The proposed teaching and assessment methodology that appear in the guide may be subject to changes as a result of the restrictions to face-to-face class attendance imposed by the health authorities.

### Prerequisites

Students should have basic knowledge linear algebra, statistical inference and linear models. We also assume the students have programming skills.

Previous experience with R and Python will be helpful.

### Objectives and Contextualisation

The aim of this course is to learn and apply various mathematical and statistical methods related to the discovery of relevant patterns in data sets. Nowadays, huge amounts of data are being generated in many fields, and the goal of this course is to learn how to extract information from such data. When dealing with large datasets, mathematical procedures should be scalable, so we will be concerned with methods that can be scaled and/or parallelized.

### Competences

- Analyse, synthesise, organise and plan projects in the field of study.
- Apply logical/mathematical thinking: the analytic process that involves moving from general principles to particular cases, and the synthetic process that derives a general rule from different examples.
- Apply techniques for solving mathematical models and their real implementation problems.
- Conceive and design efficient solutions, applying computational techniques in order to solve mathematical models of complex systems.
• Formulate, analyse and validate mathematical models of practical problems in different fields.
• Isolate the main difficulty in a complex problem from other, less important issues.
• Solve complex problems by applying the knowledge acquired to areas that are different to the original ones.

Learning Outcomes

1. Analyse, synthesise, organise and plan projects in the field of study.
2. Apply Bayesian statistical techniques to predict the behaviour of certain phenomena.
3. Apply logical/mathematical thinking: the analytic process that involves moving from general principles to particular cases, and the synthetic process that derives a general rule from different examples.
4. Identify real phenomena as models of stochastic processes and extract new information from this to interpret reality.
5. Isolate the main difficulty in a complex problem from other, less important issues.
6. Solve complex problems by applying the knowledge acquired to areas that are different to the original ones.
7. Solve real data analysis problems by identifying them appropriately from the perspective of Bayesian statistics.
8. Use appropriate statistical packages and Bayesian methods solutions to solve specific problems.

Content

Text Mining

• Fundamentals of Text Mining - From text to numbers
• Data cleaning
• Tokenization
• Stemming
• Lemmatization
• POS,NER
• Data chunking

Statistics

• The problem of multiple testing.
• Linear and Generalized linear methods: LASSO/BigLASSO, Ridge Regression and Elastic Nets.
  Feature Selection.
• Gaussian Processes for Machine Learning.

Alternatively,


Topological data analysis

Topology and data, quick review of linear algebra, from points to polyhedra, combinatorial topology, persistence Diagrams and software.

Statistical Learning

Review of basic concepts and the state-of-the-art in statistical learning techniques.

Methodology
Lectures, supervised exercises and autonomous activities directed to perform data analysis projects based on statistical and topological tools.

**Activities**

<table>
<thead>
<tr>
<th>Title</th>
<th>Hours</th>
<th>ECTS</th>
<th>Learning Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type: Directed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homework (problems &amp; computer exercises)</td>
<td>36</td>
<td>1.44</td>
<td>3, 8</td>
</tr>
<tr>
<td>Lectures</td>
<td>38</td>
<td>1.52</td>
<td>1, 4</td>
</tr>
<tr>
<td><strong>Type: Autonomous</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal study, readings</td>
<td>20</td>
<td>0.8</td>
<td>4</td>
</tr>
<tr>
<td>Project</td>
<td>44</td>
<td>1.76</td>
<td>3, 1, 5, 4, 6, 8</td>
</tr>
</tbody>
</table>

**Assessment**

Homework: Completion and presentation of the proposed exercises.

Final Project: The students must choose one of a series of topics provided by the teaching staff and undertake a data project and prepare a talk. This task can be done in group.

Due dates will be announced during the course and will be strict.

**Assessment Activities**

<table>
<thead>
<tr>
<th>Title</th>
<th>Weighting</th>
<th>Hours</th>
<th>ECTS</th>
<th>Learning Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Project</td>
<td>0.4</td>
<td>6</td>
<td>0.24</td>
<td>3, 1, 5, 4, 6, 7, 8</td>
</tr>
<tr>
<td>Homework</td>
<td>0.6</td>
<td>6</td>
<td>0.24</td>
<td>3, 1, 2, 5, 4, 6, 7</td>
</tr>
</tbody>
</table>

**Bibliography**

Basic references


Complementary references


L. Torgo. "Data Mining with R. Learning with Case Studies". Chapman & Hall, Miami. 2010


