Course guide
270650 - DAKD - Data Analysis and Knowledge Discovery

Last modified: 12/07/2021

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

Degree: MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Optional subject).

Academic year: 2021 **ECTS Credits:** 6.0 **Languages:** English

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:
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.

Generical:
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.

Transversal:
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.

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 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

Type	Hours	Percentage
Guided activities	9,0	6.00
Self study	96,0	64.00
Hours large group	45,0	30.00

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.

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.

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.

Basics of probability theory in Data Analysis and Knowledge Discovery (DAKD)

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).

Statistical Machine Learning for DAKD: 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 DAKD: 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 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 an specific DAKD-DM topic
2. Evaluation of an 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 :
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.

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 :
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.

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 :
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.
CEC3. Ability to apply innovative solutions and make progress in the knowledge that exploit the new paradigms of Informatics, particularly in distributed environments.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.

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 :
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.

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 :
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.

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 :
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.

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 :
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.

Full-or-part-time: 25h
Theory classes: 9h
Guided activities: 1h
Self study: 15h

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. Oxfrod University Press, 2008. ISBN 9780199233564.
- Spence, R. Information visualization: design for interaction. 2nd ed. Pearson/Prentice Hall, 2007. ISBN 9780132065504.
- Yau, N. Visualize this: the flowing data guide to design, visualization, and statistics. Wiley, 2011. ISBN 9780470944882.