UCL Institut de statistique

STAT

STAT2411 Data Analysis

[22.5h+7.5h exercises] 5 credits

This course is taught in the 1st semester

Teacher(s): Léopold Simar
Language: french
Level: 2nd cycle course

Aims

General objectives.
Presentation of the modern techniques for the analysis of huge multivariate data sets. Developing the basic tools for " data mining ".
Specific objectives.
At the end of this course, the students should be able to:
- Manipulate and describe the information contained in huge data sets;
- Understand why such or such method is appropriate;
- Give a correct interpretation of the resulting pictures and of the output of the software;
- Solve problems with real data sets.

Main themes

Contents:
- Reminders of algebra and geometry useful for multivariate data analysis
- Basic principles of factorial methods
- Principal components analysis (PCA)
- Canonical correlation
- Factorial discriminant analysis (FDA)
- Factorial correspondence analysis (FCA simple and multiple)
- Cluster analysis
- Data analysis in practice

Content and teaching methods

Contents:
- Reminders of algebra and geometry useful for multivariate data analysis
- Basic principles of factorial methods
- Principal components analysis (PCA)
- Canonical correlation
- Factorial discriminant analysis (FDA)
- Factorial correspondence analysis (FCA simple and multiple)
- Cluster analysis
- Data analysis in practice

Other information (prerequisite, evaluation (assessment methods), course materials recommended readings, ...)

Prerequisite:
The student should be able to:
Manipulate and read algebraic expression (matrix calculus);
Dominate the basic tools of statistical analysis.
Evaluation: two parts:
1) A project with real data (see details below). The idea is to apply the methods of the course in a real problem chosen by the student. The course is concentrated on 10 weeks to allow the students to do this project in the best conditions. Students work by groups of two students. The teaching assistant will help the students for the software problems. This work would represent 12 hours of work per student (24 h. for the team of two).
2) Written exam, with closed book. The idea is to see if the student masters all the techniques developed in the course (understanding of the techniques) but also if he is able to comment output from a software (like these presented in the manual).
Details about project:
For those who want, 2 or 3 meetings will be organized by the teaching assistant to initiate the students to the software SPADN. The teaching assistant will also help the students for their project but only for software's issues.
This project is a work on real data. The idea is to apply the techniques of the course to analyze a problem in a field chosen by the student. The project should at least contain a PCA and a FCA. It would be better if the chosen data set allows for both approaches (remember that multiple FCA is possible for most of the data sets). Often, some cluster analysis shed new light on the data that have been analyzed (detection of outliers, structure of different groups of individuals, etc#). In case, a statistical description of the obtained groups is useful.
The project will be presented in a short report, summarizing:
1 The object of the analysis
2 Description of the data (units, etc.)
3 The analysis
4 Comments on the obtained results and conclusions.
The report should not be longer than 7-10 pages (some details may be given in appendices). The evaluation criterion will be based on:
1 Appropriateness of the chosen methods.
2 Originality and interest of the chosen problem.
3 Deepness of the analysis (more than the minimum required).
4 Correctness of the comments.
5 Quality of the presentation of the report.
Manual.
L.SIMAR (2004) : " Multivariate Data Analysis ", 256 pages, Institut de Statistique, UCL.
This manual is available at the DUC (students' bookstore).
Professor : Léopold Simar, tél : 010/47 43 08, simar@stat.ucl.ac.be
References :
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
For more information:
http://www.stat.ucl.ac.be/cours/stat2411/index.html http://www.stat.ucl.ac.be/cours/stat2411/index.html

Other credits in programs

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