Universitat Autònoma de Barcelona

Mathematics for Big Data

Code: 45562
Credits: 6

2026/2027

Degree programme	Type	Course
Modelling for Science and Engineering	OP	1

Contact lecturer

Name : Amanda Fernandez Fontelo

Email : amanda.fernandez@uab.cat

Teaching staff

Sundus Zafar

Víctor Navas Portella

Group languages

You can consult this information at the end of the 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

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

- CA27 (Apply the mathematical tools for large database analysis to solve problems in the business or research field.) Apply the mathematical tools for large database analysis to solve problems in the business or research field.
- CA28 (Integrate the mathematical tools of large database analysis with other tools in multidisciplinary work environments.) Integrate the mathematical tools of large database analysis with other tools in multidisciplinary work environments.

- CA29 (Communicate to a specialised and non-specialised audience the results obtained from applying the statistical methods of large database analysis in different fields.) Communicate to a specialised and non-specialised audience the results obtained from applying the statistical methods of large database analysis in different fields.
- CA30 (Work in multidisciplinary teams on the development of projects where deep learning techniques are applied.) Work in multidisciplinary teams on the development of projects where deep learning techniques are applied.
- KA21 (Describe the deep learning techniques used in large database analysis.) Describe the deep learning techniques used in large database analysis.
- KA22 (Describe the mathematical tools used to analyse large databases and volumes of data.) Describe the mathematical tools used to analyse large databases and volumes of data.
- SA27 (Apply mathematical techniques for processing large databases to analyse particular phenomena, such as consumer behaviour patterns, market trends and social network analysis.) Apply mathematical techniques for processing large databases to analyse particular phenomena, such as consumer behaviour patterns, market trends and social network analysis.
- SA28 (Interpret the results obtained from analysing a large database.) Interpret the results obtained from analysing a large database.

Contents

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):

- Data aggregation: Sufficient statistics, meta-analysis and weighted aggregation of local estimators.
- Scalable statistical methods: Divide-and-Conquer methods, subsampling techniques and stochastic optimization (Stochastic Gradient Descent).
- Applications

Block 3. Deep Learning (10 h):

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

Learning activities and methodology

Title	Hours	ECTS	Learning outcomes
Problem-solving and practical sessions	11	0.44	CA27, CA28, CA29, CA30, SA27, SA28
Self-directed learning to deepen understanding of lecture topics	43	1.72	CA28, KA21, KA22, SA28
Lecture sessions	19	0.76	CA28, CA29, KA21, KA22
Problem-solving and practical sessions	8	0.32	CA27, CA28, CA29, CA30, SA27, SA28

Tasks to practise the concepts introduced during in-person classes	50	2	CA27, CA28, CA29, CA30, KA21, KA22, SA27

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 methods of data aggregation, including sufficient statistics, meta-analysis and methods for combining different estimations, as well as scalable statistical methods, including divide and conquer methods, subsampling techniques and stochastic optimization such as Stochastic Gradient Descent. These methods will be studied in some real data application, specially related to linear and generalised linear models. 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.

On the use of Artificial Intelligence (AI): In this course, the use of Artificial Intelligence (AI) technologies is permitted as an integral part of the development of coursework, provided that the final submission reflects a significant contribution by the student in terms of analysis and personal reflection. Students must clearly identify the parts that have been generated using AI, specify the tools used, and include a critical reflection on how these tools have influenced both the process and the final outcome of the assignment. Failure to disclose the use of AI will be considered a breach of academic integrity and may result in a reduction of the assignment grade or, in more serious cases, additional disciplinary sanctions.

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

Continuous assessment activities

Title	Weight	Hours	ECTS	Learning outcomes
Projects Block 3	26	5	0.2	CA27, CA28, CA29, CA30, KA21, KA22, SA27, SA28
Projects Block 2	48	9	0.36	CA27, CA28, CA29, KA22, SA27, SA28
Projects Block 1	26	5	0.2	CA27, CA28, CA29, 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_296006297
- 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.
- Wang, C., Chen, M.-H., Schifano, E., Wu, J., & Yan, J. (2016). Statistical methods and computing for big data. Statistics and Its Interface, 9(4), 399-409.
- Chen, X., Cheng, G., & Xie, M. (2021). Divide-and-conquer methods for big data analysis. Statistical Science, 36(4), 582-597.
- Zhu, R., Ma, P., Mahoney, M. W., & Yu, B. (2015). Optimal Subsampling Approaches for Large Sample Linear Regression. arXiv preprint arXiv:1509.05111.
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. Proceedings of COMPSTAT 2010, 177-186.

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

Course groups and languages

The information provided is provisional until November 30. After this date, you will be able to consult the language of each group through this link. To access the information, you will need to enter the course CODE

Type of teaching	Group	Language	Semester	Shift
(TEm) Theory (master)	1	English	second semester	afternoon