UAB
Universitat Autònoma de Barcelona

Mathematics for Big Data

Code: 45562
ECTS Credits: 6

2025/2026

Degree	Type	Year
Modelización para la Ciencia y la Ingeniería / Modelling for Science and Engineering	OP	1

Contact

Name: Amanda Fernandez Fontelo

Email: amanda.fernandez@uab.cat

Teachers

Sundus Zafar

Teaching groups languages

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

Prerequisites

Students should have basic knowledge of linear algebra, probability, statistical inference, and linear models. Experience using R and Python is also highly recommended.

Objectives and Contextualisation

Huge amounts of data are generated today across a wide range of fields, including health, engineering, the social sciences, economics, etc. While this exponential growth of data presents significant challenges, it also creates opportunities to extract relevant information and facilitate evidence-based decision-making, process optimisation, and the generation of new knowledge. This course aims to provide students with the mathematical, statistical, and computational knowledge, as well as the necessary tools for processing, analysing, and modelling large datasets. Special emphasis is also placed on interpreting the information obtained and using it to transform data into meaningful knowledge, leading to more accurate conclusions and better decision-making. The course focuses particularly on learning and applying some mathematical, statistical, and computational methods to identify patterns, trends, and relationships within massive and complex datasets.

Learning Outcomes

1. CA27 (Competence) Apply the mathematical tools for large database analysis to solve problems in the business or research field.

2. CA28 (Competence) Integrate the mathematical tools of large database analysis with other tools in multidisciplinary work environments.
3. CA29 (Competence) Communicate to a specialised and non-specialised audience the results obtained from applying the statistical methods of large database analysis in different fields.
4. CA30 (Competence) Work in multidisciplinary teams on the development of projects where deep learning techniques are applied.
5. KA21 (Knowledge) Describe the deep learning techniques used in large database analysis.
6. KA22 (Knowledge) Describe the mathematical tools used to analyse large databases and volumes of data.
7. SA27 (Skill) Apply mathematical techniques for processing large databases to analyse particular phenomena, such as consumer behaviour patterns, market trends and social network analysis.
8. SA27 (Skill) Apply mathematical techniques for processing large databases to analyse particular phenomena, such as consumer behaviour patterns, market trends and social network analysis.
9. SA27 (Skill) Apply mathematical techniques for processing large databases to analyse particular phenomena, such as consumer behaviour patterns, market trends and social network analysis.
10. SA28 (Skill) Interpret the results obtained from analysing a large database.

Content

Block 1. Text Mining (10 h):

- Fundamentals of Text Mining - From text to numbers.
- Data cleaning.
- Tokenization.
- Stemming.
- Lemmatization.
- POS, NER.
- Data chunking.

Block 2. Statistics for Big Data (18 h):

- Topic 1. The principle of sufficiency: Summarising the information from large, complex datasets.
  Sufficient statistics and the factorisation theorem.
  Classic examples of sufficient statistics.
- Topic 2. Classical linear models with large, complex datasets.
  A quick review of classical linear models and the ordinary least squares estimator. Examples.
  Use of sufficient statistics for the estimation of classical linear models. Examples.
- Topic 3. Generalised linear models with large, complex datasets: Logit and Poisson models.
  A quick review of logit and Poisson models, as well as the associated likelihood-based estimators.
  Introduction to the idea of segmentation: The "Divide and Recombine" method and meta-analysis-based methods.
  The estimation of logit and Poisson models using the idea of segmentation. Examples.
  Some recent advances in the idea of "Divide and Recombine".
- Topic 4. The problem of multiple testing and the false discovery rate.
  Introduction to knockoff variables.

Block 3. Deep Learning (10 h):

- Fully Connected Neural Networks.
- Convolutional Neural Networks.
- Recurrent Neural Networks.
- Keras and Tensorflow.

Activities and Methodology

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Lecture sessions	19	0.76	CA28, CA29, KA21, KA22, CA28
Problem-solving and practical sessions	11	0.44	CA27, CA28, CA29, CA30, SA27, SA28, CA27
Type: Supervised			
Problem-solving and practical sessions	8	0.32	CA27, CA28, CA29, CA30, SA27, SA28, CA27
Type: Autonomous			
Self-directed learning to deepen understanding of lecture topics	43	1.72	CA28, KA21, KA22, SA28, CA28
Tasks to practise the concepts introduced during in-person classes	50	2	CA27, CA28, CA29, CA30, KA21, KA22, SA27, CA27

The course is organised into three independent blocks, each taught by a different professor. While the first block of the course (10 hours) introduces concepts related to data mining, which are generally applied to large datasets, the second block (18 hours) focuses on the statistical methods and knowledge required for modelling such large volumes of data. Particular emphasis is placed on fitting and making inferences with classical linear models and generalised linear models (logistic and Poisson models) when working with large amounts of information. Finally, the third block of the course (10 hours) introduces students to some of the most relevant methods in deep learning, with a special focus on neural networks and their applications.

Each block generally combines lecture sessions introducing theory and technical concepts and with laboratory and problem-solving sessions, which may be instructor-led or based on independent student work. During lecture sessions, professors may use slides, which will be shared via the Moodle course page. Similarly, examples in R and/or Python may be presented during laboratory and problem-solving sessions, and these will generally be made available via Moodle. Students are also expected to independently review supplementary materials shared via Moodle in order to deepen their understanding of the concepts introduced in the in-person sessions.

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
Projects Block 1	26	5	0.2	CA27, CA28, CA29, KA22, SA27, SA28
Projects Block 2	48	9	0.36	CA27, CA28, CA29, KA22, SA27, SA28
Projects Block 3	26	5	0.2	CA27, CA28, CA29, CA30, KA21, KA22, SA27, SA28

The course evaluation is carried out independently for each of the described blocks. Each block's weighting in the final grade corresponds to its proportion of the total course hours. Each block typically consists of a number of projects, which can be completed either individually or in groups. In some cases, these projects will require an oral presentation of the content and results. The projects in blocks 1 and 3 each account for 26% of the final grade of the course, while those in block 2 account for 48% of the final grade of the course.

For each project, a document will be published on the Moodle page, containing the project description and requirements, as well as all necessary materials for its completion (datasets, additional information sources, etc.), the submission deadline and procedure, and any other information that the professor considers relevant. Grades for each project, as well as the final grades for each block and for the course as a whole, will also be published on the Moodle page.

Bibliography

Basic references:

- B. Efron, T. Hastie. Computer Age Statistical Inference, Cambridge University Press, 2018.
  https://bibcercador.uab.cat/permalink/34CSUC_UAB/1eqfv2p/alma991010753063206709
- G. James, D. Witten, T. Hastie and R. Tibshirani. An Introduction to Statistical Learning (with applications in R). Springer, 2013.
  https://bibcercador.uab.cat/permalink/34CSUC_UAB/1c3utr0/cdi_globaltitleindex_catalog_2960062
- D. Skillicorn. Understanding Complex Data. Data Mining with Matrix Decomposition. Chapman & Hall, 2007.
  https://bibcercador.uab.cat/permalink/34CSUC_UAB/1eqfv2p/alma991004136809706709

Complementary references:

- B. Everitt and T. Hothorn. An introduction to Applied Multivariate Analysis with R. Springer, 2011.
- B. Everitt. An R and S+ Companion to Multivariate Analysis. Springer, 2005.
- J. Faraway. Extending de Linear Model with R. Chapman & Hall, Miami, 2006.
- J. Faraway. Linear Models with R. Chapman & Hall, Boca Raton, 2005.
- W. Härdle and L. Simar. Applied Multivariate Statistical Analysis. Springer. 2007.
- B. Ripley. Pattern Recognition and Neural Networks. Cambridge University Press, 2002.
- L. Torgo. Data Mining with R. Learning with Case Studies. Chapman & Hall, Miami. 2010
- W Venables, B Ripley. Modern Applied Statistics with S-PLUS. Springer, New York.

The professors may provide other interesting references for each block, which will be available via the Moodle page.

Software

R Core Team (2021). R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/.

Python

Groups and Languages

Please note that this information is provisional until 30 November 2025. You can check it through this link. To consult the language you will need to enter the CODE of the subject.

Name	Group	Language	Semester	Turn
(TEm) Theory (master)	1	English	second semester	afternoon