Université Paris Descartes
UFR de Mathématiques et Informatique

Classification Master 1 Ingénierie Mathématique
Examen partiel du 8 mars 2021 - durée 1h30

On écrira tous les codes R dans un document `Nom_Prenom.R` ou `Nom_Prenom.Rmd`. Les packages suivants seront utilisés : `ade4`, `MASS`, `DiscriMiner`. Utiliser la commande `library` pour charger ces packages. Si un ou plusieurs de ces packages ne sont pas disponibles, utiliser la commande `install.packages` pour les installer.

Exercice 1

Soit le tableau de données suivant (2 variables, 5 observations) :

	e_1	e_2	e_3	e_4	e_5
X^1	0	0	0	2	6
X^2	6	2	0	0	0

1. Représenter les données sur un graphique, les axes correspondant aux variables X^1, X^2.
2. Appliquer à la main l'algorithme des K-moyennes sur les données, avec $K = 3$ et en choisissant la partition initiale suivante : $\{\{e_1, e_3\}, \{e_2, e_5\}, \{e_4\}\}$. On décrira précisément chacune des étapes et la partition obtenue à la fin.
3. La partition finale est-elle optimale selon vous ? Quelle devrait être cette partition optimale ?
4. Rappeler pourquoi l'algorithme des K-moyennes peut être vu comme une méthode pour minimiser l'inertie intra-classe des observations. Expliquer pourquoi chacune des deux étapes de l'algorithme permet de réduire cette inertie.
5. Calculer l'inertie intra-classe pour la partition initiale, pour la partition finale, et pour la partition optimale (telle que proposée à la question 3). Commenter les 3 valeurs obtenues.
6. Il y a en fait plusieurs algorithmes des K-moyennes classiques. L'algorithme de MacQueen diffère de l'algorithme classique de Forgy en ce qu'il met à jour les centres des classes à chaque fois qu'une observation est ré-affectée à une classe, et non pas après que toutes les observations aient été affectées. Quel est l'avantage de la méthode de MacQueen par rapport à l'algorithme classique ? Quel peut être son inconvénient ?
7. Reprendre l'algorithme des K-moyennes détaillé à la question 2 en appliquant cette fois la méthode de MacQueen. Aboutit-on à une partition finale différente ? La partition finale est-elle obtenue plus rapidement ?

Exercice 2

Le jeu de données fgl de la librairie MASS contient les résultats de diverses mesures chimiques et physiques effectuées sur des fragments de verre récupérés au sol. Chaque fragment est classé suivant son type correspondant à l'usage d'origine du verre : verre de fenêtre, de bouteille, de phare de véhicule, etc.

On va ici utiliser ce jeu de données pour tester des méthodes de classification non supervisée.

1. Charger les données puis les observer pour comprendre comment elles sont organisées. Créer un tableau de données contenant toutes les variables sauf la classe.
2. Quelle dissimilarité vous semble a priori la plus appropriée pour ce jeu de données ?
3. Effectuer des classifications ascendantes hiérarchiques en faisant varier les dissimilarités et les stratégies d'agrégation. Afficher à chaque fois le dendrogramme obtenu. Finalement, sélectionner une dissimilarité et une stratégie pour la suite des questions, en justifiant votre choix.
4. Déterminer le nombre de classes optimal en détectant le saut maximal du critère d'agrégation, puis calculer la partition correspondante.
5. Comparer la partition obtenue avec la partition induite par le type de verre, en affichant une table de contingence. Commenter le résultat.
6. La fonction `withinSS` du package `DiscriMiner` permet de calculer l'inertie intra-classes. Utiliser cette fonction pour calculer l'inertie intra-classes pour la partition obtenue par la CAH, et pour la partition induite par le type de verre. Laquelle de ces deux partitions a la plus petite inertie intra-classe ? Peut-on en conclure que cette partition est meilleure ?
7. Reprendre les deux dernières questions en choisissant cette fois pour la CAH un nombre de classes égal au nombre de types de verre. Peut-on comparer cette fois les deux partitions sur le critère de l'inertie intra-classe ?