Classe : 4Data Science
Proposé par : M^ed Heny Selmi

Travaux Dirigés – Les Techniques de Segmentation

Exercice 1 :

On souhaite découper un échantillon de patientes atteintes de la maladie de l'ostéoporose en groupes homogènes.

1. Un premier médecin propose de construire deux groupes de patientes via la méthode K-means : patientes atteintes de fractures de la hanche et patientes atteintes de tassements vertébraux. Afin de quantifier la qualité du découpage, le médecin propose de croiser le résultat de l'algorithme K-means sur un échantillon de patientes caractérisé par une variable qui identifie les patientes réellement fracturées à la hanche par les caractères FH et les patientes atteintes de tassements vertébraux par le caractère TV.
Le résultat du croisement génère la table de confusion suivante :

	1	2
FH	4	54
TV	72	7

En se basant sur les résultats de la table de confusion, quantifier la qualité du découpage obtenu.

2. Un deuxième médecin propose d'appliquer la classification hiérarchique ascendante afin de classifier les patientes. Après la construction du dendrogramme, le médecin propose de faire le découpage de son dendrogramme au niveau de la plus forte perte d'inertie interclasses.

2.1. Citer l'avantage de la classification hiérarchique ascendante par rapport à la méthode K-means dans le choix du nombre de classes à construire.

2.2. Le découpage du dendrogramme au niveau de la plus forte perte d'inertie interclasses génère deux groupes de patientes.

En croisant le résultat de classification avec la variable de l'échantillon d'apprentissage qui identifie les patientes réellement fracturées à la hanche et les patientes atteintes de tassements vertébraux, on obtient la table de confusion suivante :

	1	2
FH	56	2
TV	9	70

Quantifier la qualité du découpage obtenu et comparer les résultats de la classification hiérarchique ascendante aux résultats de la méthode K-means. Conclure

Exercice 2 :

1. Donner le **principe** algorithmique de la méthode de segmentation CAH.
2. On désire appliquer la méthode CAH sur les données suivantes : 5 individus caractérisés par deux variables X_1 et X_2.

Déterminer, pour chaque phase de l'algorithme, la **mise à jour des individus** et la **matrice des distances**.

NB : Utiliser la distance de Manhattan $d(I, J) = |X_1(I) - X_1(J)| + |X_2(I) - X_2(J)|$

Mise à jour des Individus	*Matrice des distances*
Phase 1 : | | X_1 | X_2 | | I_1 | 12 | 5 | | I_2 | 8 | 16 | | I_3 | 14 | 5 | | I_4 | 8 | 10 | | I_5 | 2 | 20 |	| | ..... | ..... | ..... | ..... | ..... | ..... | | | | | | | ..... | | | | | | | ..... | | | | | | | ..... | | | | | | | ..... | | | | | |
Phase 2 : | | X_1 | X_2 | | ..... | | | | ..... | | | | ..... | | | | ..... | | |	| | ..... | ..... | ..... | ..... | ..... | | | | | | ..... | | | | | | ..... | | | | | | ..... | | | | |
Phase 3 : | | X_1 | X_2 | | ..... | | | | ..... | | | | ..... | | |	| | ..... | ..... | ..... | ..... | | | | | ..... | | | | | ..... | | | |
Phase 4 : | | X_1 | X_2 | | ..... | | | | ..... | | |	| | ..... | ..... | ..... | | | | ..... | | |

Exercice 3 :

NB : Utiliser la distance de Manhattan $d(I, J) = |X_1(I) - X_1(J)| + |X_2(I) - X_2(J)|$

On donne le dataset suivant :

1. Le choix du nombre de groupe dans la méthode k-means est considéré comme le point **faible** le plus gênant. Expliquer ce point de vue dans un domaine applicatif.
2. Calculer la **matrice des distances**, En déduire les deux individus **les plus distants**.
3. On souhaite trouver une partition des individus I_1, I_2, I_3, I_4, I_5, I_6 et I_7 en deux groupes. En prenant **comme centroïdes** de départ les deux individus obtenus dans la question précédente, donner la segmentation correspondante.
4. Une **autre** segmentation a donné la composition suivante :

Groupe 1	I_1, I_2, I_3
Groupe 2	I_4, I_5, I_6
Groupe 3	I_7

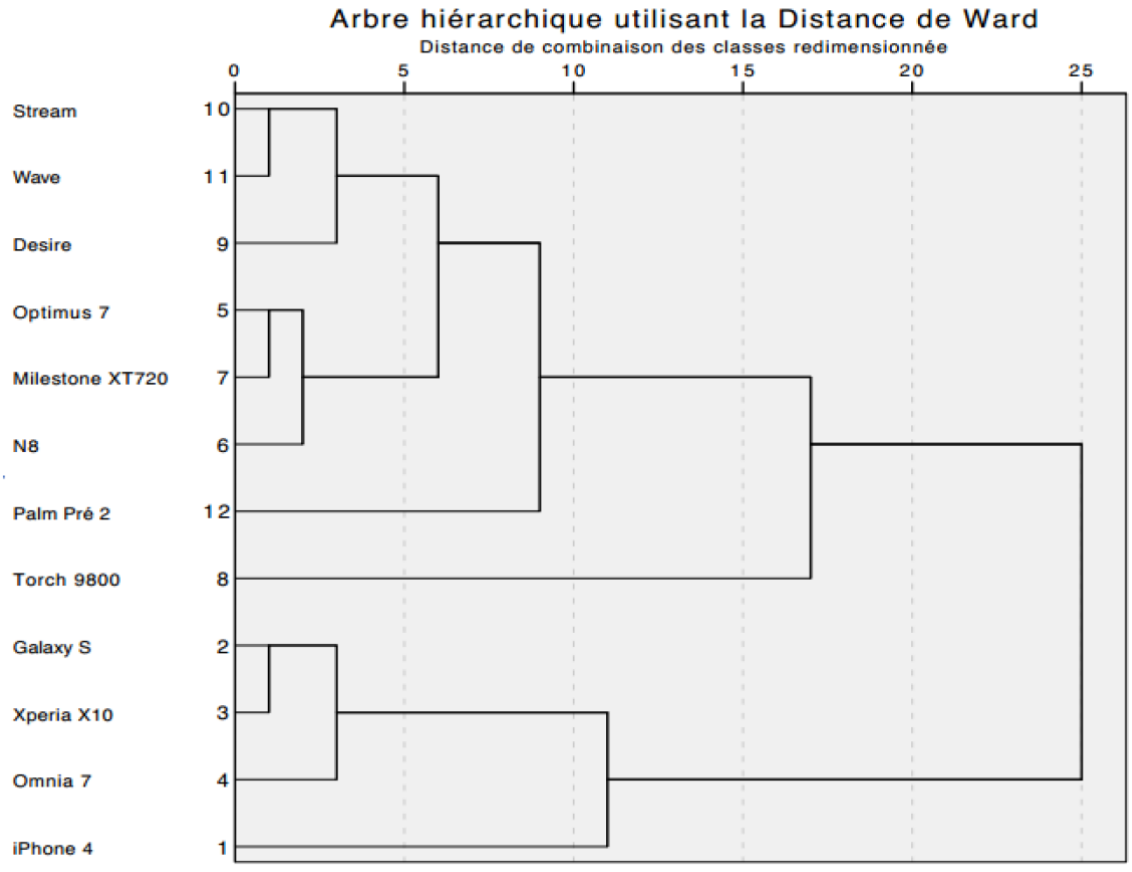
Comparer cette solution avec la vôtre, en précisant votre critère d'évaluation utilisé.

Exercice 4 :

Un magazine français a publié un comparatif des **12** principaux Smartphones (téléphones mobiles connectés disponibles sur le marché en novembre 2010.
Chaque Smartphone est décrit et évalué par les points suivants :

- **Prise en main** : facilité de prise en main (note sur 20)
- **Communication** : qualité des communications téléphoniques (note sur 20)
- **Organisation** : fonctionnalités d'organisation : agenda, carnet d'adresses, etc. (note sur 20)
- **Divertissement** : offre en divertissement : jeux, etc. (note sur 20)
- **Navigation** : offre en logiciel de navigation : Maps, GPS, etc. (note sur 20)
- **Prix** : prix public hors abonnement (en euros)
- **Autonomie** : autonomie en communication (en heures)

Source : Le Point no 2045, 11 novembre 2010 : Le guide du numérique 2011.
En premier lieu, des statistiques descriptives ont été réalisées afin de mieux comprendre les données collectées.



Une segmentation automatique des Smartphones est réalisée.

1. Est-ce qu'un centrage et une réduction vous semblent indispensables sur ces données ? Justifier ce choix ?

2. Justifier le choix d'une méthode hiérarchique ascendante pour réaliser la segmentation au lieu de k-means.

3. Selon les résultats fournis en annexe de la segmentation hiérarchique ascendante, proposer un Smartphone qui pourra remplacer le plus un Iphone 4.

4. Combien de groupes suggérez-vous de retenir selon le dendrogramme ?

5. Quel est le critère utilisé pour déterminer la meilleure segmentation possible avec la coupe du dendrogramme ?