Course guide
270964 - DAKD - Data Analysis and Knowledge Discovery

Unit in charge: Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science.

Degree: MASTER'S DEGREE IN DATA SCIENCE (Syllabus 2021). (Optional subject).
Academic year: 2022
ECTS Credits: 6.0
Languages: English

LECTURER
Coordinating lecturer: ALFREDO VELLIDO ALCACENA

Others:

PRIOR SKILLS
Students are expected to have at least some basic background in the area of artificial intelligence and, more specifically, with the areas of Machine Learning and Computational Intelligence. Some basic knowledge of probability theory and statistics would be beneficial. Other than this, the course is open to students and researchers of all types of background.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CE2. Apply the fundamentals of data management and processing to a data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature

Generical:
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

This course will build on different teaching methodology (TM) aspects, including:
TM1: Expositive seminars
TM2: Expositive-participative seminars
TM3: Orientation for individual assignments (essays)
TM4: Individual tutorization

LEARNING OBJECTIVES OF THE SUBJECT

1. Presenting DM as a process that should involve a methodology applied at its best.
2. Introducing the students to the new concept of DM for processes, called Process Mining.
3. Delving into some detail in one of the stages of DM: data exploration.
4. Dealing in detail with the problem of data visualization for exploration as a key issue in DM.
5. Introducing the students to the basics of probability theory as applied in Data Analysis and Knowledge Discovery (DAKD)
6. Introducing the students to the probabilistic variant of DAKD in the form of Statistical Machine Learning, both for supervised and unsupervised learning models.
7. Dealing in detail with different unsupervised models for data visualization, including case studies.
8. Approaching the multi-faceted concept of data mining (DM) from different perspectives.

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours large group</td>
<td>45,0</td>
<td>30.00</td>
</tr>
<tr>
<td>Guided activities</td>
<td>9,0</td>
<td>6.00</td>
</tr>
<tr>
<td>Self study</td>
<td>96,0</td>
<td>64.00</td>
</tr>
</tbody>
</table>

Total learning time: 150 h

CONTENTS

**Introduction to the concept of data mining (DM).**

Description:
DM is a multi-faceted concept that requires discussion and clarification. We will do this at the beginning of the course.

**DM as a methodology.**

Description:
We argue that DM should not be focused on the concept of data analysis/modeling, but, instead, should be treated as a methodology with diverse inter-related stages.

**DM for processes: Process Mining.**

Description:
A new development in DM methodologies is that which deals with one specifically suited for processes. It is called Process Mining and will be described and discussed in this course.
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data exploration in DM.</td>
<td>One of the main stages of well-structures DM methodologies is Data exploration. It will be discussed as a preamble to data visualization.</td>
</tr>
<tr>
<td>Data visualization for exploration.</td>
<td>One of the aspects of the problem of data exploration is data visualization. It has a research 'life' of its own as it involves not only computer-based mathematical models, but also natural perception and processing.</td>
</tr>
<tr>
<td>Basics of probability theory in Data Analysis and Knowledge Discovery (DAKD)</td>
<td>For a long time in the last half-century, multivariate statistics and artificial intelligence (mostly in the field of machine learning) have developed in parallel without fully meeting. Statistical machine learning has bridged that field over the last two decades. We introduce it by first providing some basic principles of probability theory (Bayesian inference).</td>
</tr>
<tr>
<td>Statistical Machine Learning for DAKD: supervised models.</td>
<td>Once the basics of Bayesian inference are set, we will delve into the field of Statistical Machine Learning for IDA, starting with supervised learning models, with an emphasis on feed-forward artificial neural networks.</td>
</tr>
<tr>
<td>Statistical Machine Learning for DAKD: unsupervised models.</td>
<td>Once the basics of Bayesian inference and of Statistical Machine Learning for IDA in supervised models are set, we will continue with unsupervised models, focusing on self-organizing maps and related models.</td>
</tr>
<tr>
<td>Unsupervised models for data visualization, with case studies.</td>
<td>In the final item of the contents of the course, we will bring statistical machine learning and data visualization together by discussing some probabilistic unsupervised learning models for data visualization, including some case studies as an example.</td>
</tr>
</tbody>
</table>
ACTIVITIES

Essay on DAKD for DM

Description:
Students will have to write a research essay on the topic of DAKD for DM, with different options:
1. State of the art on a specific DAKD-DM topic
2. Evaluation of a DAKD-DM software tool with original experiments
3. Pure research essay, with original experimental content

Specific objectives:
1, 2, 3, 4, 5, 6, 7, 8

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE2. Apply the fundamentals of data management and processing to a data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
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: 3h
Guided activities: 3h
Introduction to Data Mining and its Methodologies

Description:
Introduction to Data Mining as a general concept and to its methodologies for practical implementation

Specific objectives:
1

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE2. Apply the fundamentals of data management and processing to a data science problem
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
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.
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: 23h
Theory classes: 9h
Guided activities: 1h
Self study: 13h

Process Mining

Description:
Introduction to the novel concept of Process Mining and its application within the DM framework.

Specific objectives:
2

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE2. Apply the fundamentals of data management and processing to a data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
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.

Full-or-part-time: 9h
Theory classes: 3h
Guided activities: 1h
Self study: 5h
Data Visualization

Description:
As part of the DM stage of Data Exploration, we focus in the problem of Data Visualization.

Specific objectives:
3, 4

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE2. Apply the fundamentals of data management and processing to a data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
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.
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: 16h
Theory classes: 6h
Guided activities: 1h
Self study: 9h

Basics of probability theory for intelligent data analysis

Description:
Introduction to probability theory for intelligent data analysis, with a focus on Bayesian statistics

Specific objectives:
5

Related competencies:
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
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.

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: 16h
Theory classes: 6h
Guided activities: 1h
Self study: 9h
Statistical Machine Learning methods

Description:
The meeting of statistics and machine learning: Statistical Machine Learning methods, from the point of view of both supervised and supervised learning

Specific objectives:
5, 6

Related competencies:
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
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.

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: 31h
Theory classes: 12h
Guided activities: 1h
Self study: 18h
SML in data visualization, with case studies

Description:
We merge the topics of SML and data visualization, illustrating its use with some real case studies.

Specific objectives:
4, 7, 8

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE2. Apply the fundamentals of data management and processing to a data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE12. Apply data science in multidisciplinary projects to solve problems in new or poorly explored domains from a data science perspective that are economically viable, socially acceptable, and in accordance with current legislation
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
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 knowledge 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: 25h
Theory classes: 9h
Guided activities: 1h
Self study: 15h

GRADING SYSTEM
The course will include two evaluation tasks:
The first one will be a data science purely analytical task performed according to data mining principles.
The second one will involve writing an essay according to one of these three modalities:
1. State of the art on an specific IDA-DM topic
2. Evaluation of an IDA-DM software tool with original experiments
3. Pure research essay, with original experimental content

BIBLIOGRAPHY
Basic:

Complementary: