Course guide
270960 - OTDM - Optimization Techniques for Data Mining

Unit in charge: Barcelona School of Informatics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.
Degree: MASTER'S DEGREE IN DATA SCIENCE (Syllabus 2021). (Optional subject).
Academic year: 2022  ECTS Credits: 6.0  Languages: English

LECTURER
Coordinating lecturer: JORDI CASTRO PÉREZ
Others: Primer quadrimestre: JORDI CASTRO PÉREZ - 10 FRANCISCO JAVIER HEREDIA CERVERA - 10

PRIOR SKILLS
Basic background in linear algebra, calculus, and programming languages is needed for the course.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science

General:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats

Transversal:
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. 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.

Basic:
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

TEACHING METHODOLOGY
The students will have available all the course material.

About two thirds of lecture time will be devoted to optimization algorithms and their properties, and the rest will be for presenting and solving exercises and problems

Lab sessions will be devoted to the solution of some data science applications (neural networks, support vector machines, clustering) using optimization methods.
LEARNING OBJECTIVES OF THE SUBJECT

1. Discerning what is an optimization problem and its type and having a basic knowledge of optimization algorithms

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours large group</td>
<td>54,0</td>
<td>36.00</td>
</tr>
<tr>
<td>Self study</td>
<td>96,0</td>
<td>64.00</td>
</tr>
</tbody>
</table>

Total learning time: 150 h

CONTENTS

**Unconstrained Optimization**

Description:
- Line search. Acceptability of step sizes.
- General minimization algorithm.
- Gradient method. Rate of convergence.
- Newton’s method. Factorizations to ensure convergence.
- Quasi-Newton methods.

**Constrained Optimization and Support Vector Machines.**

Description:
- Introduction to the modelling language AMPL.
- Introduction to Support Vector Machines (SVM)
- KKT Optimality conditions of constrained optimization. Optimality conditions of SVM.
- Duality in Optimization. The dual of the SVM.

**Integer Programming**

Description:
- Modelling problems with binary variables.
- The branch and bound algorithm for integer programming
- Gomory’s cutting planes algorithm.
- Minimal spanning tree problem and algorithms of Kruskal and Prim. Application to clustering.
**ACTIVITIES**

**Unconstrained Optimization**

**Description:**
Line search. Acceptability of step sizes.
General minimization algorithm.
Gradient method. Rate of convergence.
Newton's method. Factorizations to ensure convergence.
Weighted least squares.
Introduction to AMPL. The Neos solver site.

**Specific objectives:**
1, 2, 3

**Related competencies:**
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CT5. 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.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

**Full-or-part-time:** 52h 18m
Theory classes: 17h 18m
Self study: 35h
Constrained Optimization and Support Vector Machines

Description:
- Introduction to Support Vector Machines (SVM)
- KKT Optimality conditions of constrained optimization. Optimality conditions of SVM.
- Duality in Optimization. The dual of the SVM.

Specific objectives:
1, 2, 3

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CT5. 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.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 52h 24m
Theory classes: 17h 24m
Self study: 35h
Integer Programming

Description:
- Modelling problems with binary variables.
- The branch and bound algorithm for integer programming
- Gomory's cutting planes algorithm.
- Minimal spanning tree problem and algorithms of Kruskal and Prim

Specific objectives:
1, 3

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE1. Develop efficient algorithms based on the knowledge and understanding of the computational complexity theory and considering the main data structures within the scope of data science
CT5. 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.
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

Full-or-part-time: 52h 18m
Theory classes: 17h 18m
Self study: 35h

GRADING SYSTEM
- Theory (40%). There will be two midterm exams (there is no final exam). The first midterm exam will consist on some practical exercises about the first two parts of the course, and it will be done before January. The second midterm exam will be about the third part of the course, and it will be done in January.

- Practical assignments (60%). There will be 3 lab assignments, one for each part of the course, all of them with the same weight on the final mark.

Additional exercises provided during the lectures may be taken into consideration to decide or to boost the final mark.

RESOURCES

Hyperlink: