Course guide
200645 - PBDE - Statistical Programming and Databases

Unit in charge: School of Mathematics and Statistics
Teaching unit: 723 - CS - Department of Computer Science.
707 - ESAII - Department of Automatic Control.
Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Optional subject).
Academic year: 2023  ECTS Credits: 5.0  Languages: English

LECTURER
Coordinating lecturer: ALEXANDRE PERERA LLUNA
Others: Primer quadrimestre:
ALEXANDRE PERERA LLUNA - A
ORIOL RICART VILARRUBIAS - A

PRIOR SKILLS
Non compulsory subject.
The student has already developed several abilities in Statistics and/or Operations Research previously.
A B2 (Cambridge First Certificate, TOEFL PBT >550) level of English is required.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES
Specific:
3. CE-1. Ability to design and manage the collection of information and coding, handling, storing and processing it.
4. CE-4. Ability to use different inference procedures to answer questions, identifying the properties of different estimation methods and their advantages and disadvantages, tailored to a specific situation and a specific context.
5. CE-5. Ability to formulate and solve real problems of decision-making in different application areas being able to choose the statistical method and the optimization algorithm more suitable in every occasion.

Transversal:
2. EFFECTIVE USE OF INFORMATION RESOURCES: Managing the acquisition, structuring, analysis and display of data and information in the chosen area of specialisation and critically assessing the results obtained.

10. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.

11. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are available.
TEACHING METHODOLOGY

The course is divided into 2 modules that are taught in succession. Each module consists roughly of a half part of the sessions. All classes are theoretical-practical and in them teachers present and discuss the basic concepts of each module. The support material will be published previously in Athena (teaching guide, contents, course slides, examples, evaluation activities schedule, bibliography, ...).

The student should devote the autonomous learning hours to the study of the subjects of the course, bibliography extension and follow-up of the laboratory practices.

LEARNING OBJECTIVES OF THE SUBJECT

This course presents and discusses tools and techniques to prepare students to data science. Main concepts introduced in class will cover tools and methods for data storage and analysis, including relational DB, noSQL and distributed databases, scientific computing, applied machine learning and deep learning with Python. Scala and Spark will also be considered. The course consists of two main modules.

MODULE 1:
First modulus will cover a crash course for scientific python for data analysis. This crash course will include include four main stages:
* Introduction to python language as a tool. ipython, ipython notebook (jupyter), basic types, mutability and immutability and object oriented programming.
* Short introduction to numerical python and matplotlib for graphical visualization.
* Introduction to scientific kits for data analysis with machine learning. Principal components analysis, clustering and supervised analysis with multivariate data.
* Introduction to Deep Learning with Python.

MODULE 2:
We introduce the Scala language and the Spark architecture.
* Scala as a functional language and the Scala collections.
* Spark and RDD (Resilient Distributed Data Sets).
* Spark and SQL.
* Introduction to MLlib.

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hours small group</td>
<td>15,0</td>
<td>12.00</td>
</tr>
<tr>
<td>Hours large group</td>
<td>30,0</td>
<td>24.00</td>
</tr>
<tr>
<td>Self study</td>
<td>80,0</td>
<td>64.00</td>
</tr>
</tbody>
</table>

Total learning time: 125 h

CONTENTS

Introduction to Python

Description:
a. Why Python?
b. Python History
c. Installing Python
d. Python resources

Full-or-part-time: 1h
Theory classes: 1h
Working with Python

Description:
- Workflow
- ipython vs. CLI
- Text Editors
- IDEs
- Notebook

Full-or-part-time: 1h
Theory classes: 1h

Getting started with Python

Description:
- Introduction
- Getting Help
- Basic types
- Mutable and in-mutable
- Assignment operator
- Controlling execution flow
- Exception handling

Full-or-part-time: 1h
Theory classes: 1h

Functions and Object Oriented Programming

Description:
- Defining Functions
- Input and Output
- Standard Library
- Object-oriented programming

Full-or-part-time: 1h
Theory classes: 1h

Introduction to NumPy

Description:
- Overview
- Arrays
- Operations on arrays
- Advanced arrays (ndarrays)
- Notes on Performance (%timeit in ipython)

Full-or-part-time: 2h
Theory classes: 2h
<table>
<thead>
<tr>
<th>Course</th>
<th>Description</th>
<th>Full-or-part-time</th>
<th>Theory classes</th>
</tr>
</thead>
</table>
| Matplotlib             | **Description:** a. Introduction  
b. Figures and Subplots  
c. Axes and Further Control of Figures  
d. Other Plot Types  
e. Animations  
**Full-or-part-time:** 2h  
Theory classes: 2h |                  |                |
| Introduction to Panda  | **Description:** a. Introduction to Pandas  
b. Series and Dataframes  
c. Importing and Exporting data through Pandas. Accessing Syntax Query Language (SQL) databases through Pandas.  
d. Aggregation, slicing, missingness  
e. Plotting within Pandas  
**Full-or-part-time:** 2h  
Theory classes: 2h |                |                |
| Python scikits         | **Description:** a. Introduction  
b. scikit-timeseries  
**Full-or-part-time:** 1h  
Theory classes: 1h |                |                |
| scikit-learn           | **Description:** a. Datasets  
b. Sample generators  
c. Unsupervised Learning  
d. Supervised Learning  
i. Linear and Quadratic Discriminant Analysis  
ii. Nearest Neighbors  
iii. Support Vector Machines  
e. Feature Selection  
**Full-or-part-time:** 8h  
Theory classes: 8h |                |                |
Practical Introduction to Scikit-learn

Description:
a. Solving an eigenfaces problem
   i. Goals
   ii. Data description
   iii. Initial Classes
   iv. Importing data
b. Unsupervised analysis
   i. Descriptive Statistics
   ii. Principal Component Analysis
   iii. Clustering
   c. Supervised Analysis
   i. k-Nearest Neighbors
   ii. Support Vector Classification
   iii. Cross validation

Full-or-part-time: 5h 30m
Theory classes: 5h 30m

Introduction to Zeppelin, Scala & Functional Programming

Description:
a. Immutable & Mutable
b. Lists and maps, filters, reductions
c. Map reduce
d. Other collections, Streams

Full-or-part-time: 5h
Theory classes: 5h

Spark architecture & Spark Core

Description:
a. Spark architecture: in particular Spark Core
b. Spark context
c. Types of operations: transformations and actions
d. RDD: Resilient Distributed Data Sets
e. Closure of a function

Full-or-part-time: 5h
Theory classes: 5h

Spark: MLlib

Description:
a. Description of the MLlib,
b. Labeled Points and features
c. Linear Regression Example

Full-or-part-time: 5h
Theory classes: 5h
Spark SQL

Description:
- Reading from a file.
- Spark Data Frame.
- Selection, filters, grouping, sorting.
- Window operations
- SQL
- Accessing and storing methods to a DB, SQL queries.
- SQL aggregates.

Full-or-part-time: 7h 30m
Theory classes: 7h 30m

GRADING SYSTEM

Final grade will be composed by:
- 1/4 Written exam first module
- 1/4 Written exam first module
- 1/2 Final practical assignment on large databases integrating concepts from both modules

BIBLIOGRAPHY

Basic:

Complementary: