Universitat Autònoma de Barcelona

2023/2024

Machine Learning 1

Code: 104870
ECTS Credits: 6

Degree	Type	Year	Semester
2503852 Applied Statistics	OB	3	1

Contact

Name: Juan Ramon Gonzalez Ruiz

Email: juanramon.gonzalez@uab.cat

Teaching groups languages

You can check it through this link. To consult the language you will need to enter the CODE of the subject. Please note that this information is provisional until 30 November 2023.

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
- Elastic net, ridge and lasso regression: improving logistic and linear regression
- Tractament de Big Data amb R
- La llibrería caret

- Mètodes d'aprenentatge automàtic
  KNN
  LDA
- Methods to deal with non-balanced outcomes
- Decision trees
-
    - Classification trees
    - Regression trees
    - Bagged trees
    - Random Forest
- Boosting
    - AdaBoost
    - GBM
    - Estochastic GBM
    - XGBoost
    - Others

**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.

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			
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 = 0.5 * NE + 0.4 * NT + 0.1*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/or NF lower than 5. In this case, the final grade will be calculated with the formula:

NF = 0.7 * NR + 0.3 * NT

where NR is the grade of the recovery exam.

Single evaluation (optional):

A comprehensive exam (4 hours) will be conducted to assess the knowledge and skills acquired throughout the course. This exam will be designed to evaluate the student's ability to apply the statistical analyses learned and their understanding of theoretical concepts.

The exam will consist of two main parts: statistical analysis and theoretical questions. In the statistical analysis section, relevant data will be provided, requiring the student to apply the statistical techniques and tools learned during the course. The following steps are expected from the student:

1. Problem identification: The student should understand the nature of the data and the analysis objectives.

2. Selection and application of techniques: The student will use the acquired knowledge to select and apply appropriate statistical techniques to analyze the data. This may include determining measures of central tendency, dispersion, correlation, regression, hypothesis testing, among others.

3. Interpretation of results: Once the analyses are performed, the student should interpret the results accurately, explaining their significance in the context of the given problem.

The second part of the exam will consist of theoretical questions that require written responses. These questions will be related to fundamental statistical concepts, their applicability in different situations, and their importance in decision-making. The student should demonstrate a clear understanding of the concepts and the ability to explain them coherently.

The evaluation of this exam will consider several criteria:

1. Accuracy and correctness in analyses: The student's ability to perform statistical analyses accurately and correctly will be evaluated, including selecting appropriate techniques and using the correct procedures.

2. Interpretation of results: The student's capacity to interpret and explain the results obtained from the statistical analyses will be assessed.

3. Completeness of theoretical responses: The student's ability to provide clear and comprehensive answers to the theoretical questions, demonstrating mastery of the concepts and their application, will be considered.

4. Organization and clarity of presentation: The overall organization of the exam, clarity of written responses, and quality of statistical result presentation will be taken into account.

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	50%	0	0	12, 2, 4, 18, 7, 8, 16, 13, 10
Self-evaluation	10%	0	0	1, 3, 6, 7, 8, 16, 13, 17
Tasks + self-learning	40%	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

- The bookdown of the topic: https://isglobal-brge.github.io/Aprendizaje_Automatico_1/

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

Software

Theory and practical exercises will be done using R