DEGREE CURRICULUM

BIG DATA PROJECT

Coordination: COMAS RODRIGUEZ, CARLOS

Academic year 2020-21

BIG DATA PROJECT 2020-21

Subject's general information

<table>
<tr><td>Subject name</td><td>BIG DATA PROJECT</td></tr>
<tr><td>Code</td><td>103090</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><td>Number of credits</td><td>3</td><td>3</td></tr>
<tr><td>Number of groups</td><td>1</td><td>1</td></tr>
</table>
</td></tr>
<tr><td>Coordination</td><td>COMAS RODRIGUEZ, CARLOS</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>Important information on data processing</td><td>Consult this link for more information.</td></tr>
<tr><td>Language</td><td>English</td></tr>
</table>

BIG DATA PROJECT 2020-21

Teaching staff	E-mail addresses	Credits taught by teacher	Office and hour of attention
COMAS RODRIGUEZ, CARLOS	carles.comas@udl.cat	3	
GARCIA GONZALEZ, ROBERTO	roberto.garcia@udl.cat	0	
VIRGILI GOMÀ, JORDI	jordi.virgili@udl.cat	3	

Subject's extra information

To follow this course, the student should have knowledge of Python and it recommended to have completed the subject Massive Data Processing.

Learning objectives

1. Understand and apply statistical techniques of data mining.
2. Use appropriately statistical packages for data analysis.
3. Understand and apply statistical techniques related to the operation of Big Data.
4. Use appropriately statistical packages for such exploitation.
5. Proposes adequate visualization of the analysed information that facilitates understanding.
6. Know the common applications in the area of big data and be able to develop solutions for these problems.
7. Systematically deploy the set of common techniques for solving big data problems.
8. Communicate effectively the results of the project to technical partners and customers.

Competences

General Competences

- **CG4**: Capacity for mathematical modelling, calculation and simulation in technologic and engineering business centres, particularly in research tasks, development and innovation in all areas 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.

Strategic Competences of UdL

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

Cross-disciplinary Competences

- **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.

Basic Competences

- **CB3**: Being able to integrate knowledge and handle the complexity to formulate judgments based on information that being incomplete or limited, include reflecting on social and ethical responsibilities linked to the application of their knowledge and judgments.
- **CB4**: Knowing how to communicate their conclusions -and the knowledge and rationale underpinning these, to specialist and non-specialist audiences clearly and unambiguously

Specific competences

- **CE1**: Capacity for the integration of technologies, applications and computer engineering systems, in general and in wider and multidisciplinary contexts.
- **CE2**: Capacity for the strategic planning, preparation, direction, coordination, and technical and economic management in the fields of the computer engineering in: systems, applications, services, networks, infrastructures or computer installations and centres or factories of software development, complying with the suitable fulfilment of the quality criteria and multidisciplinary working environments.
- **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. Introduction to CPA and the EM algorithm
2. Principal component analysis.
    1. Data matrices and associated spaces
    2. Principal component analysis
    3. Interpretation and quality of the results of a CAP
3. Expectation-Maximization Algorithm
    1. Maximum Likehood Estimation (MLE)
    2. The Expectation-Maximization (EM) Algorithm
    3. EM for Missing Data
    4. Gaussian Mixture Model (EM clustering)
4. Business Intelligence
    1. Supporting enterprise decisions on big data insights
    2. Smart Data
    3. Visualisation with matplotlib
    4. Databricks, Google Cloud,...
5. Interactive Exploration of Big Data
    1. Apache Spark for querying and interactively exploring great volumes of heterogeneous data
    2. SparkSQL
    3. Spark R and Exploratory Data Analysis

Methodology

All the courses of the Big Data Analytics block (including this one), will be graded by an unique, common, project involving all the courses topics (data gathering, processing, learning, statistics, visualization, etc.).

Students will work on that project from the beginning to the final courses. During regular courses, different topics will be introduced, showing their relation to the common project and how all topics fit together to create a real-world complex task or project.

BIG DATA PROJECT 2020-21

The three courses forming Big Data Analytics will use the same technological base setup:

- Python as the base programming language.
- Hadoop/Spark (with Java if required)
- Although during the courses other technological suites will be introduced:
  - Scala
  - NodeJS
  - Etc.

Development plan

Week	Description	Face-to-Face Activities	Autonomous Student Activities
1	Introduction to CPA and The EM algorithm	Exhibition events and methodology Lecture and participatory classes	Study and Exercises resolution
2	Data matrices and associated spaces	Lecture and participatory classes	Study and Exercises resolution
3	PCA	Lecture and participatory classes	Study and Exercises resolution
4	Interpertation and quality of the result of a CPA	Lecture and participatory classes	Study and Exercises resolution
5	Maximum Likehood Estimation (MLE)	Lecture and participatory classes	Study and Exercises resolution
6	The Expectation-Maximization (EM) Algorithm	Lecture and participatory classes	Study and Exercises resolution
7	Oral Presentations	CPA and EM Presentations	Study and Exercises resolution Project Development
8	Apache Spark Intro and Demos Big Data Project Review	Lecture and participatory classes Project Presentations	Cases Study Project Development
9	Big Data Exploration	Lecture and participatory classes	Project Development
10	Big Data Exploration	Lecture and participatory classes	Project Development
11	Big Data Exploration	Lecture and participatory classes	Project Development
12	Big Data Exploration	Lecture and participatory classes	Project Development
13	Business Intelligence	Lecture and participatory classes	Project Development
14	Business Intelligence	Lecture and participatory classes	Project Development
15	Business Intelligence	Project Presentations	Project Development

Evaluation

The assessment for this course is based on continuous evaluation. Depending on the sanitary situation some of these activities could be done as a classroom activity or virtually using tools of the CV.

ID	Evaluation activities	%	Dates	Mandatory	I/G (1)

ID	Evaluation activities	%	Dates	Mandatory	I/G (1)
OP1	Oral Presentation Resolution of a practical problem (CPA)	25%	Week 7	Y	Group
OP2	Oral Presentation Resolution of a practical problem (EM)	25%	Week 7	Y	Group
PP	Writing Work Project Planning	15%	Week 10	Y	Group
PD	Writing Work Project Deliverable	15%	Week 15	Y	Group
OP3	Oral Presentation Project	20%	Week 15	Y	Group

(1) Individual / Group

The final mark will be calculated using the following formula:

Final grade = 0,25*OP1 + 0,25*OP2 + 0,15*PP + 0,15*PD + 0,2*OP3

Bibliography

[Kar15] Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia, "Learning Spark: Lightning-Fast Big Data Analysis", O'Reilly, 2015

[Ryz15] Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills, "Advanced Analytics with Spark: Patterns for Learning from Data at Scale", O'Reilly, 2015

[Bae14] Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications"