Université Toulouse 1 Capitole *2018/2019*

DU de Statistique Appliquée et année préparatoire au M2 Statistique et Econométrie en FOAD

Examen de Data Mining 1

Sujet d'examen à réaliser avec le **logiciel R** *et à déposer sous Moodle au plus tard le dimanche 23 juin 2019 à 23h55*

Recommandations : Votre rapport doit être de **30 pages** au maximum. Insérer le code R (pas de code SAS) à la fin du rapport. Vous devez insérer les sorties R (tableaux et graphiques) que vous commentez dans le corps du rapport (pas d'annexes).

Exercice 1

Vous disposez de données concernant 57 médecins de la région Midi-Pyrénées en 1999 (fichier **medecins.txt**). Les 7 variables considérées sont les suivantes :

- ANCINS99 : expérience du médecin (en nombre de mois),
- MTH70 : pourcentage de la clientèle du médecin âgée de plus de 70 ans,
- SHAREFRE : pourcentage de la clientèle du médecin ne payant pas les frais médicaux,
- AGE : âge du médecin (en années),
- HONPAT : honoraire moyen par patient (en francs),
- CONSUPPA : nombre moyen de consultations par patient,
- VISITSHA : proportion de visites à domiciles,

1. Réaliser une classification des médecins par agrégation autour des moyennes mobiles (AMM) à l'aide des 7 variables quantitatives et commenter.
2. Réaliser une classification des médecins sur les composantes principales retenues par l'analyse en composantes principales (ACP) à l'aide d'une classification ascendante hiérarchique (CAH). Commenter.

Exercice 2

Vous disposez de données (fichier **banque.txt**) concernant 812 clients d'un établissement bancaire de la région toulousaine en 2014. Les variables d'intérêt sont les suivantes :

- NBRDEB : nombre de mouvements débiteurs sur le compte durant le mois de juin 2014,
- MTDEB : montant des mouvements débiteurs sur le compte durant le mois de juin 2014,
- SOLDCC : solde du compte courant au 30 juin 2014 en euros,
- CARTE : 1 si le client possède une carte Visa Premier, 0 sinon,
- ANCREL : ancienneté de la relation du client avec la banque en années,
- EPAR : solde de l'épargne du client au 30 juin 2014 en euros.

L'objectif est d'expliquer le fait de posséder une carte Visa Premier à l'aide des variables disponibles. On cherche donc à discriminer les détenteurs d'une carte bancaire Visa Premier et ceux qui n'en possèdent pas, à l'aide d'une analyse factorielle discriminante (AFD).

1. Donner le pouvoir discriminant des variables initiales. Illustrer par des graphiques adaptés.
2. Tirer un échantillon d'apprentissage (80% de l'échantillon total) à l'aide de la fonction `sample()`.
3. A l'aide de l'échantillon d'apprentissage, discriminer les 2 groupes de clients à partir des 5 autres variables à l'aide d'une AFD et répondre aux questions suivantes :
    a. Interpréter l'axe discriminant.
    b. Commenter la qualité de la discrimination (indicateur numérique et test d'hypothèse).
    c. Donner des graphiques représentant les 2 groupes de clients sur l'axe discriminant et commenter les groupes à l'aide de l'axe discriminant.
    d. Prédire le groupe d'affectation et calculer le taux de clients mal classés sur l'échantillon d'apprentissage.
4. Prédire le groupe d'affectation sur l'échantillon test (complémentaire de l'échantillon d'apprentissage) et calculer le taux de clients mal classés sur cet échantillon.