Université Paris Cité
UFR de Mathématiques et Informatique

Classification Master 1 IMB
Examen partiel du 10 mars 2022 - Correction de l'exercice 1

Exercice 1
Les observations suivantes ont été récoltées, on souhaite effectuer une classification ascendante hiérarchique (CAH) à partir des variables X^1="Température supérieure à 38°C", X^2="Toux", X^3="Gorge irritée" et X^4="Fumeur".

Individu	X^1	X^2	X^3	X^4
e_1	1	1	1	1
e_2	0	1	0	1
e_3	0	0	0	1
e_4	1	1	1	0
e_5	1	0	1	0

Pour chacune des variables la modalité 1 indique respectivement une température supérieure à 38°C, une présence de toux, une gorge irritée ou le fait d'être fumeur.

Dans la suite on va utiliser l'indice de dissimilarité de Jaccard, dont on rappelle la formule :

$$\delta(e_i, e_j) = 1 - \frac{a_{i,j}}{a_{i,j} + a_{i,\bar{j}} + a_{\bar{i},j}}$$

où $a_{i,j} = \text{Card}\{p, X_i^p = X_j^p = 1\}$, $a_{i,\bar{j}} = \text{Card}\{p, X_i^p = 1, X_j^p = 0\}$, $a_{\bar{i},j} = \text{Card}\{p, X_i^p = 0, X_j^p = 1\}$. On obtient le tableau de distances suivant entre les individus

Individu	e_1	e_2	e_3	e_4
e_2	0.5			
e_3	0.75	0.5		
e_4	0.25	•	1	
e_5	0.5	1	•	0.33

1. Quel(s) autre(s) indice(s) de dissimilarité vu(s) en cours aurait-on pu utiliser ?

> **Correction.** Comme toutes les variables sont binaires, on pourrait utiliser les autres dissimilarités adaptées aux données binaires, telles que l'indice de Dice. Ces indices de dissimilarités sont tous fonctions des quantités $a_{i,j}$, $a_{i,\bar{j}}$, $a_{\bar{i},j}$, $a_{\bar{i},\bar{j}}$

2. Montrer que δ est bien un indice de dissimilarité.

> **Correction.** On doit vérifier les propriétés suivantes :
> — $\forall i, j \quad \delta(e_i, e_j) \geqslant 0$: ceci est vrai car les quantités $a_{i,j}$, $a_{i,\bar{j}}$, $a_{\bar{i},j}$ sont toutes $\geqslant 0$, donc $a_{i,j} + a_{i,\bar{j}} + a_{\bar{i},j} \geqslant a_{i,j}$, donc $0 \leqslant \frac{a_{i,j}}{a_{i,j}+a_{i,\bar{j}}+a_{\bar{i},j} \geqslant a_{i,j}} \leqslant 1$, et par conséquent $1 - \frac{a_{i,j}}{a_{i,j}+a_{i,\bar{j}}+a_{\bar{i},j} \geqslant a_{i,j}} \geqslant 0$.

— $\forall i, \quad \delta(e_i, e_i) = 0$: ceci est vrai car d'après la définition de $a_{i,\bar{i}}$ on a

$$a_{i,\bar{i}} = \text{Card}\{p, X_i^p = 1, X_i^p = 0\} = 0$$

et de même $a_{\bar{i},i} = 0$, donc

$$\delta(e_i, e_i) = 1 - \frac{a_{i,i}}{a_{i,i}} = 0.$$

— $\forall i, j, \quad \delta(e_i, e_j) = \delta(e_j, e_i)$: ceci est vrai car

$$a_{j,i} = \text{Card}\{p, X_j^p = X_i^p = 1\} = a_{i,j},$$

et

$$a_{j,\bar{i}} = \text{Card}\{p, X_j^p = 1, X_i^p = 0\} = a_{\bar{i},j}.$$

Donc

$$\delta(e_j, e_i) = 1 - \frac{a_{j,i}}{a_{j,i} + a_{j,\bar{i}} + a_{\bar{j},i}} = 1 - \frac{a_{i,j}}{a_{i,j} + a_{\bar{i},j} + a_{i,\bar{j}}} = 1 - \frac{a_{i,j}}{a_{i,j} + a_{i,\bar{j}} + a_{\bar{i},j}} = \delta(e_i, e_j).$$

3. Calculer les valeurs manquantes $\delta(e_2, e_4)$ et $\delta(e_3, e_5)$ dans le tableau.

Correction. Pour e_2 et e_4 on a $a_{2,4} = 1$ (variable X^2), $a_{2,\bar{4}} = 1$ (variable X^4), et $a_{\bar{2},4} = 2$ (variables X^1 et X^3). Donc

$$\delta(e_2, e_4) = 1 - \frac{1}{1 + 1 + 2} = \frac{3}{4} = 0.75.$$

Pour e_3 et e_5 on a $a_{3,5} = 0$, donc forcément

$$\delta(e_3, e_5) = 1.$$

Le tableau devient donc

Individu	e_1	e_2	e_3	e_4
e_2	0.5			
e_3	0.75	0.5		
e_4	0.25	0.75	1	
e_5	0.5	1	1	0.33

4. Effectuer une CAH en adoptant la stratégie d'agrégation du maximum

$$d(A, B) = \max_{i \in A, j \in B} \delta(e_i, e_j).$$

Justifier chacune des étapes.

Correction.

— **partition initiale** : $\{\{e_1\},\{e_2\},\{e_3\},\{e_4\},\{e_5\}\}$.

— **1ère étape** : ici la stratégie d'agrégation ne joue pas, puisque les groupes sont des singletons. On choisit de regrouper les deux observations les plus proches au sens de δ. On doit donc regrouper e_1 et e_4, et la partition devient

$$\{\{e_1,e_4\},\{e_2\},\{e_3\},\{e_5\}\}.$$

On retient que la valeur du critère pour ce regroupement est 0.25.

— **2ème étape** : On doit à présent comparer les dissimilarités entre groupes. On a

$$d(\{e_1,e_4\},\{e_2\}) = \max(d(\{e_1\},\{e_2\}),d(\{e_4\},\{e_2\})) = \max(0.5,0.75) = 0.75,$$

$$d(\{e_1,e_4\},\{e_3\}) = \max(d(\{e_1\},\{e_3\}),d(\{e_4\},\{e_3\})) = \max(0.75,1) = 1,$$

$$d(\{e_1,e_4\},\{e_5\}) = \max(d(\{e_1\},\{e_5\}),d(\{e_4\},\{e_5\})) = \max(0.5,0.33) = 0.5,$$

$$d(\{e_2\},\{e_3\}) = 0.5,$$

$$d(\{e_2\},\{e_5\}) = 1,$$

$$d(\{e_3\},\{e_5\}) = 1.$$

On choisit les groupes les plus proches, et on voit qu'on a le choix entre regrouper $\{e_1,e_4\}$ et $\{e_5\}$, ou bien $\{e_2\}$ et $\{e_3\}$. Choisissons $\{e_1,e_4\}$ et $\{e_5\}$. La partition devient :

$$\{\{e_1,e_4,e_5\},\{e_2\},\{e_3\}\}.$$

La valeur du critère pour ce regroupement est 0.5.

— **3ème étape** : On a

$$d(\{e_1,e_4,e_5\},\{e_2\}) = \max(d(\{e_1,e_4\},\{e_2\}),d(\{e_5\},\{e_2\})) = \max(0.75,1) = 1,$$

$$d(\{e_1,e_4,e_5\},\{e_3\}) = \max(d(\{e_1,e_4\},\{e_3\}),d(\{e_5\},\{e_3\})) = \max(1,1) = 1,$$

$$d(\{e_2\},\{e_3\}) = 0.5.$$

On choisit les groupes les plus proches, donc $\{e_2\}$ et $\{e_3\}$. La partition devient :

$$\{\{e_1,e_4,e_5\},\{e_2,e_3\}\}.$$

La valeur du critère pour ce regroupement est 0.5.

— **4ème étape** : cette fois il n'y a plus qu'un seul regroupement possible. Calculons tout de même la distance, que nous devons noter pour tracer le dendrogramme. On a

$$d(\{e_1,e_4,e_5\},\{e_2,e_3\}) = \max(d(\{e_1,e_4,e_5\},\{e_2\}),d(\{e_1,e_4,e_5\},\{e_3\})) = \max(1,1) = 1.$$

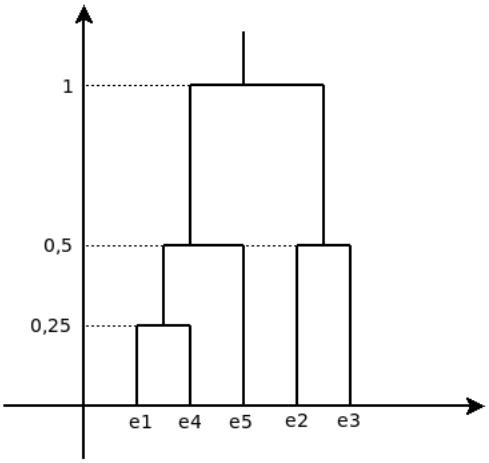
La partition finale est :

$$\{\{e_1,e_2,e_3,e_4,e_5\}\},$$

et la valeur du critère pour ce dernier regroupement est 1.

5. Tracer le dendrogramme associé.

Correction.



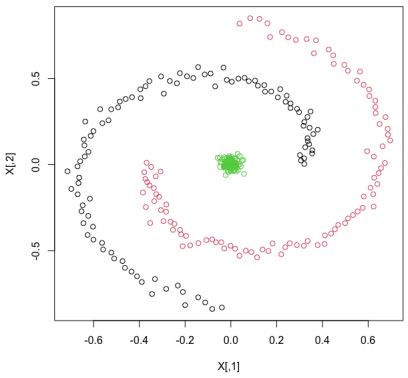
6. Combien de classes peut-on retenir ? Expliciter pour chaque individu la classe à laquelle il a été associé.

Correction. Si on suit le critère du saut maximal de l'indice, on voit nettement sur le dendrogramme qu'on doit retenir le regroupement en deux classes. La partition retenue est donc $\{\{e_1, e_4, e_5\}, \{e_2, e_3\}\}$, autrement dit e_1, e_4 et e_5 sont affectés à une classe (qu'on peut noter classe 0 par exemple), tandis que e_2 et e_3 sont affectés à une autre classe (qu'on peut noter classe 1).

Exercice 2

Dans cet exercice on va travailler sur des données synthétiques de points dans le plan, pour comparer plusieurs méthodes de partitionnement. Dans un premier temps on va construire le jeu de données, qui sera composé de trois ensembles de points : deux spirales intriquées et un amas central.

1. Le package mlbench fournit plusieurs fonctions pour construire des données synthétiques classiques. Après l'avoir chargé, regarder l'aide de la fonction mlbench.spirals, puis l'utiliser pour créer un jeu de données `X_a` contenant $n = 200$ observations à 2 variables, avec les paramètres cycles=0.75 et sd=0.025. Afficher ce jeu de données avec la commande plot(X). Enregistrer également dans une variable `cl_a` les classes associées aux observations.
2. On va ajouter à ces données 100 observations (c'est-à-dire 100 points de $\mathbb{R}^2$) qui constitueront une troisième classe. Les coordonnées de ces points seront générés indépendamment suivant une loi normale centrée d'écart-type 0.025. Créer un tableau `X_b` contenant ces nouvelles données, ainsi que le vecteur `cl_b` correspondant (qui ne doit donc contenir que des 3).
3. Enfin utiliser la fonction rbind pour concaténer les données `X_a` et `X_b` afin d'obtenir le jeu de données final X de 300 observations et 2 variables. Définir également le vecteur cl correspondant aux classes associées aux observations, en concaténant `cl_a` et `cl_b`.
4. Afficher les données avec la commande plot(X,col=cl). Le résultat devrait ressembler au graphique suivant :



N.B. Pour la suite, si vous n'avez pas pu construire correctement le jeu de données, vous pouvez le charger directement avec la commande suivante :

```
load(url("https://tinyurl.com/yckk9caf"))
```

5. Utiliser la fonction dist.quant du package ade4 pour calculer les distances euclidiennes entre les observations de X. Pourquoi le choix de la distance euclidienne simple est a priori justifié ici ?
6. Effectuer des classifications ascendantes hiérarchiques en faisant varier les stratégies d'agrégation. Afficher à chaque fois le dendrogramme obtenu. Quelle stratégie vous semble appropriée au vu des dendrogrammes, et quel nombre de classes sélectionnerait-on pour cette stratégie sur le critère du saut maximal ?
7. On décide de toute manière de créer des partitions en 3 classes, puisque ça correspond à la construction des données. Pour chacune des stratégies testées à la question précédente, définir la partition en 3 classes correspondante et afficher le nuage de points coloré suivant

la partition obtenue (comme à la question 3). Calculer également la table de contingence pour comparer la partition obtenue avec la partition originale cl.

8. Commenter les résultats obtenus à la question précédente. Quelle stratégie fonctionne bien pour ce jeu de données, et pourquoi à votre avis ?
9. A présent tester la méthode des K-moyennes sur le jeu de données, en choisissant toujours 3 classes, et en gardant les paramètres par défaut de la fonction kmeans. Comme précédemment, afficher le nuage de points correspodant à la partition obtenue et calculer la table de contingence pour la comparer avec cl. Commenter le résultat.
10. En fait, on peut montrer que la méthode des K-moyennes ne pourrait pas de toute façon produire une partition proche de la vraie partition pour de telles données. Essayer d'expliquer pourquoi en réfléchissant à quoi correspondent géométriquement les zones où se situent les points correspondant à chaque groupe dans la partition finale obtenue par les K-moyennes.