DATA MINING

TD1 : SEGMENTATION

Exercice 1 : Segmentation avec K-Means

Un opérateur téléphonique souhaite analyser les données de ces clients afin d'identifier ceux qui sont susceptibles de changer d'opérateur (les clients susceptibles de churner). Les données disponibles sont composées des variables quantitatives Age_C : les âges des clients, Durees_A : les durées d'appels par jour, Nbre_A : les nombres d'appels par jour, Nbre_SMS : les nombres des SMS envoyés par jour et une variable qualitative Churn_C qui prend la valeur 1 si le client a déjà churné et la valeur 0 si le client est encore fidèle à son opérateur téléphonique. L'échantillon étudié est composé de 87 clients de la classe 0 et 56 clients de la classe 1. On propose tout d'abords d'appliquer une méthode descriptive de groupage (Clustering) des clients via la méthode K-means en utilisant les variables quantitatives disponibles.

1. Décrire les étapes de l'algorithme de la méthode K-means.
2. Citer deux inconvénients de la méthode K-means.
3. L'application de la méthode K-means, en fixant le nombre de groupe K=2, génère la matrice de confusion suivante en croisant la variable Churn_C avec les résultats de la classification de K-means:

	Groupe 1	Groupe 2
Classe 0	7	80
Classe 1	52	4

   Calculer les taux de bonne classification de chaque classe de la variable Churn_C et le taux de bonne classification totale.
4. Est-ce que la méthode K-means génère une bonne classification des clients ? Justifier votre réponse.

Exercice 2 : Segmentation avec Kmeans et CAH

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 3 : Segmentation avec Kmeans

On souhaite découper la clientèle d'une banque en groupes homogènes. On propose d'appliquer le découpage via la méthode K-means.

1) Expliquer l'utilité d'un tel découpage dans le domaine de la banque, en se basant sur un exemple.
2) Parmi les critères d'arrêt de l'algorithme K-means, on cite la stabilisation de l'inertie totale des groupes résultants. Expliquer comment cette stabilisation de l'inertie permet de générer des groupes homogènes.
3) Citer deux inconvénients de la méthode K-means.
4) Un expert de la banque propose de construire deux groupes de clients : clients à tendance épargne et clients à tendance non épargne. Afin de quantifier la qualité du découpage, on propose de croiser le résultat de l'algorithme K-means sur un échantillon de clients caractérisé par une variable qui identifie les clients réellement à tendance épargne par le caractère E et les autres clients par le caractère NE.
   Le résultat du croisement génère la table de confusion suivante :

	1	2
E	4	54
NE	72	7

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

Exercice 4 : Segmentation avec CAH

Un laboratoire d'écologie étudie par prélèvement répétés la présence d'espèces sur quatre sites. Il obtient la matrice suivante qui contient pour chaque paire de sites A et B , le nombre d'espèces communes aux deux sites .

	S1	S2	S3	S4
S1	10	5	8	5
S2	5	15	4	11
S3	8	4	11	3
S4	5	11	3	12

Par exemple d'après le dataset, on peut lire qu'il y a 10 espèces présentes sur le site S1 et 5 espèces communes entre S1 et S2.

On souhaite classer ces sites en utilisant une distance adaptée. Ainsi, la distance entre deux sites A et B est donné par : d(A,B) = $\frac{nA+nB-2n\,AB}{nA+nB}$ ou nA (respectivement nB) désigne le nombre d'espèces présentes sur A (respectivement au site B) et nAB le nombre d'espèces communes aux sites A et B .

1. Calculer d(S1, S2) puis d(S1, S3) en expliquant vos calculs
2. Quelle est la distance de deux sites ayant les mêmes espèces exactement ?
3. Quelle est la distance de deux sites n'ayant pas d'espèces en commun ?
4. Calculer la matrice des distances
5. Tracer le dendrogramme de la classification hiérarchique ascendante en graduant l'axe vertical
6. Ou faut-il faire la coupe ? justifier votre réponse
7. Quelle sont les classes ainsi obtenues ? Interpréter le résultat de la segmentation.