Universitat Autònoma de Barcelona

2020/2021

Machine Learning 1

Code: 104870
ECTS Credits: 6

Degree	Type	Year	Semester
2503852 Applied Statistics	OB	3	1

The proposed teaching and assessment methodology that appear in the guide may be subject to changes as a result of the restrictions to face-to-face class attendance imposed by the health authorities.

Contact

Name: Juan Ramón González Ruíz
Email: JuanRamon.Gonzalez@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

Joan Valls Marsal

Prerequisites

This course assumes that the student has obtained the knowledge taught in different courses on the following topics:

- Calculus in several variables.

- Probability

- Linear models.

- R programming.

Objectives and Contextualisation

This course aims to familiarize the student with different methods of machine learning by applying the point of view used when large amounts of data are available.

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.
- 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.
- Select statistical models or techniques for application in studies and real-world problems, and know the tools for validating them.

- Select the sources and techniques for acquiring and managing data for statistical processing purposes.
- 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.
- Students must develop the necessary learning skills to undertake further training with a high degree of autonomy.
- Summarise and discover behaviour patterns in data exploration.
- Use quality criteria to critically assess the work done.
- Work cooperatively in a multidisciplinary context, respecting the roles of the different members of the team.

Learning Outcomes

1. Analyse data using an automatic learning methodology.
2. Characterise homogeneous groups of individuals through multivariate analysis.
3. Critically assess the work done on the basis of quality criteria.
4. Describe the advantages and disadvantages of algorithmic methods compared to the conventional methods of statistical inference.
5. Develop a study based on multivariate methodologies and/or data mining to solve a problem in the context of an experimental situation.
6. Discover individuals' behaviours and typologies through data-mining techniques.
7. Identify the statistical assumptions associated with each advanced procedure.
8. Identify, use and interpret the criteria for evaluating compliance with the requisites for applying each advanced procedure.
9. Implement programmes in languages suitable for data mining.
10. Make effective use of references and electronic resources to obtain information.
11. Obtain and manage complex databases for subsequent analysis.
12. Reappraise one's own ideas and those of others through rigorous, critical reflection.
13. 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.
14. 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.
15. Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
16. Students must develop the necessary learning skills to undertake further training with a high degree of autonomy.
17. Use data mining methods to validate and compare possible models.
18. Use summary graphs of multivariate or more complex data.
19. Work cooperatively in a multidisciplinary context, accepting and respecting the roles of the different team members.

Content

These are the contents of the subject*

- Introduction to Tidyverse
- Introduction to machine learning
- Linear regression and logistics
- Steps prior to creating a predictive model and validation measures
- Machine learning methods
    - Analysis using classification trees
    - K-nearest neighbours
    - Random Forest
    - Boosting

• Learning methods for data n << p

- Shrinkage methods
- Regularization methods
- The 'caret' library

• Learning methods for big data

- XGBoost
- Lasso
- The 'H20' library

**Unless the requirements enforced by the health authorities demand a prioritization or reduction of these contents.*

Methodology

The course has two hours of theory and two hours of practices each week.

- Theory: the different methods with their particular characteristics are defined and explained and concrete examples are shown.

- Practices: working with the methods explained in theory class using different data sets and the R programming language.

It is considered that, for each hour of theory and practice, the student must dedicate an additional hour for the preparation and/or finalization of the session. Self-evaluating questionaires will be filled-in to check whether the main concepts are adquired after each session.

NOTE:

*The proposed teaching methodology may experience some modifications depending on the restrictions to face-to-face activities enforced by health authorities.

Activities

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Lab sessions	50	2	1, 3, 2, 6, 4, 18, 7, 8, 9, 11, 5, 16, 13, 14, 17
Type: Supervised			
Theory sessions	50	2	1, 12, 2, 6, 4, 18, 7, 8, 5, 16
Type: Autonomous			
Weekly tasks + self-evaluation	50	2	1, 12, 3, 2, 6, 4, 18, 7, 8, 9, 11, 5, 16, 15, 13, 14, 19, 10, 17

Assessment

The evaluation of the course will be carried out with one exam (final) some weekly tasks and self-evaluation questions. The final grade will be calculated with the formula:

NF = 0.3 * NE + 0.5 * NT + 0.2*NS

where NT is the average grade of weekly tasks, NS the average grade of self-evaluated questions and NE the grade of the examen that should be greater than 5.

At the end of the semester there will be a recovery examen for those students whose NE is less than 5 and NF lower than 5. In this case, the final grade will be calculated with the formula:

NF = 0.5 * NR + 0.5 * NT

where NR is the grade of the recovery exam.

NOTE: Student's assessment may experience some modifications depending on the restrictions to face-to-face activities enforced by health authorities.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
Final exam	30%	0	0	12, 2, 4, 18, 7, 8, 16, 13, 10
Self-evaluation	20%	0	0	1, 3, 6, 7, 8, 16, 13, 17
Tasks + self-learning	50%	0	0	1, 12, 3, 2, 6, 4, 18, 7, 8, 9, 11, 5, 16, 15, 13, 14, 19, 10, 17

Bibliography

Basic bibliography:

- An Introduction to Statistical Learning with Applications in R - Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

Complementary bibliography:

- The Elements of Statistical Learning: Data Mining, Inference, and Prediction - Trevor Hastie, Robert Tibshirani and Jerome Friedman

- Data Science from Scratch - Joel Grus

- Computer Age Statistical Inference: Algorithms, Evidence and Data Science - Trevor Hastie and Bradley Efron