UCL Institut de statistique

STAT

STAT2411A Analyse multivariée des données.

[15h+7.5h exercices]

Langue d'enseignement : français
Niveau : cours de 2ème cycle

Objectifs (en terme de compétences)

Objectifs Généraux.
Présenter les techniques modernes de l'analyse de grands ensembles de données et développer les outils de base du " data mining ".
Objectifs Spécifiques.
A l'issue de ce cours, les étudiants doivent être capables de :
Traiter et décrire l'information contenue dans des grands ensembles de données ;
Comprendre les mécanismes qui justifient l'emploi de telle ou telle méthode ;
Interpréter correctement les graphiques et résultats fournis par les logiciels ;
Résoudre des problèmes avec données réelles.

Objet de l'activité (principaux thèmes à aborder)

Contenu
- Rappels d'algèbre et de géométrie utiles à l'analyse des données
- Principes de base des méthodes factorielles
- Analyse en composantes principales et ses variations
- Analyse des corrélations canoniques
- Analyse factorielle discriminante
- Introduction aux méthodes de classification
- L'analyse des données, en pratique.

Résumé : Contenu et Méthodes

Contenu
- Rappels d'algèbre et de géométrie utiles à l'analyse des données
- Principes de base des méthodes factorielles
- Analyse en composantes principales et ses variations
- Analyse des corrélations canoniques
- Analyse factorielle discriminante
- Introduction aux méthodes de classification
- L'analyse des données, en pratique.

Autres informations (Pré-requis, Evaluation, Support, ...)

Prérequis:
L'étudiant doit être capable de
manipuler et lire les expressions algébriques (calcul matriciel) ;
comprendre et dominer les éléments de base de l'analyse statistique.
L'évaluation se fait :
1) par un travail sur données réelles selon les modalités précisées ci-dessous. Il s'agit de mettre en #uvre certaines des méthodes vues au cours dans un domaine d'application choisi par l'étudiant. Pour permettre aux étudiants de réaliser ce travail dans les meilleures conditions, le cours magistral sera concentré sur 10 semaines. Les étudiant travaillent, en principe, par paire. L'assistant du cours encadrera les étudiants pour ce travail (mise au courant du logiciel). Ce travail devrait prendre environ 12 heures de travail PAR étudiant (soit 24 h. pour la paire).
2) Par un examen écrit à livre fermé : il s'agira ici de voir si l'étudiant maîtrise les concepts abordés au cours, s'il comprend les méthodes utilisées (questions d'ordre général mais aussi commentaires sur des expressions matricielles importantes) et s'il peut interpréter correctement des résultats obtenus par les logiciels (du type de ceux présentés dans le syllabus).
Modalités du projet:
Pour ceux qui le désirent, deux (ou trois) séances d'initiation à SPADN seront organisées par l'assistant du cours selon un horaire à préciser.
L'assistant encadrera également les étudiants pour le projet. Attention : il s'agit uniquement des aides pour l'utilisation du logiciel ou donner quelques conseils ponctuels d'ordre général. Ce projet reste VOTRE projet.
Ce projet est un travail sur données réelles. Il s'agit de mettre en oeuvre certaines des méthodes vues au cours dans un domaine d'application choisi par l'étudiant. Il faut que ce projet contienne au moins une ACP et une AFC (simple ou multiple). Si possible, le même ensemble de données sera analysé par ces deux types de méthodes (l'AFCM est possible sur la plupart des ensembles de données). Souvent, une analyse de classification apporte un regard complémentaire utile sur les données analysées (confirmation ou non de goupes d'individus similaires, d'outliers,#). Le cas échéant, il est toujours utile de décrire les caractéristiques des différents " clusters " obtenus.
Le projet fera l'objet d'un bref rapport présentant de façon claire et concise:
1 l'objet de l'analyse
2 la description des données (unités utilisées, etc...)
3 l'analyse proprement dite
4 les commentaires sur les résultats obtenus.
Ce rapport ne devrait pas dépasser 7 à 10 pages (des résultats peuvent être mis en annexe). Le projet sera jugé selon les critères suivants:
1 Adéquation des méthodes utilisées aux données et problème étudiés.
2 Originalité et intérêt du problème.
3 Richesse des analyses proposées (au delà du minimum requis).
4 Justesse des commentaires sur les résultats.
5 Qualité de la présentation du rapport.
Support:
Syllabus de L.SIMAR (2004) : " Multivariate Data Analysis ", 256 pages, Institut de Statistique, UCL.
Ce manuel est disponible à la DUC.

Autres crédits de l'activité dans les programmes

BIR22/0A	Deuxième année du programme conduisant au grade de bio-ingénieur: Sciences agronomiques (Technologies et gestion de l'information)	Obligatoire
BIR22/0C	Deuxième année du programme conduisant au grade de bio-ingénieur: Chimie et bio-industries (Technologies & gestion de l'information)	Obligatoire
BIR22/0E	Deuxième année du programme conduisant au grade de bio-ingénieur: Sciences et technologies de l'environnement (Technologies et gestion de l'information)	Obligatoire
BIR22/4E	Deuxième année du programme conduisant au grade de bio-ingénieur : Sciences et technologie de l'environnement (Technologies environnementales: eau, sol, air)	Obligatoire
BIR22/5E	Deuxième année du programme conduisant au grade de bio-ingénieur : Sciences et technologie de l'environnement (Aménagement du territoire)	Obligatoire
BIR22/6E	Deuxième année du programme conduisant au grade de bio-ingénieur : Sciences et technologie de l'environnement (Nature, eau & forets)	Obligatoire
BIR22/7E	Deuxième année du programme conduisant au grade de bio-ingénieur : Sciences et technologie de l'environnement (Ressources en eau et en sol)	Obligatoire