Analyse de données multidimensionnelles
M1 Statistique et économétrie, 2013

Mini projet (préparation du contrôle)

V. Monbet

Cette étude porte sur les données "Loisirs" (`loisirs.csv`). Ce fichier contient des données (de l'insee) issues d'une enquête réalisée en 2003 auprès de n = 8403 individus sur leurs activités-loisirs. 18 questions concernent leurs activités au cours des 12 derniers mois. Les réponses à ces questions sont oui ou non.

- Lecture : Avez-vous pratiqué une activité de lecture?
- Voyage : Avez-vous réalisé des voyages?
- Joue musik : Avez-vous une activité artistique (musique, peinture, danse, théâtre, écriture, photo...)?
- Collection : Avez-vous pratiqué une activité de collection?
- Bénévolat : Avez-vous pratiqué une activité de bénévolat?
- Mécanique : Avez-vous effectué des travaux de mécanique, de bricolage, de décoration?
- Jardinage : Avez-vous pratiqué une activité de jardinage?
- Tricot : Avez-vous pratiqué une activité de tricot, de la broderie, de la couture?
- Cuisine : Avez-vous fait la cuisine pour le plaisir?
- Pêche : Avez-vous été à la pêche, à la chasse?
- Ecoute musik : Avez-vous écoutez de la musique?
- Cinéma : Avez-vous été au cinéma?
- Spectacle : Avez-vous été à des spectacles (théâtre, concert, danse, cirque...)?
- Exposition : Avez-vous été à des expositions, visité des musées des monuments?
- Télé : Avez-vous regardé la télévision?
- Ordinateur : Avez-vous utilisé un ordinateur ou une console de jeu?
- Sport : Avez-vous pratiqué du sport?

- Marche : Avez-vous réalisé des marches ou randonnées?

Nous disposons aussi d'information de type signalétique :

- CSP : Manoeuvre ou ouvrier spécialisé ; Ouvrier quali

  fié ou hautement quali

  fié, technicien(ne) d'ate- lier ; Technicien(ne) (non cadre) ; Agent de maîtrise, maîtrise administrative ou commerciale, VRP (non cadre), personnel de catégorie B de la fonction publique ; Ingénieur ou cadre, personnel de catégorie A de la fonction publique ; Employé de bureau, employé de commerce, personnel de ser- vice, personnel de catégorie C; autre cas

- Age

- Sexe

Nous disposons enfi

n d'information plus précises sur le type de sports qu'ils pratiquent (natation ; vélo ; marche ; gym ; foot ; sportco ; tennis ; sportindiv). Le but de l'étude est de caractériser, de décrire la population de cette enquête, de faire une typologie des enquêtés par rapport à leur loisirs.

1 Analyse factorielle

1.1 Analyse des correspondances multiples

1. Faire une analyse descriptive multivariée du jeu de données. Préparez vous à commentez les résultats obtenus. Comment interprétez-vous la position des modalités "0" des types de sports?

2. Pour mieux connaitre la population, on peut s'intéresser exclusivement à une partie du jeu de données. Par exemple, on pourra étudier

   - les dépendances entre les variables démographiques et en tirer les principales tendances,
   - puis les dépendances entre les variables de sport pratiqué,
   - et le lien entre ces deux sous ensembles, etc.

   Réfléchir à/commenter la proximité ou l'opposition des différentes modalités.
   Identifier et caractériser les individus "extrêmes".

1.2 Analyse en composantes principales

Il est intéressant de comparer les résultat de l'AFCM avec ceux d'une ACP. Cette dernière est possible uniquement si on est capable de définir des distantes (euclidiennes) pertinentes entre les modalités d'une même variable. Proposez une représentation des variables qui conduise à des distances euclidiennes interprétables. Et conduire l'ACP sur la table obtenue via cette représentation.
Aide : on pourra par exemple conserver les variables binaires, recoder la variable age de 1 à 9 et recoder la variable CSP en variables binaires. Expliquer pourquoi ce recadoge est pertinent en terme de distance euclidienne.

2 Classification

L'objectif est d'obtenir des groupes homogènes et bien distinct des individus de la base de données. On va comparer plusieurs approches.

1. Réaliser une classification sur les composantes de l'analyse des correspondances multiples. Justifier ce choix.

2. Réaliser une classification directement sur les données d'origine en choisissant un codage des données et une distance pertinentes.

Dans les deux cas vous expliquerez votre démarche (choix de la (ou des) méthode(s), différentes étapes), justifierez le nombre de classes retenues et interpréterez les résultats.

3 Analyse discriminante

Choisir une des variables "activités" et proposer un modèle pour la prédire à partir des variables démographiques et signalitiques. Estimer l'erreur de classement et le risque de Bayes. Peut-on interpréter le modèle?
Peut-on améliorer significativement le modèle en utilisant aussi les autres variables d'activité? Détailler votre démarche, jsutifiez vos choix.