Universitat Autònoma de Barcelona

Complex Data Modelling

2022/2023

Code: 104864
ECTS Credits: 6

Degree	Type	Year	Semester
2503852 Applied Statistics	OB	3	2

Contact

Name: Rosario Delgado de la Torre

Email: rosario.delgado@uab.cat

Use of Languages

Principal working language:	catalan (cat)
Some groups entirely in English:	No
Some groups entirely in Catalan:	Yes
Some groups entirely in Spanish:	No

Teachers

Rosario Delgado de la Torre

Prerequisites

It is assumed that the student taking this subject has acquired the skills of the subjects of

- Càlcul 1,
- Eines informàtiques per a l'Estadística i Introducció a la Programació,
- Introducció a la Probabilitat i Inferència Estdística 1, i
- Aprenentatge Automàtic 1.

You will need a good level and practice in programming with R.

Objectives and Contextualisation

Learn what Bayesian Networks (BN) are and how they are used: BN are a probabilistic model used in Supervised Machine Learning that describe the probabilistic relationships between variables that affect a given phenomenon of interest (which can be a complex system) and can be used as classifiers.

Understand how Bayesian Networks are used to assess and quantify risks, among other applications.

Know different methodologies that will have to be applied, or not, when working with these models, in the pre-process phase of the database depending on its characteristics or in the construction phase of the predictive model.

Know different behavioral metrics to validatethemodel and understand its usefulness and adequacy, depending on the characteristics of the database.

Learn how to build R scripts that allow you to learn these models from a database and do their validation, using the relevant libraries. Apply it with real data.

Competences

- Analyse data using statistical methods and techniques, working with data of different types.
- Correctly use a wide range of statistical software and programming languages, choosing the best one for each analysis, and adapting it to new necessities.
- Critically and rigorously assess one's own work as well as that of others.
- Design a statistical or operational research study to solve a real problem.
- Formulate statistical hypotheses and develop strategies to confirm or refute them.
- Interpret results, draw conclusions and write up technical reports in the field of statistics.
- Make efficient use of the literature and digital resources to obtain information.
- Select and apply the most suitable procedures for statistical modelling and analysis of complex data.
- Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
- Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
- Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
- Summarise and discover behaviour patterns in data exploration.
- Use quality criteria to critically assess the work done.

Learning Outcomes

1. Analyse data through inference techniques using statistical software.
2. Analyse data using other models for complex data (functional data, recount data etc.).
3. Critically assess the work done on the basis of quality criteria.
4. Establish the experimental hypotheses of modelling.
5. Identify the stages in problems of modelling.
6. Identify the statistical assumptions associated with each advanced procedure.
7. Make effective use of references and electronic resources to obtain information.
8. Make slight modifications to existing software if required by the statistical model proposed.
9. Prepare technical reports within the area of statistical modelling.
10. Reappraise one's own ideas and those of others through rigorous, critical reflection.
11. Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
12. Students must be capable of collecting and interpreting relevant data (usually within their area of study) in order to make statements that reflect social, scientific or ethical relevant issues.
13. Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
14. Use graphics to display the fit and applicability of the model.
15. Validate the models used through suitable inference techniques.

Content

1. Introduction to Bayesian Networks (BNs).
   Definition.
   Inference with BNs.
   Learning BNs (both structure and parameters).
2. The BNs as classifiers.
   The classification task within Supervised Machine Learning.
   The MAP criterion.
   Types of BN (Naive Bayes, Augmented Naive, TAN).
   Classification type: binary, multi-class, multi-label.
3. Validation and behavioral metrics.
   Cross-validation.
   Metrics for the binary and multi-class case.
   Metrics for the case of ordinal classification.
4. Other aspects.
   Multi-label classification: the chains of classifiers.
   The cost-sensitive approach.

The problem of database imbalance: oversampling, thresholding, ...
Ensemblesof classifiers.
BN Gaussians and hybrids.
Dynamics BN.

Methodology

The subject is structured around theoretical classes, problems and practices. The follow-up of the subject is face-to-face, but it will be necessary to extend the teacher's explanations with the student's autonomous study, with the support of the reference bibliography and the material provided by the teacher.

The problem class will focus on solving some of the proposed problems. In the practical classes we will work with R and his libraries. Student participation in problem and practice classes will be especially valued.

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			
Practices (deliveries, controls)	12	0.48	1, 3, 9, 8
Problems	14	0.56	2, 4, 14, 5, 6, 13, 15
Theory	26	1.04	2, 1, 10, 3, 9, 4, 14, 5, 6, 8, 13, 11, 12, 7, 15
Type: Supervised			
Tutorials	10	0.4	10, 3, 11, 12, 7
Type: Autonomous			
Practical work with computer tools	30	1.2	1, 3, 9, 8, 15
Study and think problems	40	1.6	4, 14, 5, 6, 13, 15

Assessment

The final grade for this subject is obtained as the weighted average of the grades of:

- PAC1 (20%)
- PAC2 (20%)
- Exam (60%)

The PAC1 and PAC2 continuous assessment tests consist of a delivery of problems/practical exercises/work with R, which will be specified throughout the course.

Only those notes that are at least 3.5 out of 10 will be taken into account in the calculation of the weighted average (those that do not comply will weight 0).

To pass the subject it is necessary that this average is at least 5.0 out of 10. In case of not passing the subject in the first call, the student can present himself for recovery. The retake exam represents 100% of the final grade for those students who take the retake, which can only be students who have not passed the subject on the first call (the retake exam does not serve to improve the grade for students who have already passed).

The student who has presented the PAC1 or PAC2 deliveries, or has presented the exam or the recovery exam will be considered evaluable. Otherwise, it will be recorded in the minutes as Not Assessable.

For the eventual assignment of Honors, the marks of the second call will not be taken into account.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Exam	60%	3	0.12	2, 1, 10, 3, 9, 4, 14, 5, 6, 8, 13, 11, 12, 7, 15
PAC1	20%	6	0.24	2, 1, 10, 3, 9, 4, 14, 5, 6, 8, 13, 11, 12, 7, 15
PAC2	20%	9	0.36	2, 1, 10, 3, 9, 4, 14, 5, 6, 8, 13, 11, 12, 7, 15

Bibliography

- Norman Fenton and Martin Neil, "Risk Assessment and Decision Analysis with Bayesian Networks", CRC Press. A Chapman & Hall Book, 2013. (Available online)
- Radhakrishnan Nagarajan, Marco Scutari and Sophie Lèbre, "Bayesian Networks in R with applications in Systems Biology", Springer, 2013.(Available online)
- Oliver Porret, Patrick Naïm and Bruce Marcot, "Bayesian Networks. A practical guide to applications". Series: Statistics in Practice. Wiley, 2008.(Available online)
- Richard E. Neapolitan, "Learning Bayesian Networks", Prentice Hall Series in Artificial Intelligence, 2004.

Software

The R software will be used with some libraries that will be indicated in due course throughout the course. Preferably in the RStudio environment.