Course guide
270717 - IDADM - Intelligent Data Analysis and Data Mining

Last modified: 21/07/2022

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

Degree: MASTER'S DEGREE IN ARTIFICIAL INTELLIGENCE (Syllabus 2017). (Optional subject).

Academic year: 2022 **ECTS Credits:** 4.5 **Languages:**

LECTURER

Coordinating lecturer: ALFREDO VELLIDO ALCACENA

Others: Primer quadrimestre:
LUIS ANTONIO BELANCHE MUÑOZ - 10
ALFREDO VELLIDO ALCACENA - 10

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 Leaning 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:
CEA11. Capability to understand the advanced techniques of Computational Intelligence, and to know how to design, implement and apply these techniques in the development of intelligent applications, services or systems.
CEA4. Capability to understand the basic operation principles of Computational Intelligence main techniques, and to know how to use in the environment of an intelligent system or service.
CEA7. Capability to understand the problems, and the solutions to problems in the professional practice of Artificial Intelligence application in business and industry environment.
CEP1. Capability to solve the analysis of information needs from different organizations, identifying the uncertainty and variability sources.
CEP5. Capability to design new tools and new techniques of Artificial Intelligence in professional practice.

Generical:
CG3. Capacity for modeling, calculation, simulation, development and implementation in technology and company engineering centers, particularly in research, development and innovation in all areas related to Artificial Intelligence.

Transversal:
CT4. 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.

CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..
CT7. ANALISIS Y SINTESIS: Capability to analyze and solve complex technical problems.

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 id 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 Intelligent Data Analysis (IDA)
6.Introducing the students to the probabilistic variant of IDA 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.
9.Approaching the multi-faceted concept of data mining (DM) from different perspectives.

STUDY LOAD

Type	Hours	Percentage
Self study	72,0	63.94
Hours large group	33,8	30.02
Guided activities	6,8	6.04

Total learning time: 112.6 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.

Data exploration in DM.

Description:
One of the main stages of well-structures DM methodologies is Data exploration. It will be discussed as a preamble to data visualization.

Basics of probability theory in Intelligent Data Analysis (IDA)

Description:
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).

Data visualization for exploration.

Description:
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.

Statistical Machine Learning for IDA: supervised models.

Description:
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.

Statistical Machine Learning for IDA: unsupervised models.

Description:
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.

Unsupervised models for data visualization, with case studies.

Description:
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.

ACTIVITIES

Essay on IDA for DM

Description:
Students will have to write a research essay on the topic of IDA for DM, with different options:
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

Specific objectives:
1, 2, 3, 4, 5, 6, 7, 9

Related competencies :
CG3. Capacity for modeling, calculation, simulation, development and implementation in technology and company engineering centers, particularly in research, development and innovation in all areas related to Artificial Intelligence.
CEP5. Capability to design new tools and new techniques of Artificial Intelligence in professional practice.
CEA11. Capability to understand the advanced techniques of Computational Intelligence, and to know how to design, implement and apply these techniques in the development of intelligent applications, services or systems.
CEP1. Capability to solve the analysis of information needs from different organizations, identifying the uncertainty and variability sources.
CEA4. Capability to understand the basic operation principles of Computational Intelligence main techniques, and to know how to use in the environment of an intelligent system or service.
CEA7. Capability to understand the problems, and the solutions to problems in the professional practice of Artificial Intelligence application in business and industry environment.
CT4. 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.

CT7. ANALISIS Y SINTESIS: Capability to analyze and solve complex technical problems.
CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..

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, 9

Related competencies :
CG3. Capacity for modeling, calculation, simulation, development and implementation in technology and company engineering centers, particularly in research, development and innovation in all areas related to Artificial Intelligence.
CEP5. Capability to design new tools and new techniques of Artificial Intelligence in professional practice.
CEA7. Capability to understand the problems, and the solutions to problems in the professional practice of Artificial Intelligence application in business and industry environment.
CT4. 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.

CT7. ANALISIS Y SINTESIS: Capability to analyze and solve complex technical problems.
CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..

Full-or-part-time: 15h
Theory classes: 6h
Self study: 9h

Process Mining

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

Specific objectives:
2

Related competencies :
CG3. Capacity for modeling, calculation, simulation, development and implementation in technology and company engineering centers, particularly in research, development and innovation in all areas related to Artificial Intelligence.
CEP5. Capability to design new tools and new techniques of Artificial Intelligence in professional practice.
CEP1. Capability to solve the analysis of information needs from different organizations, identifying the uncertainty and variability sources.
CEA7. Capability to understand the problems, and the solutions to problems in the professional practice of Artificial Intelligence application in business and industry environment.
CT4. 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.

CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..

Full-or-part-time: 9h
Theory classes: 3h
Self study: 6h

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 :
CG3. Capacity for modeling, calculation, simulation, development and implementation in technology and company engineering centers, particularly in research, development and innovation in all areas related to Artificial Intelligence.
CEP5. Capability to design new tools and new techniques of Artificial Intelligence in professional practice.
CEA11. Capability to understand the advanced techniques of Computational Intelligence, and to know how to design, implement and apply these techniques in the development of intelligent applications, services or systems.
CEP1. Capability to solve the analysis of information needs from different organizations, identifying the uncertainty and variability sources.
CEA4. Capability to understand the basic operation principles of Computational Intelligence main techniques, and to know how to use in the environment of an intelligent system or service.
CT4. 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.

CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..

Full-or-part-time: 15h
Theory classes: 6h
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 :
CEP1. Capability to solve the analysis of information needs from different organizations, identifying the uncertainty and variability sources.
CT4. 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.

CT7. ANALISIS Y SINTESIS: Capability to analyze and solve complex technical problems.
CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..

Full-or-part-time: 15h
Theory classes: 6h
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

Full-or-part-time: 28h
Theory classes: 12h
Self study: 16h

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:
7

Related competencies :
CG3. Capacity for modeling, calculation, simulation, development and implementation in technology and company engineering centers, particularly in research, development and innovation in all areas related to Artificial Intelligence.
CEP5. Capability to design new tools and new techniques of Artificial Intelligence in professional practice.
CEA11. Capability to understand the advanced techniques of Computational Intelligence, and to know how to design, implement and apply these techniques in the development of intelligent applications, services or systems.
CEP1. Capability to solve the analysis of information needs from different organizations, identifying the uncertainty and variability sources.
CT4. 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.

CT7. ANALISIS Y SINTESIS: Capability to analyze and solve complex technical problems.
CT6. REASONING: Capability to evaluate and analyze on a reasoned and critical way about situations, projects, proposals, reports and scientific-technical surveys. Capability to argue the reasons that explain or justify such situations, proposals, etc..

Full-or-part-time: 15h
Theory classes: 6h
Self study: 9h

GRADING SYSTEM

The course will be evaluated through a final essay that will take 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:
- MacKay, D.J.C. Information theory, inference, and learning algorithms. Cambridge University Press, 2003. ISBN 0521642981.
- Hand, D.; Mannila, H.; Smyth, P. Principles of data mining. MIT Press, 2001. ISBN 026208290X.
- Bishop, C.M. Pattern recognition and machine learning. New York: Springer, 2006. ISBN 0387310738.

Complementary:
- Hand, D.J. Statistics: a very short introduction. Oxford University Press, 2008. ISBN 9780199233564.
- Spence, R. Information visualization: an introduction. 3rd ed. Switzerland: Springer, 2020. ISBN 9783319073408.
- Yau, N. Visualize this: the flowing data guide to design, visualization, and statistics. Wiley, 2011. ISBN 9780470944882.