Universitat Autònoma de Barcelona

2017/2018

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

Albert Ruíz Cirera

External teachers

Isabel Serra

Prerequisites

Students should have basic knowledge of statistics, linear algebra and linear models and programming skills. A previous experience with statistical software "R" 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. This process is often called learning from data.

To begin with, we shall discuss two basic tools: k-nearest neighbours and linear regression. Then we shall move to other linear methods, both classical and more modern (such as lasso).

Another topic well be non-linear statistical learning, mainly tree-based methods, and support vector machines. We shall also consider a setting in which we only have input variables, but no output. In particular, we present principal component analysis, K-means clustering and hierarchical clustering.

We will also focus in clustering methods but with a markedly different approach, using topology based methods to extract insights from the shape of complex data sets.

Skills

- 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

1. Introduction: Statistical learning, concept, methods, and basic examples: knn and regression.

2. Linear methods (logistic regression, linear discriminant analysis, lasso).

3. Non-linear methods.

4. Tree-based methods (regression and clssification trees, bagging, boosting and random forests).

5. Support vector Machines

5. Unsupervised learning: PCA and KNN

6. Topological data analysis.

Methodology

Lectures, supervised exercices and autonomous activities directed to realise a data analysis project based on statistical and topological learning tools.

Activities

Title	Hours	ECTS	Learning outcomes
Type: Directed			
Lectures	38	1.52	1, 4, 7
Type: Supervised			
Completion of exercices	36	1.44	2, 3, 7, 8
Type: Autonomous			

Personal study, readings	20	0.8	2, 4, 7
Project	44	1.76	1, 2, 3, 4, 5, 6, 7, 8

Evaluation

The following factors will be taken into account:

Exercises (60%): Completion and presentation of the proposed exercises. Due dates will be anounced during the course and will be strict.

Project (40%) (in pairs): The student proposes a project related to the contents of the course that must be approved by the professors. It might be executed in pairs. A final report and a public presentation are compulsory.

Evaluation activities

Title	Weighting	Hours	ECTS	Learning outcomes
Exercices	0,6	6	0.24	2, 3, 4, 5, 6, 7, 8
Project	0.4	6	0.24	1, 2, 3, 4, 5, 6, 7, 8

Bibliography

Basic references

[JWHT] G. James, D. Witten, T. Hastie and R. Tibshirani, "An Introduction to Statistical Learning (with applications in R)". Springer, 2013.

[C] Gunnar Carlsson, "Topology and data". Bull. AMS 46,2 (2009), 255-308.

Complementary references

[EH] 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).

[F1] J Faraway, " Extending de Linear Model with R", Chapman & Hall, Miami, 2006.
[F2] J Faraway, "Linear Models with R", Chapman & Hall, Boca Raton, 2005.

[**HS]** W. Härdle and L. Simar, "Applied Multivariate Statistical Analysis". Springer. 2007.

[R] B. Ripley, "Pattern Recognition and Neural Networks". Cambridge University Press, 2002.

[T] L. Torgo. "Data Mining with R. Learning with Case Studies". Chapman & Hall, Miami. 2010
[EH] W Venables, B Ripley, "Modern Applied Statisticswith S-PLUS", Springer, New York.

[CV] Collins FS and Varmus H, "A new initiative on precision medicine". N Engl J Med. 2015 Feb 26;372(9):793-5 .

[HTF] T. Hastie R. Tibshirani and J. Friedman, "Elements of Statistical Learning (Data Mining,Inference and Prediction)". Springer, 2009.

[J] 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.

[Jo] J.D. Jobson, "Applied Multivariate Analysis". Vol I i II. Springer, 1992.

[JW] R. Johnson and D.W. Wichern, "Applied Multivariate Statistical Analysis". Pearson Education International, 2007.

[L] P.Y.Lum et al., "Extracting insights from the shape of complex data using topology". Sci. Rep. 3, 1236; DOI:10.1038/srep01236 (2013).

[Re] A. Rencher, "Methods of Multivariate Analysis". Wiley Series in Probability and Mathematical Statistics, 2002.

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

[SMC] 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/

Dealing with Data (2011) Special Issue. Science 11 February 2011:692-789