2022/2023

Mathematics and Big Data

Code: 43478
ECTS Credits: 6

Degree	Type	Year	Semester
4313136 Modelling for Science and Engineering	OT	0	2

Contact

Name: Alejandra Cabaña Nigro

Email: anaalejandra.cabana@uab.cat

Use of Languages

Principal working language: english (eng)

Teachers

Pere Puig Casado

Antonio Lozano Bagen

Sundus Zafar

Prerequisites

Students should have basic knowledge linear algebra, statistical inference and linear models. We also assume the students have programming skills.

Previous experience with R and Python will be helpful.

Objectives and Contextualisation

The aim of this course is to learn and apply various mathematical and statistical methods related to the discovery of relevant patterns in data sets. Nowadays, huge amounts of data are being generated in many fields, and the goal of this course is to learn how to extract information from such data. When dealing with large datasets, mathematical procedures should be scalable, so we will be concerned with methods that can be scaled and/or paralelized.

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
- Lemmatizattion
- POS,NER
- Data chunking

Statistics

- The problem of multiple testing.
- Linear and Generalized linear methods: LASSO/BigLASSO, Ridge Regression and Elastic Nets. Feature Selection.
- Summarising the information of large data sets: sufficient statistics. Application to linear models. The Biglm package.
- Likelihood estimation problems for large data sets. Segmentation, analysis of chunks of data, methods based on meta-analysis. Applications to Generalised linear models.

Alternative topics,

- Functional Data Analysis: Observed functional data and its computational representation,descriptive statistics and dimensionality reduction, depth measures for FD, two-sample problem for FD, Functional linear models, class cation techniques.
- Gaussian processes for Machine Learning
- Model Explainability

Deep Learning

- Fully Connected Neural Networks.
- Convolutional Neural Networks.
- Recurrent Neural Networks.

- Keras and Tensorflow.

Methodology

Lectures, supervised exercices and autonomous activities directed to perform 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.

Activities

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	3, 1, 5, 4, 6, 8
Personal study, readings	20	0.8	4

Assessment

Homework: Completion and presentation of the proposed exercises.

Final Project: The studensts must choose one of a series of topics provided by the teaching staff and undertake a data project and prepare a talk. This task can be done in group.

Due dates will be anounced during the course and will be strict.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Deep Learning	0,25	3	0.12	3, 1, 5, 4, 6, 8
Homework Estadística Part B	0.25	3	0.12	3, 1, 5, 4, 6, 7, 8
Homework Statistics Part A	0,25	3	0.12	3, 1, 2, 5, 4, 6, 7
Homework Text Mining	0,25	3	0.12	3, 1, 5, 6, 8

Bibliography

Basic references

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.

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 Statisticswith 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.

D. Skillicorn, "Understanding Complex Data. Data Mining with Matrix Decomposition". Chapman&Hall, 2007.

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

Journal of Statistical Software, http://www.jstatsoft.org/

P. Kokoszka, M. Reimherr, *Introduction to Functional Data Analysis*. CRC Press.(2017).

Ramsay, J. , B. W. Silverman,*Functional Data Analysis Springer* (2nd Ed. 2005).
Dealing with Data (2011) Special Issue. Science 11 February 2011:692-789

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