UAB Universitat Autònoma de Barcelona

Unsupervised Learning

Code: 104869
ECTS Credits: 6

2024/2025

Degree	Type	Year
2503852 Applied Statistics	OB	3

Contact

Name: Maria Merce Farre Cervello

Email: merce.farre@uab.cat

Teaching groups languages

You can view this information at the end of this document.

Prerequisites

A previous course in linear algebra is essential, as well as courses in probability, multidimensional dustributions and statistical inference. It is also assumed that you know how to use the R language with agility.

Objectives and Contextualisation

The need to process a large amount of data with many variables of a diverse nature, while reducing information that is not relevant and discovering patterns of association between variables and/or cases, have led to the development of a large number of procedures that are us they place in the multivariate scenario. Unsupervised Learning deals with the methods that are closest to describing, exploring and modeling vector data. The subject is designed as the student's first contact with the world of so-called "statistical learning", so that he understands the power and applicability, and at the same time the limitations, of the methods, some of the which are based on rather intuitive heuristic ideas. Most of the methods worked on in the course are unsupervised, that is to say, there is no set of cases with known answers that allow the method to be evaluated. The approach of the subject is eminently applied in terms of working with data using the potential of the free software R, accompanied by the appropriate rigor and generality in the definition of theoretical models and the corresponding methods of analysis and validation of the results.

Learning Outcomes

1. CM11 (Competence) Create new machine learning models, running experiments to demonstrate their feasibility and improved performance compared to the state of the art.
2. CM12 (Competence) Assess the existence of inequalities on the grounds of gender in databases, to avoid bias in automatic (algorithmic) decision-making.
3. KM16 (Knowledge) Recognise supervised and unsupervised, profound and generic machine learning models, fostering innovation in the field of statistics.
4. KM16 (Knowledge) Recognise supervised and unsupervised, profound and generic machine learning models, fostering innovation in the field of statistics.

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 under constrains: 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 (CA) as a MDS method.
- Profiles and inertias. Decomposing inertia.
- Graphical representation and interpretation of results in CA.

Cluster analysis (CLA)

- Comparing different approaches. Examples.
- Analyzing and validating the clusters.
- 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)

- Objectives and criteria of discriminant analysis.
- Discriminant analysis in Gaussian models.
- Fisher's linear discriminant analysis.

Activities and Methodology

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Computer lab sessions	26	1.04	
Theoretical classes	26	1.04	
Type: Autonomous			
Personal work	42	1.68	
Tasks solving and delivery	44	1.76	

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.

Annotation: Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Assessment

Continous Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Partial exam 1 (theory & comput)	0,35	4	0.16	CM11, KM16
Partial exam 2 (theor & comput)	0,45	4	0.16	CM11, KM16
Tasks delivery	0,2	4	0.16	CM11, CM12, KM16

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• P1 + 0.5 • P2 + 0.10 • 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 • F + 0.10• 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.

Single assessment

The single assessment will be a synthesis test of the skills of the two partials, based on: (1) An exam with theory and practical questions (weight: 50%). (2) A practice test in front of the computer (weight: 40%). (3) The delivery of the scheduled tasks that are indicated, with the possibility of the professor asking the student to explain details of these deliveries (weight: 10%).

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.

Wehrens, R. (2020). Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences. Heidelberg: Springer. https://link-springer-com.are.uab.cat/book/10.1007/978-3-662-62027-4

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.

Software

R and RStudio.

Language list

Information on the teaching languages can be checked on the CONTENTS section of the guide.