Syllabus

Subject

Subject / Group	21746 - Data Mining / 2
Degree	Degree in Computer Engineering (2014) - Third year Degree in Computer Engineering (2010) - Third year
Credits	6
Period	First semester
Language of instruction	English

Professors

Lecturers	**Office hours for students**					
	Starting time	**Finishing time**	**Day**	**Start date**	**End date**	**Office / Building**
Margarita María Lourdes Miró Juliá *(Responsible)* margaret.miro@uib.es	10:30	11:30	Wednesday	10/09/2018	21/12/2018	Anselm Turmeda D164
	11:30	12:30	Thursday	11/02/2019	31/05/2019	Anselm Turmeda D164

Context

21746 - Data Mining is an elective subject of the Computing specialization, and is taught in the first semester of the fourth year. Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. The ability to process and extract insightful information from large amounts of data has become a desired and necessary skill in almost every field of industry and science. Among other benefits, such information can provide useful knowledge, support decision-making, uncover hidden trends, and enable deeper understanding of observed phenomena.

This course will cover some of the main problems and challenges encountered in data analysis and applications, and provide fundamental tools and techniques for solving them. Students will become familiar with a variety of machine learning and data mining approaches, both supervised approaches and unsupervised ones.

The aim of the course is to acquire knowledge of the techniques and tools for the automatic extraction of information from large volumes of data. The topics include data extraction, exploratory data analysis, visualization, classification, clustering, frequent itemsets, recommender systems, dimensionality reduction etc. At the end of the semester you should:

* have a solid understanding of the basic concepts, principles, and techniques in data mining;
* be familiar with most of the classical data mining algorithms;
* be able to perform systematic analysis of real world data mining problems end to end;
* be able to model real-life problems and evaluate, visualize and interpret data mining models.

Requirements

Date of publication: 26/07/2018

The course assumes basic prior knowledge in probabilities, statistics, linear algebra, data structures, algorithms, and programming.

Essential

It is advisable to have successfully completed the following courses of the Computer Engineering degree:

* *20305 - Mathematics III - Statistics*
* *21722 - Artificial Intelligence*

Skills

Specific

* Ability to know and develop computational learning techniques and design and implement applications and systems that use them, including those dedicated to automatic extraction of information and knowledge from large volumes of data (CI307) .

Generic

* Capacity for analysis and synthesis, organization, planning and decision making (CTR01) .
* Ability to search for resources and information management in the field of information technology (CTR04) .
* Ability to acquire new knowledge autonomously (CTR03) .

Basic

* You may consult the basic competencies students will have to achieve by the end of the degree at the following address: http://www.uib.eu/study/grau/Basic-Competences-In-Bachelors-Degree-Studies/

Content

Data Mining studies algorithms and computational paradigms that allow computers to find patterns and regularities in databases, perform prediction and forecasting, and generally improve their performance through interaction with data. It is currently regarded as the key element of a more general process called Knowledge Discovery from Databases that deals with extracting useful knowledge from raw data. The knowledge discovery process includes data selection, cleaning, coding, using different statistical and machine learning techniques, and visualization of the generated structures. The course will cover all these issues and will illustrate the whole process by solving examples. Special emphasis will be given to the Machine Learning methods as they provide the real knowledge discovery tools.

The numbering of the topics does not imply a temporal sequence.

Range of topics

1. Introduction to Data Mining
   Comprehensive overview of the background and general themes of data mining. The Knowledge Discovery process.
2. Introduction to Data Science
   Data Science process: Extraction, Exploration, Visualization and Communication

Date of publication: 26/07/2018

3. Multilinear models
   Relationship between variables
4. Classification
   Classification models such as Decision trees, kNN, Naive Bayes, ANN and others
5. Clustering
   Clustering algorithms such as k-Means, k-medoids, hierarchical clustering and others
6. Association Rules
   Frequent itemsets and Apriori algorithm
7. Mining real data
   Applications: Examples of practical cases using specific software

Teaching methodology

Following below are the different types of activities to be performed by the students, both in the classroom and autonomously (at home, library, group-study, ...).

With the purpose of making easier the student's personal work, it has been requested that the course be part of the Aula Digital project that allows for flexibility in distance teaching and learning. Through this platform students will have at their disposal online communication with the teachers, a calendar with news of interest, electronic documents, proposed problems or assignments for both individual and group work, as well as a suitable environment for submitting assignments and access to their grades.

Important dates are available at the beginning of the semester through the UIB digital platform, these dates are tentative, except the dates of the final exam which is set by the Polytechnical School. The due dates of the final project and proposed assignments will be notified to the students in class and by announcements through Aula Digital.

Workload

The distribution of the proposed in-class work volume is illustrative and represents the planning made by the teachers, without taking into account all the contingencies that may arise during the course.

The distribution of out-of-class work, which is also illustrative, represents the ideal distribution planned by the teachers. The various activities of the course are planned so that for each hour of in-class work, the average student should work an hour and a half autonomously (individual study, problem solving, ...). Without an out-of-classroom work of this magnitude it will be difficult to reach the desired level of knowledge and obtain the desired skills.

In-class work activities (2.4 credits, 60 hours)

Modality	Name	Typ. Grp.	Description	Hours
Theory classes	Lecture class	Large group (G)	Concepts, procedures and their application to exercises and problems are introduced at master classes. The lecturer will describe the theoretical and practical foundations of the different topics covered in the course.	18

Date of publication: 26/07/2018

Modality	Name	Typ. Grp.	Description	Hours
Seminars and workshops	Workshop	Medium group (M)	In the workshop sessions, problems will be solved by the lecturer. Also, proposed problems will be solved by the students, individually or in small groups, with or without the support of the professor.	18
Practical classes	Final Project	Large group (G)	In the practical classes, the lecturer will model practical cases using software and tools specially designed for the treatment of data and its application to the resolution of data mining problems. The final project will be completed using the R software environment for statistical computing and graphics.	20
Assessment	Final Exam	Large group (G)	The final comprehensive exam evaluates the acquisition of the topics and competencies of the course.	4

At the beginning of the semester a schedule of the subject will be made available to students through the UIBdigital platform. The schedule shall at least include the dates when the continuing assessment tests will be conducted and the hand-in dates for the assignments. In addition, the lecturer shall inform students as to whether the subject work plan will be carried out through the schedule or through another way included in the Aula Digital platform.

Distance education tasks (3.6 credits, 90 hours)

Modality	Name	Description	Hours
Group or individual self-study	Self-study	Individual self-study to assimilate the contents presented in the master classes or to review autonomously proposed assignments. Followed by individual or group study focused on consolidating what has been assimilated in the individual self-study through the resolution of exercises and problems, and exam preparation.	35
Group or individual self-study	Project implementation	Individually or in small groups, the student will be required to complete the final project using specialized software.	55

Specific risks and protective measures

The learning activities of this course do not entail specific health or safety risks for the students and therefore no special protective measures are needed.

Student learning assessment

The lectures and discussions in class will be accompanied by workshops that combine theoretical questions, which emphasize the understanding of underlying data mining principles, together with programming tasks in R that demonstrate practical implementations of studied data mining techniques. Grades in this course will be based on these workshops, a project, and an exam.

* COMPREHENSIVE EXAM: final exam of the entire course, it may have questions on theoretical concepts and will always have a part consisting in the resolution of simple problems.

Date of publication: 26/07/2018

* WORKSHOPS: throughout the course, students will complete proposed assignments in small groups or individually. These assignments may consist in the submission of problems proposed by professor and resolved autonomously, proposed questionnaires, reseach on topics, ...
* FINAL PROJECT: the final project of the course integrates the knowledge acquired throughout the semester, it will consist in the resolution of a problem using the R software environment for statistical computing and graphics.

Frau en elements d'avaluació

In accordance with article 33 of Academic regulations, "regardless of the disciplinary procedure that may be followed against the offending student, the demonstrably fraudulent performance of any of the evaluation elements included in the teaching guides of the subjects will lead, at the discretion of the teacher, a undervaluation in the qualification that may involve the qualification of "suspense 0" in the annual evaluation of the subject".

Workshop

Modality	Seminars and workshops
Technique	Papers and projects **(non-retrievable)**
Description	In the workshop sessions, problems will be solved by the lecturer. Also, proposed problems will be solved by the students, individually or in small groups, with or without the support of the professor.
Assessment criteria	The students will solve the workshops proposed by the teacher. The correctness of the approach, the resolution of the problem, the clarity in the exposition, the rigor in the reasoning, ... will be assessed. Assessed skills: CI307, CTR01, CTR03 and CTR04.

Final grade percentage: 20%

Final Project

Modality	Practical classes
Technique	Papers and projects **(non-retrievable)**
Description	In the practical classes, the lecturer will model practical cases using software and tools specially designed for the treatment of data and its application to the resolution of data mining problems. The final project will be completed using the R software environment for statistical computing and graphics.
Assessment criteria	The students will complete a final project. The correctness of the strategy used, the ability to express and defend concepts learned throughout the course, the quality of the submitted report or documentation, ... will be assessed. Assessed skills: CI307, CTR01, CTR03 and CTR04.

Final grade percentage: 30%

Final Exam

Modality	Assessment
Technique	Objective tests **(retrievable)**
Description	The final comprehensive exam evaluates the acquisition of the topics and competencies of the course.
Assessment criteria	The final exam will assess the level of acquisition of the contents and the specific competences of the course.The correctness of the approach, the resolution of the problem, the clarity in the exposition, the rigor in the reasoning, ... will be evaluated.

Date of publication: 26/07/2018

Assessed skills: CI307, CTR01 and CTR03.

Final grade percentage: 50%with a minimum grade of 4

Resources, bibliography and additional documentation

Basic bibliography

An Introduction to Statistical Learning with applications in R. G. James, D. Witten, T. Hastie, R. Tibshirani, Springer 2013.
R and Data Mining. Examples and Case Studies Y. Zhao, Academic Press 2013.
Data Mining. A Knowledge Discovery Approach K.J. Cios, W. Pedrycz, R.W. Swiniarski, L.A. Kurgan Springer 2007.

Complementary bibliography

Data Mining: Concepts and Techniques, 3rd ed. Jiawei Han, Micheline Kamber and Jian Pei, Elsevier Inc. 2012.
Introduction to Data Mining. Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. Pearson Education Inc. 2006.
Barry J. Zimmerman. Self-regulated learning and academic achievement: an Overview. Educational Psychologist, 25(1), pp. 3-17. Springer, 2008