DEGREE CURRICULUM

DATA MINING

Coordination: BEJAR TORRES, RAMON

Academic year 2019-20

Subject's general information

<table>
<tr><td>Subject name</td><td>DATA MINING</td></tr>
<tr><td>Code</td><td>103089</td></tr>
<tr><td>Semester</td><td>1st Q(SEMESTER) CONTINUED EVALUATION</td></tr>
<tr><td>Typology</td><td>
<table>
<tr><th>Degree</th><th>Course</th><th>Character</th><th>Modality</th></tr>
<tr><td>Master's Degree in Informatics Engineering</td><td>2</td><td>OPTIONAL</td><td>Attendance-based</td></tr>
</table>
</td></tr>
<tr><td>Course number of credits (ECTS)</td><td>6</td></tr>
<tr><td>Type of activity, credits, and groups</td><td>
<table>
<tr><th>Activity type</th><th>PRALAB</th><th>TEORIA</th></tr>
<tr><th>Number of credits</th><td>3</td><td>3</td></tr>
<tr><th>Number of groups</th><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordination</td><td>BEJAR TORRES, RAMON</td></tr>
<tr><td>Department</td><td>COMPUTER SCIENCE AND INDUSTRIAL ENGINEERING</td></tr>
<tr><td>Teaching load distribution between lectures and independent student work</td><td>30% of the time are lectures (3 hours/week) and 70% is based on autonomous work.</td></tr>
<tr><td>Language</td><td>English</td></tr>
</table>

Teaching staff

Teaching staff	E-mail addresses	Credits taught by teacher	Office and hour of attention
BEJAR TORRES, RAMON	ramon@diei.udl.cat	3	
GUITART BRAVO, FRANCESC JOSEP	fguitart@diei.udl.cat	3	

Subject's extra information

To follow this subject, the student should have solid knowledge of structured, object and functional programming in python

Learning objectives

1. To know the current tools for Data Cleaning and Data Analyisis
2. To know the basics for the development of data-centric procedures using interactive programming tools
3. To know how to transform raw data from any source to consistent data for its analysis
4. To know how to implement and debug procedures for the transformation of massive data sets using Big Data approaches
5. To acquire a sceptic spirit in front of sets of data to incentive the exploratory analysis using computer science tools
6. To acquire the tools and knowledge for the descriptive analysis of potentially massive and intractable sets of data
7. To know and use the most basic data mining algorithms for discovering relevant features from big data sets
8. To know and use basic and advanced algorithsm for machine learning suitable for big data applications
9. To know the basics of recommender systems

Significant competences

University of Lleida strategic competences

- **UdL2:** Command of a foreign language.

Cross-disciplinary Competences EPS

- **EPS1:** Capacity of planning and organizing the personal work.
- **EPS3:** Capacity to convey information, ideas, problems and solutions to both a specialized and no specialized public.

- **EPS4:** Capacity to conceive, design and implement projects and/or contribute to new solutions, using engineering tools.
- **EPS5:** To be motivated for the quality and steady improvement.

General Competences

- **CG4:** Capacity to mathematically model, calculate and simulate in technological companies and engineering centres, particularly with regard to research, development and innovation tasks in all fields related to computer engineering.
- **CG8:** Capacity to apply the knowledge acquired for solving problems in new and unfamiliar situations within broader and more multidisciplinary contexts, and to be capable of integrating this knowledge.

Basic Competences

- **CB3:** Students are able to integrate knowledge and handle complexity, and formulate judgments based on information that was incomplete or limited, include reflecting on social and ethical.
- **CB4**: Students can communicate their conclusions -and the knowledge and rationale underpinning these, to specialist and non-specialist audiences clearly and unambiguously.

Degree-specific competences

- **CE1:** Capacity to understand and apply advanced knowledge in high-performance computing and numerical or computational methods to problems of engineering.
- **CE2:** Capacity to apply mathematical, statistical and artificial intelligence methods, design and develop applications, services, intelligent systems and systems based on knowledge.
- **CE4**: Capacity to model, design, define the architecture, implant, manage, operate, administer and keep applications, networks, systems, services and computer contents.
- **CE5**: Capacity to understand and know how to apply the operation and organisation of the Internet, the technologies and new generation network protocols, the models of components, middleware software and services.
- **CE7**: Capacity to design, develop, manage and evaluate mechanisms to certificate and guarantee the security in the treatment and access to the information in a processing or distributed local system.

Subject contents

1. Data cleaning
    - Introduction to Data Cleaning / Data Structures for Data Analysis with Python
    - Python Data Cleaning
    - Apache Spark
    - Exploratory analysis, summarization and data visualization

2. Data mining and learning
    - Mining frequent items/item sets and distinct elements
    - Dimensionality reduction
    - Linear and Logistic regression with SGD

- Naive Bayes classifiers
- Clustering
- Recommender systems

Methodology

Every week, each student will receive:

- Three hours of class attendance. Lectures are conducted by theoretical explanations accompanied by illustrative examples in the first part, finalizing with practical exercices in the second part. As a support material of the class, we will follow the slides or python notebooks of the course.
- Other support materials to follow the subject in a non-attendance way.

The evaluation is continuous throughout the semester and consists of four different parts:

- Two practices: Extending the bigdata application started at the previous subject “Massive data processing" with data cleaning and big data tools.
- Two reports and oral presentations about integrating data cleaning and data minning tools in their bigdata application.

Development plan

Week	**Description**	**Face-to-Face Activity**	**Autonomous Activity**	Hours (F and A)
1	Introduction to Data Cleaning / Data Analysis with Python	Lecture and participatory classes	Study and exercises resolution	3 2
2	Python Data Cleaning (i)	Lecture and participatory classes	Study	3 3
3	Python Data Cleaning (ii)	Lecture and participatory classes	Study and exercises resolution	3 4
4	Big Data with Apache Spark (i)	Lecture and participatory classes	Study and exercises resolution	3 2
5	Big Data with Apache Spark (ii)	Lecture and participatory classes	Study and Exercises resolution	3 3
6	Exploratory analysis, summarization and data visualization (i)	Lecture and participatory classes	Study	3 3
7	Exploratory analysis, summarization and data visualization (ii)	Lecture and participatory classes	Study and Exercises resolution	3 3
8	Data cleaning projects	Participatory classes	Work on bigdata application	3 10
9	Mining frequent items/item sets	Lecture and participatory classes	Study	3 3
10	Mining distinct elements	Lecture and participatory classes	Study and Exercises resolution	3 4
11	Dimensionality reduction	Lecture and participatory classes	Study	3 3

12	Linear and Logistic regression	Lecture and participatory classes	Study and Exercises resolution	3 5
13	Naive Bayes classifiers	Lecture and participatory classes	Study and Exercises resolution	3 3
14	Clustering - crisp and probabilistic	Lecture and participatory classes	Study and Exercises resolution	3 4
15	Recommender systems	Lecture and participatory classes	Study	3 3
16	Data mining projects	**Participatory classes**	Work on bigdata application	3 10
17	Big data application		Work on bigdata application	- 10
18	Tutorial classes			3
19	**Final presentation of big data application**	**Participatory classes**	**Preparattion of presentation**	3 3

Evaluation

The assessment for this course is based on continuous evaluation.

Evaluation activities	**%**	**Dates**	**O/V (1)**	**I/G (2)**
Data mining application (3)	40%	End Sem.	M	G
Data mining application presentation (4)	10%	End Sem.	M	G
Data cleaning exercises	25%	Middle Sem.	M	G
Data mining exercises	25%	Middle Sem.	M	G

(1) Mandatory / Voluntary

(2) Individual / Group

(3) There will be a revision of the work performed by each member of the group

(4) Each member of the group has to participate equally in the presentation and answer questions from the evaluator

Bibliography

- Wes McKinney. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Ipython. O'Really, 2012
- Holden Karau, Andy Konwinski, Patrick Wendell & Matei Zaharia. Learning Spark. O'Really, 2015
- Jure Leskovec, Anand Rajaraman & Jeffrey David Ullman. Mining of Massive Datasets. Cambridge University Press , 2014 (Find a copy at http://www.mmds.org/) It is also available at our Library, but only one user can borrow the book at the same time).