Mathematics and Big Data

2015/2016

Code: 43478
ECTS Credits: 6

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

Contact

Name: Mercè Farré Cervelló

Email: Merce.Farre@uab.cat

Use of languages

Principal working language: english (eng)

Teachers

Jaume Aguadé Bover

Joan Valls Marsal

Carme Font Moragon

Prerequisites

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

Objectives and Contextualisation

The aim of the subject 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 is to understand what the data say. This process is often called learning from data.

The first part of the course deals with the spectral and singular value decomposition of matrices from standpoints algebraic and geometric. These decompositions are the basis of the principal component analysis (PCA) and other factorial methods that could be applied to reduce the data dimension and visualize some patterns. A second part is devoted to classical clustering methods, a broad class of methods for discovering unknown subgroups in data. PCA and clustering are two particular types of unsupervised statistical learning. In a third step, we 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.The final part of the course will be devoted to supervised statistical learning: regression analysis, classification and regression trees and neural networks, among others.

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

Introduction: Statistical learning, concept, methods, and examples

1. Matrix decompositions and factorial methods.

1.1. Spectral decomposition.

1.2. Singular value decomposition.
1.3. Principal component analysis.
1.4. Multidimensional scaling.

1.5. Exploratory factor analysis.
1.6. Correspondence analysis.

2. Classical clustering methods

2.1. Hierarchical clustering.

2.2. K-means Clustering.
2.3. Model-based methods.

3. Introduction to Topological Data Analysis

3.1. What is Topology and why it may be useful to data analysis.
3.2. Simplicial complexes and homology.
3.3. The Cech complex.
3.4. Discretization and persistence.
3.5. A classic example: Mumford's photographs.
3.6. Implementing topological data analysis.
4.7. Visualization through stratification and clustering.

4. Supervised learning methods.

4.1. Introduction to statistical modelling.
4.2. Tree-based methods.
4.3. Neural networks.
4.4. Random forests.
4.5. Performance assessment: cross-validation, train & testing and other validation. Procedures.
4.6. BIG DATA in research: some examples in biomedicine.

Methodology

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

Activities

Title	Hours	ECTS	Learning outcomes
Type: Directed			
Lectures	34	1.36	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	48	1.92	1, 2, 3, 4, 5, 6, 7, 8

Evaluation

First, it is necessary to attend at least 80% of all sessions. In the evaluation, the following factors will be taken into account:

Exercises (40%): Completion and presentation of the proposed exercises.

Exams (20%): At the end of each block, a test in which the achievement of the objectives will be assessed is performed.

Project (40%) (in pairs): The work consists in finding a appropriate database, analyze the data using two or more techniques learned during the course and write a final report to be presented publicly.

Evaluation activities

Title	Weighting	Hours	ECTS	Learning outcomes
Exercices	0,4	4	0.16	2, 3, 4, 5, 6, 7, 8
Project presentation	0,4	4	0.16	1, 2, 3, 4, 5, 6, 7, 8
Tests	0,2	4	0.16	4, 5

Bibliography

Basic references

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

[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.

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

[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.

Complementary references

[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