Licence Informatique 3 – AD Développeur
4) *Apprentissage Non Supervisé* (TD) – EC 160-512

Pr. 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 des 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 (`S`), amélioration (`A`) et amélioration significative (`A+`). Les résultats sont donnés ci-contre.

	S	A	A+	ML
x	3	13	17	Green
y	12	6	11	Yellow
z	10	14	20	Green
t	1	14	2	Blue
u	7	15	9	Blue

$\overline{x}$	6.6	12.4	11.8
s	4.13	3.26	6.30

1. Distances

1-1) Calculez les distances de Manhattan, de Chebychev et euclidienne entre les apprenants x et y.

(HW) De même entre les apprenants t et u.

1-2) Calculez la distance *cosinus* entre les variables `A` et `A+`. ***Orange ?***

(HW) Calculez leur distance *corrélation*.

2. Algorithme *K-means*

2-1) Comme il y a 3 modalités d'utilisation des logiciels, on a exécuté 3 itérations de l'algorithme *K-means* sur le tableau X des données numériques, et on a choisi $c = 3$ et la distance euclidienne usuelle. Que pensez-vous de ce choix ?

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

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

$d^2(\overline{x}_j, x_i)$	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	
$\overline{x}_3$	59	149		405	131
$Y^{(1)}$	1	2		1	1

$\rightarrow C^{(2)} = \begin{bmatrix} \overline{x}_1 = (\quad,\quad,\quad), \\ \overline{x}_2 = (12, 6, 11), \\ \overline{x}_3 = (\quad,\quad,\quad) \end{bmatrix} \rightarrow$

$d^2(\overline{x}_j, x_i)$	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			1	1

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

$d^2(\overline{x}_j, x_i)$	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
$Y^{(3)}$					

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

2-3) Peut-on affirmer que $Y^{(3)}$ est la meilleure de toutes les partitions possibles de X en $c = 3$ clusters ?

3. Autour d'une Partition

On donne ci-contre la matrice de covariance des données X, et on s'intéresse à la partition finale $Y^{(3)}$ dont les centres sont dans $C^{(3)}$.

$V =$

17.04	-7.84	11.72
-7.84	10.64	-0.32
11.72	-0.32	39.76

3-1) On rappelle que le critère optimisé par l'algorithme *K-means* est l'inertie intra-*clusters* $\mathcal{D}(U, C) = \frac{1}{n} \sum_{i=1}^{c} \sum_{x_k \in C_i} d_2^2(x_k, \overline{x}_i)$. Calculez $\mathcal{D}(Y^{(3)}, C^{(3)})$.

3-2) Calculez, à partir de $C^{(3)}$, le tableau C' des *centroids* centrés. attention à la notation C'

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) Imaginez qu'une fonction (la vôtre par exemple) retourne non seulement U (ou Y) et C mais aussi $\mathcal{D}(U,C)$. Comment trouver la valeur de l'inertie totale de la partition en appelant une seule fois cette fonction ? Que contient alors le tableau C ?

3-7) On rappelle la formule de l'indice de Dunn permettant d'évaluer une partition

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

Calculez sa valeur pour la partition $Y^{(3)}$=(3,2,3,1,1) obtenue par *K-means*. Pour cela, on donne les carrés des distances entre *squareform*

- points : [166. 59. 230. 84. 149. 266. 110. 405. 131. 86.] ,
- barycentres : [166.5 176.25 142.75].

(HW) A priori, peut-on affirmer/infirmer qu'entre 2 itérations de l'algorithme *K-means* on a

- $DI(Y^{(t-1)}, C^{(t)}) \geq DI(\mathrm{Y}^{(t-1)}, C^{(t-1)})$
- $DI(Y^{(t)}, C^{(t)}) \geq DI(\mathrm{Y}^{(t-1)}, C^{(t)})$
- $DI(Y^{(t)}, C^{(t)}) \geq DI(\mathrm{Y}^{(t-1)}, C^{(t-1)})$

(HW) A priori, peut-on affirmer/infirmer que $DI(U,C) \leq DI(U',C')$ si U et U' sont deux partitions en c et c' clusters avec $c < c'$?

(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.

En plus d'évaluer la qualité d'une partition, 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 $n_{i\bullet} = \sum_{j=1}^{c'} n_{ij}$ $\quad \sum_{i=1}^{c} n_{i\bullet} = n$
- $v = \sum_{j=1}^{c'} n_{\bullet j}^2 - n$, où $n_{\bullet j} = \sum_{i=1}^{c} n_{ij}$ $\quad \sum_{j=1}^{c'} n_{\bullet j} = n$

alors l'accord entre les partition P et Q peut être mesuré par l'indice de Rand $RI(P,Q) = \frac{2\,t-(u+v)}{n\,(n-1)} + 1$.
On montre facilement que $RI(P,Q) \in [0,1]$ et bien sûr que $R(P,P) = RI(Q,Q) = 1$.

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) Donnez la matrice d'accord $N(P^{(3)}, P')$.

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

4-5) A priori, pensez-vous que $Y^{(1)} = (1,2,3,1,1)$ est plus compatible avec $Y^{(3)}$ que Y' ?

(HW) Calculez $R(P^{(3)}, P^{(1)})$ pour vérifier votre intuition.

(HW) Soient deux partitions $Y_1 = (1,1,1,1,1)$ et $Y_n = (1,2,3,4,5)$. À quel point sont-elles compatibles/concordantes ?