UAB
Universitat Autònoma
de Barcelona

Mathematics and Big Data

Code: 43478
ECTS Credits: 6

2024/2025

Degree	Type	Year
4313136 Modelling for Science and Engineering	OT	0

Contact

Name: Amanda Fernandez Fontelo

Email: amanda.fernandez@uab.cat

Teachers

Sundus Zafar

Carles Barril Basil

Teaching groups languages

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

Prerequisites

Students should have a basic knowledge of linear algebra, statistical inference and linear models. We also expect students to have programming skills. Previous experience with R and Python is helpful.

Objectives and Contextualisation

The aim of this course is to learn and apply different mathematical and statistical methods related to the discovery of relevant patterns in large data sets. Nowadays, huge amounts of data are being generated in many fields and the aim of this course is to learn how to extract information from such data.

Competences

- Analyse, synthesise, organise and plan projects in the field of study.
- Apply logical/mathematical thinking: the analytic process that involves moving from general principles to particular cases, and the synthetic process that derives a general rule from different examples.
- Apply techniques for solving mathematical models and their real implementation problems.
- Conceive and design efficient solutions, applying computational techniques in order to solve mathematical models of complex systems.
- Formulate, analyse and validate mathematical models of practical problems in different fields.
- Isolate the main difficulty in a complex problem from other, less important issues.

- Solve complex problems by applying the knowledge acquired to areas that are different to the original ones.

Learning Outcomes

1. Analyse, synthesise, organise and plan projects in the field of study.
2. Apply Bayesian statistical techniques to predict the behaviour of certain phenomena.
3. Apply logical/mathematical thinking: the analytic process that involves moving from general principles to particular cases, and the synthetic process that derives a general rule from different examples.
4. Identify real phenomena as models of stochastic processes and extract new information from this to interpret reality.
5. Isolate the main difficulty in a complex problem from other, less important issues.
6. Solve complex problems by applying the knowledge acquired to areas that are different to the original ones.
7. Solve real data analysis problems by identifying them appropriately from the perspective of Bayesian statistics.
8. Use appropriate statistical packages and Bayesian methods solutions to solve specific problems.

Content

Text Mining

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

Statistics

- Summarising the information from large data sets:
  The principle of sufficiency and sufficient statistics.
  Applications to classical and generalised linear models.
  The Biglm package.
- Problems of likelihood estimation problems for large data sets:
  The method of "Divide and Recombine" and generalisations.
  The idea of segmentation, analysis of chunks of data, and methods based on meta-analysis.
  Applications to linear and generalised linear models.
- The problem of multiple testing and false discovery rate:
  The idea of knockoff variables.
- Functional Data Analysis:
  Observed functional data and its computational representation.
  Descriptive statistics and dimensionality reduction.
  Depth measures for functional data.
  Functional linear models and classification techniques.

Deep Learning

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

Activities and Methodology

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Homework (problems & computer excercises)	36	1.44	3, 8
Lectures	38	1.52	1, 4
Type: Autonomous			
Homework	44	1.76	1, 3, 4, 5, 6, 8
Personal study, readings	20	0.8	4

Lectures, supervised exercises and individual activities to work on data analysis projects based on statistical and computational tools.

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
Deep Learning	0,25	3	0.12	1, 3, 4, 5, 6, 8
First Homework Statistics	0.25	3	0.12	1, 3, 4, 5, 6, 7, 8
Homework Text Mining	0,25	3	0.12	1, 3, 5, 6, 8
Second Homework Statistics	0,25	3	0.12	1, 2, 3, 4, 5, 6, 7

Homework: Completion and presentation of the proposed exercises.
Final project: The students must choose one of a series of topics provided by the teaching staff, undertake a data project and prepare a talk. This task can be done in groups.
The deadlines will be announced during the course and will be strictly adhered to.

Bibliography

Referències bàsiques

- B. Efron, T. Hastie, *Computer Age Statistical Inference*, Cambridge University Press (2016) (5th Ed 2017) https://web.stanford.edu/~hastie/CASI/index.html

- G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning (with applications in R)*. Springer, 2013.
- D. Skillicorn, "Understanding Complex Data. Data Mining with Matrix Decomposition". Chapman&Hall, 2007.

Referències Complementàries

- 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.
- Collins FS and Varmus H, "A new initiative on precision medicine". N Engl J Med. 2015 Feb 26;372(9):793-5 .
- Jensen A.B. et al, "Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients". Nat Commun 2014 Jun 24; 5:4022.
- J.D. Jobson, "Applied Multivariate Analysis". Vol I i II. Springer, 1992.
- R. Johnson and D.W. Wichern, "Applied Multivariate Statistical Analysis". Pearson Education International, 2007.
- P.Y.Lum et al., "Extracting insights from the shape of complex data using topology". Sci. Rep. 3, 1236; DOI:10.1038/srep01236 (2013).
- A. Rencher, "Methods of Multivariate Analysis". Wiley Series in Probability and Mathematical Statistics, 2002.
- G. Singh, F. Mémoli, G. Carlsson, "Topological methods for the analysis of High dimensional data sets and 3D object recognition". Eurographic Symp. on Point-Based Graphics, 2007
- P. Kokoszka, M. Reimherr, *Introduction to Functional Data Analysis*. CRC Press.(2017).
- Ramsay, J. , B. W. Silverman,*Functional Data Analysis Springer* (2nd Ed. 2005).

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

Language list

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