Course guide
200642 - ODS - Optimization in Data Science

Last modified: 01/06/2023

Unit in charge: School of Mathematics and Statistics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.

Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Optional subject).

Academic year: 2023 **ECTS Credits:** 5.0 **Languages:** Spanish, English

LECTURER

Coordinating lecturer: JORDI CASTRO PÉREZ

Others: Primer quadrimestre:
DANIEL BAENA MIRABETE - A
JORDI CASTRO PÉREZ - A

PRIOR SKILLS

Basic concepts of Statistics and Optimization/Operations Research.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
6. CE-2. Ability to master the proper terminology in a field that is necessary to apply statistical or operations research models and methods to solve real problems.
7. CE-3. Ability to formulate, analyze and validate models applicable to practical problems. Ability to select the method and / or statistical or operations research technique more appropriate to apply this model to the situation or problem.
8. CE-5. Ability to formulate and solve real problems of decision-making in different application areas being able to choose the statistical method and the optimization algorithm more suitable in every occasion.
Translate to english
9. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.
10. CE-7. Ability to understand statistical and operations research papers of an advanced level. Know the research procedures for both the production of new knowledge and its transmission.
11. CE-8. Ability to discuss the validity, scope and relevance of these solutions and be able to present and defend their conclusions.
12. CE-9. Ability to implement statistical and operations research algorithms.

Transversal:
1. ENTREPRENEURSHIP AND INNOVATION: Being aware of and understanding how companies are organised and the principles that govern their activity, and being able to understand employment regulations and the relationships between planning, industrial and commercial strategies, quality and profit.

2. SUSTAINABILITY AND SOCIAL COMMITMENT: Being aware of and understanding the complexity of the economic and social phenomena typical of a welfare society, and being able to relate social welfare to globalisation and sustainability and to use technique, technology, economics and sustainability in a balanced and compatible manner.

3. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are available.

4. EFFECTIVE USE OF INFORMATION RESOURCES: Managing the acquisition, structuring, analysis and display of data and information in the chosen area of specialisation and critically assessing the results obtained.

5. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

TEACHING METHODOLOGY

Theory:
The contents of the subject are presented and discussed with a combination of explanations on the board and with transparencies.
Training:
Laboratory sessions which demonstrate the use of software.

LEARNING OBJECTIVES OF THE SUBJECT

The aim of the topic is to introduce students to some applications of "data science" that can be formulated and solved by optimization techniques. The topic has three parts, each one representing 1/3 of the total:

1. Solution of statistical and machine learning models using integer and combinatorial optimization: design of experiments (orthogonal Latin squares), optimal clustering (k-medoids), heuristic clustering (k-means).

2. Continuous optimization methods for the solution of machine learning problems: regression, support vector machines, neural networks.

3. Introduction to the field of statistical disclosure control or statistical data protection: microdata and tabular data. This discipline includes a set of methods to ensure the confidentiality of individual data when disseminating statistical data, either microdata or aggregate data in tabular form. This issue is of great importance for national statistical offices, and in general, for any public or private entity that has to release data.

Skills to be acquired

* To formulate some "data science" applications as optimization problems (clustering, support vector machines, neural networks ...)
* To learn how to solve the formulated "data science" problems using optimization and machine learning software (scikit-learn, tensorflow).
* To know what is the field of statistical disclosure control or statistical data protection.
* To know software for data protection and to be able to protect data using some existing technique.
* To become familiar with literature of optimization for "data science".

STUDY LOAD

Type	Hours	Percentage
Hours small group	15,0	12.00
Self study	80,0	64.00
Hours large group	30,0	24.00

Total learning time: 125 h

CONTENTS

Integer and combinatorial ptimization in statistical and data science problems.

Description:
Basic background in integer and combinatorial optimization. Modelling integer and combinatorial optimization problems. Applications: design of experiments (orthogonal latin squares), optimal clustering (k-medoids), heuristic clustering (k-means).

Full-or-part-time: 15h
Theory classes: 10h
Practical classes: 5h

Continuous optimization in data science applications.

Description:
Ridge and LASSO regression. Support Vector machines (SVMs): primal formulation, KKT optimality conditions, duality in SVMs, dual formulation, optimization methods for SVMs, software for SVMs. Neural networks: structure and modeling of NNs as an optimization problem, optimization methods for NNs, software for NNs (scikit-learn, tensorflow).

Full-or-part-time: 15h
Theory classes: 5h
Practical classes: 10h

Introduction to statistical data protection.

Description:
Introduction. Definitions. Data types and methods. Methods for microdata. Methods for tabular data. Software for data protection.

Full-or-part-time: 15h
Theory classes: 10h
Practical classes: 5h

GRADING SYSTEM

A midterm exam about the contents of the first part of the subject (40% of the final mark) and practical assignments (60% of the final mark).

BIBLIOGRAPHY

Basic:
- Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron. Deep learning [on line]. Cambridge, Massachusetts: The MIT Press, [2016] [Consultation: 05/07/2023]. Available on: https://www.deeplearningbook.org/. ISBN 9780262035613.
- Arthanari, T.S. Mathematical programming in statistics. Wiley, 1993. ISBN 0471592129.
- Willenborg, Leon; Waal, Ton de. Elements of statistical disclosure control. New York: Springer, 2001. ISBN 0387951210.
- Cristianini, Nello; Shawe-Taylor, John. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press, 2000. ISBN 0521780195.
- Luenberger, David G; Ye, Yinyu. Linear and nonlinear programming [on line]. 3rd ed. New York: Springer, cop. 2008 [Consultation: 05/07/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-0-387-74503-9. ISBN 0387745033.
- Nocedal, Jorge; Wright, Stephen J. Numerical optimization [on line]. New York: Springer, 1999 [Consultation: 05/07/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-0-387-40065-5. ISBN 9780387987934.