Course guide
270962 - AMA - Advanced Multivariate Analysis

Unit in charge: Barcelona School of Informatics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.
Degree: MASTER'S DEGREE IN DATA SCIENCE (Syllabus 2021). (Optional subject).
Academic year: 2022 ECTS Credits: 6.0 Languages: English

LECTURER

Coordinating lecturer: PEDRO FRANCISCO DELICADO USEROS

Others: Primer quadrimestre:
PEDRO FRANCISCO DELICADO USEROS - 10

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
CE5. Model, design, and implement complex data systems, including data visualization
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CE8. Extract information from structured and unstructured data by considering their multivariate nature.

General:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats

Transversal:
CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CT5. 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.

Basic:
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.
CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

TEACHING METHODOLOGY

There are two weekly 2 hours session.
The first three hours are devoted to the exposition of the theoretical subjects by the teacher.
The last hour is dedicated to implement these contents: Each student has his laptop in class and he or she performs the tasks proposed by the teacher.
Each week ends with an assignment to students who must be delivered in 7 days. The software used will be primarily R.
LEARNING OBJECTIVES OF THE SUBJECT

1. Know the structure of the main unsupervised learning problems.
2. Learn different methods for dimensionality reduction when the standard assumptions in classical Multivariate Analysis are not fulfilled.
3. Learn how to combine dimensionality reduction techniques with prediction algorithms.
4. At the end of the course the student will be able to propose, estimate, interpret and validate non-parametric versions of linear regression models and generalized linear models.
5. At the end of the course the student will know properly how to choose the smoothing parameters which in nonparametric regression models control the trade-off between good fit to the observed sample and good generalization.
6. At the end of the course the students will be able to identify situations in which they can treat their data as functional, to represent them computationally, to apply simple FDA techniques (descriptions, dimensionality reduction, regression) and to visualize the results.
7. At the end of the course, the student will be aware of the need to provide interpretability to machine learning algorithms, he/she will know the most common interpretability techniques, he/she will know how to classify them and what relationships there are between them, and he/she will know how to use them in R and/or Python.

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self study</td>
<td>96.0</td>
<td>64.00</td>
</tr>
<tr>
<td>Hours large group</td>
<td>54.0</td>
<td>36.00</td>
</tr>
</tbody>
</table>

Total learning time: 150 h

CONTENTS

Unsupervised Learning through Advanced Multivariate Analysis

Description:
  i. Principal curves.
  ii. Local Multidimensional Scaling.
  iii. ISOMAP.
  iv. t-Stochastic Neighbor Embedding.
- c. Dimensionality reduction with sparsity.
  i. Matrix decompositions, approximations, and completion. Nuclear norm.
  ii. Sparse Principal Components.
  iii. Applications: (i) Recommender systems. (ii) Estimating causal effects.

Nonparametric regression models

Description:
Functional Data Analysis

Description:
a. Introduction to Functional Data Analysis (FDA). An overview of FDA. Concepts of Functional Analysis useful in FDA.
b. Observed functional data and its computational representation.
   i. Developments in bases of functions.
   ii. Smoothing: Kernel, Local Polynomials, Splines.
   iii. Registration and transformations of functional data.
c. Exploratory analysis of functional data.
   i. Location and dispersion statistics.
   ii. Depth measurements.
   iii. Outliers detection.
d. Dimensionality reduction.
   i. Functional Principal Components.
   ii. Multidimensional Scaling.
e. Regression with explanatory functional data and scalar response.
   i. Scalar response and functional regressor.
   ii. Functional response.
f. Applications: FDA in Demography.

Interpretable Machine Learning

Description:

a. Introduction to interpretability in machine learning.
   i. Transparent models versus black-box models.
   ii. Global methods (relevance of variables) versus local methods (explainability).
b. Interpretability methods for specific models.
   i. Random forests.
   ii. Neural networks.
c. Model-agnostic interpretability methods.
   i. Global methods (Importance of variables through disturbances. Importance based on the Shapley Value. Partial dependency graph. Cumulative local effects graphs.)
   ii. Local methods (LIME: Local interpretable model-agnostic explanations. Local importance based on the Shapley Value. SHAP: SHapley Additive ExPlanations. Break down graphics. ICE: Individual conditional expectation, or ceteris paribus chart.)
d. Interpretability in deep image learning.
   i. Gradient-based methods (Grad-CAM, Saliency maps).
   ii. Perturbation-based methods (LIME for images, SHAP’s DeepExplainer).
Unsupervised Learning through Advanced Multivariate Analysis

**Description:**
Unsupervised Learning through Advanced Multivariate Analysis

**Specific objectives:**
1, 2, 3

**Related competencies:**
- CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
- CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
- CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
- CE5. Model, design, and implement complex data systems, including data visualization
- CE8. Extract information from structured and unstructured data by considering their multivariate nature.
- CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
- CT5. 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.
- CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
- CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

**Full-or-part-time: 44h 24m**
- Theory classes: 12h
- Laboratory classes: 4h
- Self study: 28h 24m
Nonparametric regression models

**Description:**
Nonparametric regression models

**Specific objectives:**
4, 5

**Related competencies:**
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CTS. 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.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.

**Full-or-part-time:** 44h 24m
Theory classes: 12h
Laboratory classes: 4h
Self study: 28h 24m
Functional Data Analysis

Description:
Functional Data Analysis

Specific objectives:
6

Related competencies:
CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
CE10. Identify machine learning and statistical modeling methods to use and apply them rigorously in order to solve a specific data science problem
CE3. Apply data integration methods to solve data science problems in heterogeneous data environments
CE5. Model, design, and implement complex data systems, including data visualization
CE8. Extract information from structured and unstructured data by considering their multivariate nature.
CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
CTS. 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.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.
CB10. Possess and understand knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, often in a research context.
CB6. Ability to apply the acquired knowledge and capacity for solving problems in new or unknown environments within broader (or multidisciplinary) contexts related to their area of study.

Full-or-part-time: 33h 18m
Theory classes: 9h
Laboratory classes: 3h
Self study: 21h 18m
### Interpretable Machine Learning

**Description:**
Interpretable Machine Learning

**Specific objectives:**
7

**Related competencies:**
- CG2. Identify and apply methods of data analysis, knowledge extraction and visualization for data collected in disparate formats
- CE13. Identify the main threats related to ethics and data privacy in a data science project (both in terms of data management and analysis) and develop and implement appropriate measures to mitigate these threats
- CE8. Extract information from structured and unstructured data by considering their multivariate nature.
- CE6. Design the Data Science process and apply scientific methodologies to obtain conclusions about populations and make decisions accordingly, from both structured and unstructured data and potentially stored in heterogeneous formats.
- CT5. 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.

CT4. INFORMATION LITERACY: Capacity for managing the acquisition, the structuring, analysis and visualization of data and information in the field of specialisation, and for critically assessing the results of this management.

CB7. Ability to integrate knowledges and handle the complexity of making judgments based on information which, being incomplete or limited, includes considerations on social and ethical responsibilities linked to the application of their knowledge and judgments.

**Full-or-part-time:** 33h 18m
- Theory classes: 9h
- Laboratory classes: 3h
- Self study: 21h 18m

### GRADING SYSTEM

Homeworks will be assigned during the course. Homework grades will be worth 40% of your course grade.

There will be an exam at the end of the semester and will evaluate the assimilation of the basic concepts on the whole subject. The final exam will have a first short theoretical part (closed books) and a second longer practical part (open books, to be done by the students with their own laptops, with structure similar to homeworks).

Course Grade = 0.4 * Hwk Grade + 0.6 * Exam Grade
BIBLIOGRAPHY

Basic: