ANALYSE DE DONNÉES
TRAVAUX DIRIGÉS ET PRATIQUES

Fiche n°3 : Classification automatique

Exercice 1

On considère les données suivantes :

Observation	A	B	C	D	E
X	2	3	1	2	4
Y	2	0	5	4	0
Groupe	1	2	1	1	2

1) Représenter les données dans $\mathbb{R}^2$ en tenant compte des groupes fournis.

2) Calculez les variances intra-groupes I_{intra}, inter-groupes I_{inter} et totale I_{tot} pour les données ci-dessus.

3) Montrer que la relation $I_{tot} = I_{inter} + I_{intra}$ est bien vérifiée dans ce cas.

4) Trouver une autre partition des données qui fournisse un rapport I_{inter}/I_{tot} supérieur à celui de la partition actuelle.

Exercice 2

On considère à nouveau le jeu de données précédent et on utilise la méthode de classification ascendante hiérarchique (CAH) pour regrouper les données. Dans tous les cas, on utilise la distance euclidienne comme mesure de distance entre les points.

1) Construire le dendrogramme associé aux données pour la CAH avec le lien complet comme mesure de dissimilarité.

2) Sans refaire tous les calculs de la question précédente, indiquer si les résultats de la CAH seraient différents avec le lien simple comme mesure de dissimilarité ? Justifier.

3) CAH avec critère de Ward :

a- Construire le dendrogramme associé aux données dans ce cas.

b- Calculer pour chacun des regroupements successifs le rapport I_{inter}/I_{tot}.

c- Déterminer le nombre optimal de groupes en vous appuyant sur l'évolution de I_{inter}/I_{tot}.

Exercice 3

On travaille à présent avec la méthode des centres mobiles et l'on considère encore le jeu de données de l'exercice **1**.

1) Construire la partition en 2 classes associée aux données en prenant les observations A et B comme centres initiaux.

2) Calculer le rapport I_{inter}/I_{tot} associé à cette partition.

3) Construire la partition en 3 classes associée aux données en prenant les observations A, B et C comme centres initiaux.

4) Calculer le rapport I_{inter}/I_{tot} associé à cette partition et comparer à celui de la partition précédemment obtenue.

5) Recommencer l'étude en prenant A, C et D comme centres initiaux.

Exercice 4

Cet exercice sera effectué sous le logiciel R.

On considère le jeu de données `iris` à 150 lignes et 5 colonnes. Les 5 variables considérées sont 4 mesures sur les pétales et sépales des iris, plus une variable indiquant l'espèce de ces iris. Nous allons comparer différentes classification automatiques des iris suivant les 4 variables de mesure avec la répartition des espèces.

1) Sélectionner les variables d'intérêt pour la classification et les étudier succinctement.

2) Charger le package `FactoMineR`, effectuer une ACP sur ces variables et choisir les facteurs principaux. Que peut-on remarquer ? En déduire qu'une classification automatique est particulièrement bien adaptée pour ces données.

3) Effectuer une CAH en utilisant la fonction `hclust` sur la matrice des distances D calculée sur les variables d'intérêt centrées et réduites. Choisir soigneusement le critère d'agrégation des classes parmi les 4 suivants : `ward.D2, single, complete, average`.

4) Récupérer les codes pour tracer les critères du R^2 et du PseudoF sur Moodle ou le COMMUN. Choisir le nombre de classes en justifiant ce choix par les indicateurs appropriés.

5) Créer les profils d'iris adéquats en utilisant les codes permettant de calculer les statistiques par classe à récupérer sur Moodle ou sur le COMMUN.

6) Effectuer une comparaison de la classification obtenue avec la variable expliquée *Species* en calculant les distributions conditionnelles des espèces dans chaque classe, et en comparant la répartition des classes et des espèces dans le premier plan factoriel.
On utilisera les codes suivants pour les représentations graphiques :
— *Représentation des espèces*

```
> plot(iris.acp, choix="ind", habillage = "ind", col.hab=rainbow(3)[iris$Species])
> legend("top", legend = unique(iris$Species), col = rainbow(3)[unique(iris$Specie
pch = 19, cex = 0.75, ncol = 3)
```

— *Représentation des classes*

```
> plot(iris.acp, choix="ind", habillage = "ind", col.hab=rainbow(nclass)[classes])
> legend("top", legend = unique(classes), col = rainbow(nclass)[unique(classes)],
pch = 19, cex = 0.75, ncol = nclass)
```

où `iris.acp` est la sortie de la fonction `PCA` appliquée aux variables d'intérêt, `classes` est le vecteur des classes sélectionnées à l'issue de la CAH, et `nclass` est le nombre de classes sélectionnées.

Exercice 5

Cet exercice sera effectué sous le logiciel R.

1) Effectuer une classification des iris par les centres mobiles à l'aide de la fonction `kmeans` en choisissant judicieusement le nombre de classes (cf. résultats de l'exercice **4**).

2) Pourquoi peut-on se passer de l'étape 1 de la classification mixte pour ces données ?

3) Faire une comparaison de la classification obtenue en la croisant avec la variable *Species*.

4) Visualiser les différentes classifications dans le premier plan factoriel.

5) Comparer la variable expliquée *Species* avec la classification obtenue en sélectionnant le même nombre de classes qu'il y a d'espèces différentes.

Exercice 6

Cet exercice sera effectué sous le logiciel R.

1) Effectuer une classification sur les facteurs principaux choisis à l'exercice **4)**.

2) Faire une comparaison de la classification obtenue avec la variable *Species*.

3) Reprendre les résultats des exercices précédents et les comparer avec ceux obtenus après la classification sur facteurs.

Exercice 7

Classification mixte sur les données `Employés`

Charger les données d'employés de banque contenues dans le fichier *banque.dat* disponible sur Moodle ou dans le COMMUN.

1) Effectuer une ACP sur les variables quantitatives et une ACM sur les variables qualitatives. Dans les 2 cas, sélectionner le nombre de facteurs à conserver pour la classification.

2) Construire la nouvelle table de données formée par les facteurs de l'ACP et de l'ACM sélectionnés à la question précédente.

3) Effectuer une classification par les centres mobiles sur cette nouvelle table en prenant un nombre de classes approprié (à choisir judicieusement en fonction du nombre de données) afin de construire une première partition sur laquelle sera effectuée la CAH.

4) Enregistrer les coordonnées des centres de classes finaux dans une variable locale et les afficher.

5) Effectuer une CAH sur ces nouvelles données en utilisant la méthode de Ward. Choisir une partition.

6) Obtenir la classification finale à l'aide des centres mobiles appliqués aux données d'origine, où le nombre de classes sera celui choisi grâce à la CAH.

7) Proposer des profils d'employés issus de la classification précédente.

8) Visualiser les profils obtenus sur les graphiques appropriés.

ANNEXE : Description des variables de la base de données **Employés**.

- *id* : identifiant.
- *saldeb* : salaire d'embauche.
- *temps* : ancienneté dans la banque (en jours).
- *salact* : salaire actuel.
- *sexe* : sexe.
- *exp* : expérience de la fonction (en années).
- *catemp* : catégorie d'employé.
- *satis1* : satisfait de votre emploi.
- *satis2* : satisfait de votre chef de service.
- *satis3* : satisfait de votre salaire.
- *satis4* : satisfait de vos collègues.
- *satis5* : satisfait de votre comité d'entreprise.
- *age* : age.