Course guides
200642 - ODS - Optimization in Data Science

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: 2020  ECTS Credits: 5.0  Languages: English, Spanish

LECTURER

Coordinating lecturer: JORDI CASTRO PÉREZ

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

PRIOR SKILLS

Basic concepts of Statistics and 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.
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 course is to introduce students to some applications of "data science" that can be formulated and solved by optimization techniques. The course has three parts:
1. The first part of the course shows how to model and solve some statistical problems by optimization techniques (orthogonal Latin squares, classification problems k-median, etc).
2. The second part presents the mathematical foundations of optimization required to formulate and solve "support vector machines".
3. The third part is an introduction to the field of statistical disclosure control or statistical data protection. 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 ...)
* To learn how to solve the formulated "data science" problems using optimization software.
* To know what is the field of statistical disclosure control or statistical data protection.
* To know software for data protection.
* The ability to protect data using any existing technique.
* To become familiar with literature of optimization for "data science".

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours large group</td>
<td>30,0</td>
<td>24.00</td>
</tr>
<tr>
<td>Hours small group</td>
<td>15,0</td>
<td>12.00</td>
</tr>
<tr>
<td>Self study</td>
<td>80,0</td>
<td>64.00</td>
</tr>
</tbody>
</table>
Total learning time: 125 h

CONTENTS

Optimization in statistical problems.

Description:

Full-or-part-time: 11h 15m
Theory classes: 7h 30m
Practical classes: 3h 45m

Introduction to SVMs

Description:
Primal formulation of support vector machines (SVMs). KKT conditions of SVMs. The dual formulation of SVMs. Optimization approaches for SVMs.

Full-or-part-time: 11h 15m
Theory classes: 7h 30m
Practical classes: 3h 45m

Statistical data protection.

Description:

Full-or-part-time: 22h 30m
Theory classes: 15h
Practical classes: 7h 30m

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: