Universitat Autònoma de Barcelona

2020/2021

Unsupervised Learning

Code: 104869
ECTS Credits: 6

Degree	Type	Year	Semester
2503852 Applied Statistics	OB	2	2

The proposed teaching and assessment methodology that appear in the guide may be subject to changes as a result of the restrictions to face-to-face class attendance imposed by the health authorities.

Contact

Name: Mercè Farré Cervelló

Email: Merce.Farre@uab.cat

Use of Languages

Principal working language: catalan (cat)

Some groups entirely in English: No

Some groups entirely in Catalan: Yes

Some groups entirely in Spanish: No

Prerequisites

A previous course of Linear Algebra is essential, as well as courses in Probability and Statistical Inference. Also, a good knowledge of the R software is assumed.

Objectives and Contextualisation

Most of collected data sets are multivariate, that is, for the same experimental unit, perhaps a complex nature object, we observe simultaneously the values of several variables. Multivariate Analysis deals with the methods that are most appropriate for describing, exploring and modelling vector data, as well as for applying statistical inference. The interest in processing large amounts of observations in many variables of a diverse nature, together with the aim of reducing the information that is not relevant or discovering patterns of association between variables or between cases, they have recently promoted the development of a series of multivariate techniques. This subject is intended as a first contact of the student with the statistical learning theory. Students must understand the power and applicability as well as the limitations of the multivariate tools, some of which are based on very simple heuristic ideas. The subject focuses in the applications, mostly in the computer work sessions using the R free software resources. Theoretical and problems sessions are devoted to formalize the models, derive their properties, and study some models validation techniques.

Competences

- Analyse data using statistical methods and techniques, working with data of different types.
- Critically and rigorously assess one's own work as well as that of others.
- Make efficient use of the literature and digital resources to obtain information.
- Select and apply the most suitable procedures for statistical modelling and analysis of complex data.
- Select the sources and techniques for acquiring and managing data for statistical processing purposes.
- Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
- Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
- Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.

- Students must develop the necessary learning skills to undertake further training with a high degree of autonomy.
- Summarise and discover behaviour patterns in data exploration.
- Use quality criteria to critically assess the work done.
- Work cooperatively in a multidisciplinary context, respecting the roles of the different members of the team.

Learning Outcomes

1. Analyse data using an automatic learning methodology.
2. Characterise homogeneous groups of individuals through multivariate analysis.
3. Critically assess the work done on the basis of quality criteria.
4. Describe the advantages and disadvantages of algorithmic methods compared to the conventional methods of statistical inference.
5. Identify the statistical assumptions associated with each advanced procedure.
6. Identify, use and interpret the criteria for evaluating compliance with the requisites for applying each advanced procedure.
7. Make effective use of references and electronic resources to obtain information.
8. Obtain and manage complex databases for subsequent analysis.
9. Reappraise one's own ideas and those of others through rigorous, critical reflection.
10. Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
11. Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
12. Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
13. Students must develop the necessary learning skills to undertake further training with a high degree of autonomy.
14. Use summary graphs of multivariate or more complex data.
15. Work cooperatively in a multidisciplinary context, accepting and respecting the roles of the different team members.

Content

Statistical learning and dimension reduction

- Supervised and unsupervised learning. Multivariate methods. Examples.
- Random vectors. Expectation vector and covariance-correlation matrices. Properties.
- Multivariate data. Sample expectation and covariance-correlation matrices. Maximum likelihood estimation in the Gaussian case.
- Spectral decomposition (SD) and singular value decomposition (SVD).
- Maximizing quadratic forms: The fundamental theorem.

Factorial methods I: Principal components analysis (PCA)

- Introduction to PCA. Definition of components. The fundamental result.
- Criteria for deciding on the number of components: The principal components.
- Variables and individuals plots. Standardizations.
- Row and column analysis of the eigenvectors matrix and other related matrices.
- A geometric point of view of the principal components.

Factorial methods II: Factorial analysis (FA)

- The factorial model. Communalities and specificities.
- The covariance matrix decomposition theorem.
- Discussing the existence and uniqueness of the factorial model. Rotations.
- Parameters estimation methods. Factorial scores estimation or prediction.
- Interpreting the results. Comparing PCA and FA.

Factorial methods III: Multidimensional scaling (MDS) and correspondence analysis (CA)

- Objectives and methods.
- Classic and metric multidimensional scaling.
- Non-metric multidimensional scaling.
- Distances, proximities and dissimilarities.
- Categorical data: Chi-square distance and others.
- Correspondence analysis as a MDS method.
- Profiles and inertias. Decomposing inertia.
- Graphical representation and interpretation of results in CA.

Cluster analysis (CLA)

- Comparing different approaches. Examples.
- Results' analysis and validation.
- Hierarchical clustering: Link functions.
- Centroid based methods: The k-means algorithm.
- Model based methods: Expectation and maximization (EM).

Multivariate inference basics

- The likelihood ratio test.
- Tests for mean vectors.
- Tests for covariance matrices. ANOVA and MANOVA.

Discriminant analysis (DA) and other supervised methods

- Objectives and criteria of discriminant analysis.
- Discriminant analysis in Gaussian models.
- Fisher's linear discriminant analysis.
- Partial least squares regression (PLS) and other methods.

Methodology

The theoretical sessions, where the multivariate methods will be exposed in detail and discussed on the bases of appropriate examples. The classroom presentations will be posted on the virtual campus. The revision and expansion of contents using the course bibliography will be encouraged.

The computer lab sessions are designed to be implemented in statistical software R. The exercises statements and other auxiliary material will be made available to the students in the Virtual Campus. Extension exercises will be proposed to be solved autonomously.

The theoretical sessions, where the multivariate methods will be exposed in detail and discussed on the bases of appropriate examples. The classroom presentations will be posted on the virtual campus. The revision and expansion of contents using the course bibliography will be encouraged.

The computer lab sessions are designed to be implemented in statistical software R. The exercises statements and other auxiliary material will be made available to the students in the Virtual Campus. Extension exercises will be proposed to be solved autonomously.

The collaboration and participation of all students will be sought, without discrimination based on sex or any other cause.

Activities

Title	Hours	ECTS	Learning Outcomes

Type: Directed			
Computer lab sessions	26	1.04	1, 4, 14, 5, 6, 8, 10, 11, 15, 7
Theoretical classes	26	1.04	1, 9, 3, 2, 4, 14, 5, 6, 8
Type: Autonomous			
Personal work	42	1.68	9, 4, 5, 6, 13, 7
Tasks solving and delivery	44	1.76	1, 2, 4, 14, 8, 13, 12, 10, 11, 15, 7

Assessment

The course grade (NC) will be calculated on the basis of the delivered tasks and the marks in two partial exams (P1 and P2), including both theoretical and computational exercices:

$NC = 0.4 \cdot P1 + 0.5 \cdot P2 + 0.10 \cdot Lli$

where P1 and P2 correspond to the first and second partial grades, respectively, and Lli is based on the delivered tasks and will not be recoverable.

In order to succeed in this course, it is mandatory that $NC>=5$ and $P1>3.5$ and $P2>3.5$. Besides that, the students will have the option of taking an additional recovery exam (F) with the same format (theoretical and computational questions). The final qualification will be:

$NF = Max (NC, 0.90 \cdot F + 0.10 \cdot Lli)$

Observation: Only students who have participated in 2/3 of the continuous assessment activities will have the recovery option. Honor grades will be granted at the first complete evaluation. Once given, they will no be withdrawn even if another student obtains a larger grade after consideration of the final exam.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Partial exam 1	0,4	5	0.2	1, 14, 5, 6, 13, 12
Partial exam 2	0,5	5	0.2	1, 2, 4, 5, 6, 8, 13, 12, 10, 7
Tasks delivery	0,1	2	0.08	9, 3, 8, 13, 12, 10, 11, 15, 7

Bibliography

Everitt, B., Hothorn, T. ; An introduction to Applied Multivariate Analysis with R. Springer, 2011.

Härdle, W., Simar, L.; Applied Multivariate Statistical Analysis. Springer,2007.

Peña, D.; Análisis de datos multivariantes. McGraw Hill, 2002.

Rencher, A., Christensen, W.; Methods of Multivariate Analysis. Wiley Series in Probability and Mathematical Statistics, 2012.

Complementary references

Coghlan, A.; Little book of R for Multivariate Analysis.

https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/

Cuadras, C.; Nuevos Métodos de Análisis Multivariante (web), 2014.

Greenacre, M.; La pràctica del análisis de correspondencias. Fundacion BBA, 2003.

James, G., Witten, D., Hastie, T., Tibshirani, R.; An Introduction to Statistical Learning. Springer, 2014.

Mardia, K.V, Kent, J.T., Bibby, J.M.; Multivariate Analysis. Academic Press, 2003.

Rencher, A.; Multivariate Statistical Inference and Applications. John Wiley &Sons, 1998.