Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

Coordinating unit: 480 - IS.UPC - University Research Institute for Sustainability Science and Technology

Teaching unit: 715 - EIO - Department of Statistics and Operations Research

Academic year: 2015

Degree: MASTER'S DEGREE IN SUSTAINABILITY SCIENCE AND TECHNOLOGY (Syllabus 2013). (Teaching unit Optional)

ECTS credits: 5 Teaching languages: English

Teaching staff

Coordinator: KARINA GIBERT OLIVERAS

Others: Karina Gibert Oliveras
Miquel Sànchez-Marrè

Opening hours

Timetable: Please send a mail to the corresponding lecturer to schedule a meeting

Prior skills

Basics knowledge of R package
Basic programming skills
Basic Statistics

Requirements

Fonaments d?Estadística Aplicada i Mesura de la Sostenibilitat i el Desenvolupament

Degree competences to which the subject contributes

Specific:

CE04. The ability to apply, critically and effectively, conceptual frameworks, data collection and processing techniques, applied statistics, mathematical modelling, systems analysis, geographic information systems, information and communication technologies and industrial ecology to meeting the challenges of sustainability and sustainable development.

Teaching methodology

MD1: Lecture or conference (EXP): Sharing knowledge through lectures by professors or by external guest speakers.

MD4: Tutorials of practical or theoretical works (TD): to perform an activity in the classroom, or a theoretical or practical exercise, individually or in small groups, with the advice of the teacher

and

MD6: Extensive project (PA): learning based in the design, planning and realisation in groups of a complex or extensive project or piece of work, applying and extending knowledge and writing a report on this approach and the results and conclusions

Learning objectives of the subject

Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

The main goal of this course is to provide a global view of the application of Data Science to real socio-environmental problem solving. The use of Data Mining techniques is presented in a complete Knowledge Discovery process devoted to extract relevant information from different kind of socio-environmental data (surveys, monitoring, data-warehouses...) to support decision-making from phenomena or organizations with high degrees of complexity. The course is focused to real socio-environemntal problems and to provide the proper elements to design efficient and correct Data Mining processes, according to the real problem targeted at every application, as well as to analyze the Data Scientist competences required to deal with. Main Data Mining methods are presented; training on several important practical aspects is provided, like effects on wrong pre-processing, wrong selection of data mining method, wrong interpretation of results or assumption of false hypothesis for the analyzed process; effective communication of results to decision-makers and reporting is also carefullly analyzed. This issues will help to guarantee the validity and utility of final results, as well as real impact of the analysis into the target domain. Real cases from socio-environmental field, like water management, sustainable touristic activities, pollution or land uses will be discussed to show the versatility of the discipline to provide better knowledge and decision suppport to a wide spectrum of very difficult real socio-environmental problems.

Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

Content

Introduction	Learning time: 2h 30m Theory classes: 2h 30m

Description:
1.1. Data Science, Data Mining, Knowledge Discovery from Databases and Intelligent decision support.
1.2. Data Mining Pillars: Statistics, Artificial Intelligence, Information Systems, Visualization

Related activities:
Presentation of the project to be developed along the course and working teams definition

Specific objectives:
The Data Science and the overall process of Knowledge Discovery from Databases is presented, together with its steps and including Data Mining itself.
The disciplinary pillars of Data Mining are introduced: Statistics and Artificial Intelligence, Information Systems and Data Visualization
Finally, the basic schema of a Knowledge Discovery process is presented.

Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

Scope, KDD process and Data Structures	Learning time: 2h 30m Theory classes: 2h 30m

Description:
2. Scope of the discipline
2.1. Types of Problems suitable of Data Science
2.2. Ill-structured domains
2.3. A priori knowledge; Implicit knowledge. Causes and consequences
2.4. Main Data Mining Softwares (R, weka, rapid miner)

3. Formalising the Data Science problem and designing the complete Knowledge Discovery process

4. Data Structures
4.1 Main Socio-environmental data sources
4.2. Data and Metadata Representation

Related activities:
Projects proposal approval, dowload dataset

Specific objectives:
Scope of the discipline
Different natures of real socio-environmental problems and their different levels of complexity are discussed according to the classification proposed by Simpson. Ill-structured domains are introduced, as well as a priori and implicit knowledge management, causes and consequences.
Some software tools for developing data mining tasks are presented, with special focus on R system.

3. Formalising the Data Science problem and designing the complete Knowledge Discovery process
The steps of the Data Science process and the Knowledge Discovery process involved are introduced.
4. Data Structures
Main data structures analyzed by Data Mining techniques in socio-environmental fields.
Importance of metadata, formats and contents

480093 - TDS - Socio-Environmental Data Science

Preprocessing

Learning time: 5h

Theory classes: 5h

Description:

5. Preprocessing
5.1. Data quality issues
5.2 Filtering and Sampling
5.3 Missing data treatment
5.4 Outliers
5.5 Data transformation and Derived data
5.6. Feature weighting and dimensionality reduction

Related activities:

Preprocess your data for the project

Specific objectives:

Discussion on the importance of data quality and consequences of quality lack. Introduction of relevant aspects in data preparation step: Missing data, outliers detection and treatment, derived variables, transformed variables, filtering, sampling, feature weighting, dimensionality reduction (feature selection and factorial methods), all of them critical to guarantee the validity of the analysis. Good practice guidelines will be provided

Choosing the proper Data Mining method

Learning time: 2h 30m

Theory classes: 2h 30m

Description:

6. Choosing the proper Data Mining method
6.1. The problem-oriented approach
6.2 Criteria determining the suitability of a Data Mining method
6.3 The Data Mining Methods Conceptual Map (DMMCM-map)

Related activities:

Designing the complete KDD process for your project and working plan

Specific objectives:

The course follows a problem-oriented Data Science approach, where the nature of the problem mainly determines the analysis process and non vice-versa. Factors determining a correct choice of data mining method in real cases are discussed. The DMMCM typology of methods is presented as a conceptual basis for selection.

Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

Data Mining Step: Descriptive Methods	Learning time: 2h 30m Theory classes: 2h 30m

Description:
7. Data Mining step
7.1. Descriptive Methods
Clustering: partitioning methods, hierarchical, scalability. Hybrid methods, introduction of prior expert knowledge. Knowledge elicitation

Related activities:
Cluster your data

Data Mining: Associative Methods	Learning time: 2h 30m Theory classes: 2h 30m

Description:
.2. Associative Methods
Association Rules mining, factorial methods, bayesian networks

Related activities:
Use some associative method on your data

Data Mining: Discriminant Methods	Learning time: 2h 30m Theory classes: 2h 30m

Description:
7.3. Discriminant Methods
Decision trees, rule induction, support vector machines, discriminant analysis, random forest, Ensemble methods and bagging, hybrid methods.

Related activities:
Predict a qualitative variable

Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

Data Míning: Predictive Methods	Learning time: 2h 30m Theory classes: 2h 30m

Description:
7.8. Predictive methods
Regressión, statistical modelling in general. Temporal methods, Artificial Neural Networks, Swarm Intelligence.

Related activities:
Predict (one or more) numerical variables

Spatio-temporal data mining	Learning time: 2h 30m Theory classes: 2h 30m

Description:
7.9. Spatio-temporal management

Related activities:

General review of project advances

Specific objectives:
Some tools to deal with spatio-temporal data will be introduced

Post-processing and validation	Learning time: 2h 30m Theory classes: 2h 30m

Description:
8. Post-processing and validation
8.1. Post-processing tools
8.2. Model validation
8.3. Results validation

Related activities:
Validation of models in your project.

Specific objectives:
Post-processing tools and validation tools for both models and results adapted to different Data Mining methods.
Case wastewater treatment

480093 - TDS - Socio-Environmental Data Science

Reporting and results communication	Learning time: 2h 30m Theory classes: 2h 30m

Description:

9. Reporting and results communications

Related activities:
Review of reporting the project

Specific objectives:

Crucial to guarantee that the results of the Data Science process provide effective decision support to the end-user and the analysis have real impact on the target domain

Planning of activities

Progress presentation of projects	Hours: 2h 30m Theory classes: 2h 30m

Description:
Oral presentation of first part of project and discussion
Written deliverable

Specific objectives:
Milestone to sincronize all students with a suitable working plan
Communication and reporting skills are evaluated together with technical skills and organization of the working team

Final projects presentation	Hours: 2h 30m Theory classes: 2h 30m

Description:
Oral presentation and written deliverable of the complete project. General and individual discussion with the teacher

Specific objectives:
Evaluation of the technical, communication and reporting skills, as well as the organizational performance of the working team

Last update: 08-09-2015

480093 - TDS - Socio-Environmental Data Science

Qualification system

A big project will be developed by groups, by applying the methods lectured in class, under the teacher supervision. Some intermediate deliverables will be required in order to make easier the long term planning of the total work. For each of them, a mark is given on the following way

NP= 0.4 * quality of partial document + 0.3 * quality of oral presentation and discussion + 0.2 * individual performance at lab sessions

The final score will be computed on the basis of the final delivery (which is a compendium of all intermediate parts):
N= 0.4 * quality of final document + 0.3 * quality of oral presentation and discussion + 0.2 * individual performance at lab sessions

NF= 0.6*N+0.4 *(sum(all NP)/ n), where n is the number of Partial deliveries (1 or two, to be determined in class)

Bibliography