonaux.
3-4) Comment appelle-t-on cette matrice de covariance ?
3-5) La trace d'une matrice carrée est la somme de ses termes diagonaux ; par exemple ici $trace(V) = 67.44$.
C'est un opérateur linéaire, c'est-à-dire que : $trace(\alpha\, A + \beta\, B) = \alpha\, trace(A) + \beta\, trace(B)$.
Calculez astucieusement la trace de la matrice de covariance intra-*clusters* W. Que retrouvez-vous ?

3-6) Calculez l'indice de Dunn (dont la formule est rappelée ci-desous) de la partition $Y^{(3)}$=(3,2,3,1,1).

$$DI = \frac{min_{1\leq i<i'\leq c} d(\overline{x}_i, \overline{x}_{i'})}{max_{j=1,c}\Delta_j}$$ où $\Delta_j = max_{x_k,x_l\in C_j}\, d(x_k, x_l)$ est le *diamètre* du *cluster* C_j. *min ou max ?*

On donne les carrés des distances entre points : [166. 59. 230. 84. 149. 266. 110. 405. 131. 86.] et ceux entre barycentres : [166.5 176.25 142.75]. *squareform*

(HW) Calculez les indices de Dunn des partitions $Y^{(1)}$ et $Y' = (1, 1, 1, 2, 2)$ obtenue à la suite d'une autre exécution de l'algorithme *C-means*. Quelle est la meilleure des trois ?

4. Autour des Partitions

4-1) Rappelez pourquoi l'algorithme *C-means* ne produit pas toujours la même partition finale.

Il est donc légitime de vouloir comparer des partitions. Il existe pour cela des mesures ou indices dits relatifs. Soient P $(n \times c)$ et Q $(n \times c')$ deux (matrices de) partition stricte en respectivement c et c' *clusters*, on définit la matrice d'accord $N(P,Q) = {}^tP\,Q$ de dimension $(c \times c')$. Si on note :

- $t = \sum_{i=1}^{c}\sum_{j=1}^{c'} n_{ij}^2 - n$
- $u = \sum_{i=1}^{c} n_{i\bullet}^2 - n$, où $n_{i\bullet} = \sum_{i=1}^{c'} n_{ij}$
- $v = \sum_{j=1}^{c'} n_{\bullet j}^2 - n$, où $n_{\bullet j} = \sum_{j=1}^{c} n_{ij}$

alors l'accord entre les partition P et Q peut être mesuré par l'indice de Rand défini par :
$RI(P,Q) = \frac{2\,t-(u+v)}{n\,(n-1)} + 1$.

$RI(P,Q) \in [0,1]$ et bien sûr, $R(P,P) = RI(Q,Q) = 1$.

4-2) Soient la partition finale $Y^{(3)} = (3, 2, 3, 1, 1)$ et une autre $Y' = (1, 1, 1, 2, 2)$ obtenue à la suite d'une autre exécution de l'algorithme *C-means*. Donnez les matrices de partition stricte $P^{(3)}$ et P' correspondantes.

4-3) Calculez $R(P^{(3)}, P')$, puis interprétez le résultat.

(HW) Calculez $R(P^{(3)}, P^{(1)})$ où $P^{(1)}$ est la matrice partition associée à $Y^{(1)}$.