Note de lecture rédigée par Jean-Jacques Droesbeke[1]

EXPLORATION DE DONNEES ET METHODES STATISTIQUES

Data analysis et Data mining avec le logiciel R

Lise BELLANGER et Richard TOMASSONE

Livre (479 pages)
Édition : Ellipses – 2014

Ce livre est né de l'expérience d'enseignant de statistique appliquée et de chercheur accumulée par les auteurs. Il est construit sur cinq axes.

La première partie traite des préalables à un traitement statistique. Après un chapitre introductif explicitant les caractéristiques d'une démarche scientifique, deux chapitres en constituent la structure. Les outils de représentation d'un ensemble de données sont tout d'abord présentés : structure d'un tableau de données, résumés unidimensionnels et multidimensionnels, décomposition d'une matrice de données. Viennent ensuite un certain nombre de pratiques utiles avant leur traitement : transformations de Box et Cox, ré-échantillonnage des données, données suspectes et manquantes.

La deuxième partie porte sur l'étude d'un échantillon. On y trouve les principes et interprétations de l'analyse en composantes principales, de l'analyse factorielle des correspondances et des correspondances multiples, le modèle factoriel et les méthodes de classification.

La troisième partie traite de l'étude de deux groupes de variables. Les auteurs décrivent tout d'abord les bases du modèle de régression et ses limites. La colinéarité fait l'objet d'un chapitre dans lequel la régression sur composantes principales, la régression PLS et la régression biaisée ou pénalisée sont décrites. Cette partie s'achève avec les relations entre deux groupes de variables : l'analyse des corrélations canoniques est au centre de ce chapitre mais d'autres méthodes sont aussi signalées.

La quatrième partie porte sur l'étude de plusieurs échantillons sur le thème « discrimination et classement ». Enfin, les arbres binaires et quelques conclusions et perspectives générales complètent cet ouvrage qui s'achève par une annexe contenant quelques renseignements sur les bibliothèques de programmes *R* utilisés dans l'ouvrage ainsi que sur les fichiers de données traités.

Les auteurs s'adressent à des lecteurs ayant déjà une connaissance de base en inférence statistique (estimation et tests d'hypothèses). Ils sont aussi censés posséder une pratique de l'algèbre linéaire et être familiers avec le logiciel *R*. Chaque méthode présentée est

[1] Université libre de Bruxelles, jjdroesb@ulb.ac.be

accompagnée d'au moins un exemple dont les données sont accessibles, avec une description succincte, sur la page personnelle de l'un des auteurs.

L'ouvrage est très clairement écrit. Les exemples sont attractifs et utilement commentés. Les auteurs ont introduit quelques illustrations et portraits de quelques personnages qui ont participé à la construction des outils statistiques utilisés. Si le lecteur possède les caractéristiques nécessaires pour aborder cet ouvrage, il y trouvera certainement une aide précieuse pour aborder ses propres données de recherche et la manière d'interpréter leurs analyses. Par ailleurs, les enseignants en quête d'exemples divers traités avec *R* puiseront aussi dans cet ouvrage de quoi alimenter judicieusement certaines parties de leurs cours.