Année 2019-2020 Nom, Prénom :
Ecole des Mines de Nancy
Tronc commun 2A

Test Analyse de Données

7 janvier 2020 - Durée : 2h
Notes de cours et de TP autorisées
Sujet de S.Ferrigno (séances 1 à 7)

Consignes : Veuillez répondre directement aux questions sur la feuille d'énoncé dans les espaces laissés libres à cet effet. Toutes les réponses devront être justifiées.

"Les iris de Fisher" sont des données collectées par Edgar Enderson, et proposées en 1933 par le statisticien Ronald Aylmer Fisher comme données de référence pour l'analyse discriminante et la classification. Il s'agit de reconnaître le type d'iris (setosa, virginica et versicolor) à partir seulement de la longueur et de la largeur de ses pétales et sépales. Le fichier contient 50 fleurs de chaque type.

Les données se trouvent dans le data frame **iris** qui fait partie du package de base du logiciel R. Pour charger ce data frame, utiliser la commande **data(iris)**.

En mettant en oeuvre au mieux vos connaissances acquises en cours et en TD/TP et avec l'aide du logiciel R, répondez aux questions suivantes.

Partie 1 (7 points) : Etude bivariée et Analyse en composantes principales (ACP).

1. Existe-t-il un lien entre les variables Sepal.Length et Sepal.Width ? Pour répondre à cette question, vous utiliserez un test statistique et vous donnerez les hypothèses du test, la statistique de test et la valeur de la pvalue. Vous concluerez en utilisant un risque de 5%.

2. Reprendre la question 1. pour les variables Petal.Length et Petal.Width.

Lancer l'Analyse en composantes principales normée sur ce jeu de données en prenant soin de ne conserver que les variables qui se prêtent à ce type d'étude.

3. Quels sont les individus de cette étude d'ACP ? Dans quel espace sont-ils définis ?

4. Une ACP **normée** était-elle forcément nécessaire dans le contexte de cette étude ? Justifier votre réponse.

5. Pour cette étude, combien d'axes allons-nous conserver pour effectuer la projection des données de départ selon le critère de Kaiser ? selon votre "bon sens" ? Justifier votre réponse.

Dans la suite, nous travaillerons dans le plan $(1, 2)$.

6. Quelles sont les variables qui contribuent le plus à la construction de l'axe 1 ? de l'axe 2 ? Justifier votre réponse.

7. Quelles sont les variables qui sont bien représentées dans le plan $(1, 2)$? Justifier votre réponse.

8. Parmi les 150 iris, donner un exemple d'un iris qui est mieux représenté dans le plan $(1, 3)$ que dans le plan $(1, 2)$. Vous justifierez votre réponse.

Partie 2 (8 points) : Classification Automatique Hiérarchique (CAH).

Lancer à présent une classification hiérarchique ascendante sur le jeu de données iris en ne prenant pas en compte, comme pour la Partie 1, la variable Species et en utilisant la méthode de Ward. Dans cette partie, nous travaillerons avec la distance Euclidienne qui est la distance utilisée par défaut dans les diverses fonctions R dont vous aurez à vous servir.

1. Quelle est la distance entre les iris 6 et 3 ?

2. Donner le code R qui vous a permis d'effectuer cette classification à partir de la matrice des distances.

3. Combien de classes choisiriez-vous à partir du dendrogramme ? Justifier votre réponse.

   Dans la suite nous travaillerons avec 3 classes.

4. Parmi les 150 iris de l'étude, combien sont affiliés à la classe 1, la classe 2, la classe 3 ?

5. A quelle classe appartient l'iris numéro 70 ?

6. Quelle est la moyenne de la variable Sepal.Length dans chacune des trois classes ?

Lancer à présent une classification hiérarchique ascendante sur l'ensemble des composantes principales obtenues en effectuant l'ACP sur le jeu de données iris (sans la variable Species).

7. Combien de classes suggère le dendrogramme ? Justifier votre réponse.

8. Pour chacune des 3 classes obtenues, donner le numéro de l'iris le plus représentatif de la classe. Justifier votre réponse.

9. Pour chacune des 3 classes obtenues, donner le numéro de l'iris le plus éloigné des centres de gravité des deux autres classes. Justifier votre réponse.

Partie 3 (5 points) : Analyse discriminante.

Dans cette partie, nous travaillerons avec le jeu de données iris de départ (avec cette fois l'ensemble des variables). L'idée est d'utiliser les quatres variables quantitatives du jeu de données pour retrouver les espèces d'iris (variable Species).

1. Classer les quatre variables quantitatives de la plus discriminante à la moins discriminante. Justifier votre réponse.

2. Réaliser un test de Wilks sur ces données. Vous préciserez les hypothèses du test, la valeur de la statistique de test, la pvalue et vous concluerez en prenant un risque de 5%.

3. Réaliser une analyse discriminante avec l'ensemble des variables du jeu de données iris. Combien d'axes discriminants sont retenus ? Quelle est la proportion de variance récupérée par chaque axe ?

4. Si on applique le modèle de la question précédente aux données de départ, quel est la taux d'erreur ?

5. Prédire à quelle espèce (et avec quelle probabilité) appartient un iris ayant les caractéristiques suivantes $(6.3, 3, 5.2, 1.8)$ dans l'ordre d'apparition des variables dans le jeu de données initial ?