UCL Institut de statistique

STAT

STAT2411 Analyse des données

[22.5h+7.5h exercices] 5 crédits

Cette activité se déroule pendant le 1er semestre

Enseignant(s): Léopold Simar
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 ensemble 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 ensemble 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)

- 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.
- Analys edes corrélations canoniques.
- Analyse factorielle discriminante.
- Analyse factorielle des correspondances.
- Introduction aux méthodes de classification.
- L'analyse des données, en pratique.

Résumé : Contenu et Méthodes

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

Méthodes
Le cours comprend des exposés magistraux et un travail sur ordinateur à faire individuellement.

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

Pré-requis:
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.

Evaluation
L'évaluation se fait :
1) par un travail sur données réelles selon les modalités qui seront précisées ci-dessous. Il s'agit de mettre en oeuvre 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 sur10 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, Institu de Statistqiue, UCL.
Ce manuel est disponible à la DUC.
Encadrement
Titulaire : Léopold Simar, tél : 010/47 43 08, simar@stat.ucl.ac.be
Ouvrages de référence
Lebart, L., Morineau, A. et J.P. Fenelon (1982) : Traitement des données statistiques. Dunod, Paris.
Saporta, G. (1990) : Probabilités, analyse des données et statistiques. Ed. Tecnip, Paris.
Romeder, J.M. (1973) : Méthodes et programmes d'analyse discriminante. Dunod, Paris
Pour plus d'informations :
http://www.stat.ucl.ac.be/cours/stat2411/index.html http://www.stat.ucl.ac.be/cours/stat2411/index.html

Autres crédits de l'activité dans les programmes

ACTU21MS	Première année du master en sciences actuarielles, à finalité spécialisée	(5 crédits)	
ELME23/M	Troisième année du programme conduisant au grade d'ingénieur civil électro-mécanicien (mécatronique)	(5 crédits)	
ESP3DS/EP	Diplôme d'études spécialisées en santé publique (recherche clinique)	(5 crédits)	
MAP23	Troisième année du programme conduisant au grade d'ingénieur civil en mathématiques appliquées	(3 crédits)	
MATH21/S	Première licence en sciences mathématiques (Statistique)	(3 crédits)	Obligatoire
MATH22/G	Deuxième licence en sciences mathématiques	(3 crédits)	
STAT2MS	Master en statistique, orientation générale, à finalité spécialisée	(5 crédits)	
STAT3DA/B	diplôme d'études approfondies en statistique (biostatistique et épidémiologie)	(5 crédits)	
STAT3DA/E	diplôme d'études approfondies en statistique (statistique et économétrie)	(5 crédits)	
STAT3DA/P	diplôme d'études approfondies en statistique (pratique de la statistique)	(5 crédits)	