Course guide
200645 - PBDE - Statistical Programming and Databases

Last modified: 01/06/2023

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.
Translate to english
6. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.
7. CE-7. Ability to understand statistical and operations research papers of an advanced level. Know the research procedures for both the production of new knowledge and its transmission.
8. CE-8. Ability to discuss the validity, scope and relevance of these solutions and be able to present and defend their conclusions.

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

Type	Hours	Percentage
Hours small group	15,0	12.00
Hours large group	30,0	24.00
Self study	80,0	64.00

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:
a. Workflow
b. ipython vs. CLI
c. Text Editors
d. IDEs
e. Notebook

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

Getting started with Python

Description:
a. Introduction
b. Getting Help
c. Basic types
d. Mutable and in-mutable
e. Assignment operator
f. Controlling execution flow
g. Exception handling

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

Functions and Object Oriented Programming

Description:
a. Defining Functions
b. Input and Output
c. Standard Library
d. Object-oriented programming

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

Introduction to NumPy

Description:
a. Overview
b. Arrays
c. Operations on arrays
d. Advanced arrays (ndarrays)
e. Notes on Performance (\%timeit in ipython)

Full-or-part-time: 2h
Theory classes: 2h

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.
c. Aggregation, slicing, missingness
d. 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 contex
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:
a. Reading form a file.
b. Spark Data Frame.
c. Selection, filters, grouping, sorting.
d. Window operations
c. SQL
d. Accesing and storing methods to a DB, SQL queries.
e. 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:
- Zaharia, M.; Karau, H.; Konwinski, A.; Wendell, P. Learning spark lightning-fast big data analysis. 2015. O'Reilly Media, ISBN 9781449358624.
- Swartz, Jason. Learning scala: practical functional programming for the JVM [on line]. 2014. O'Reilly Media, [Consultation: 28/06/2023]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=1888253. ISBN 9781449367930.
- Langtangen, H.P. A Primer on scientific programming with Python [on line]. Springer, 2011 [Consultation: 28/06/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-3-642-30293-0. ISBN 9783642183652.
- Shapiro, B.E. Scientific computation: Python hacking for math junkies. Sherwood Forest Books, 2015. ISBN 9780692366936.
- Baumer, Benjamin; Kaplan, Daniel; Horton, Nicholas. Modern data science in R. Primera. Boca Raton: CRC, 2017. ISBN 9781498724487.

Complementary:
- Spector, P. Concepts in computing with data (Stat 133, UC Berkeley) [on line]. Berkeley, 2011 [Consultation: 28/06/2023]. Available on: http://www.stat.berkeley.edu/~s133/.