270215 - AD - Data Analysis

Coordinating unit: 270 - FIB - Barcelona School of Informatics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research
Academic year: 2019
Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Teaching unit Compulsory)
ECTS credits: 6
Teaching languages: Catalan

Prior skills

Knowledge of basic statistical concepts, descriptive statistics, hypothesis testing. Familiarity with the statistical software R.

Degree competences to which the subject contributes

Basic:
CB2. That the students know how to apply their knowledge to their work or vocation in a professional way and possess the skills that are usually demonstrated through the elaboration and defense of arguments and problem solving within their area of study.
CB4. That the students can transmit information, ideas, problems and solutions to a specialized and non-specialized public.

Specific:
CE1. Skillfully use mathematical concepts and methods that underlie the problems of science and data engineering.
CE2. To be able to program solutions to engineering problems: Design efficient algorithmic solutions to a given computational problem, implement them in the form of a robust, structured and maintainable program, and check the validity of the solution.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CE4. Use current computer systems, including high performance systems, for the process of large volumes of data from the knowledge of its structure, operation and particularities.
CE8. Ability to choose and employ techniques of statistical modeling and data analysis, evaluating the quality of the models, validating and interpreting them.

General:
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG3. Work in multidisciplinary teams and projects related to the processing and exploitation of complex data, interacting fluently with engineers and professionals from other disciplines.
CG4. Identify opportunities for innovative data-driven applications in evolving technological environments.

Transversal:
CT4. Teamwork. Be able to work as a member of an interdisciplinary team, either as a member or conducting management tasks, with the aim of contributing to develop projects with pragmatism and a sense of responsibility, taking commitments taking into account available resources.
CT7. Third language. Know a third language, preferably English, with an adequate oral and written level and in line with the needs of graduates.
Teaching methodology

The learning process is a combination of theoretical explanation and practical application. The theory classes are used to explain the basic scientific contents of the course, whereas the laboratory sessions work on their application to solve real-life problems.

Practicals and project form the basis for working out the transversal competences of the students, related to team-work and public presentation of results. Practicals and project also serve to integrate the different pieces of knowledge of the course.

For hands-on computer training we use the R statistical environment.

Learning objectives of the subject

1. Exploratory Data Analysis
2. Discriminant Analysis with probabilistic hypothesis
3. Multivariate modeling
4. Time series

Study load

<table>
<thead>
<tr>
<th>Total learning time: 150h</th>
<th>Hours large group: 30h</th>
<th>20.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours small group: 30h</td>
<td>20.00%</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Self study: 90h</td>
<td>60.00%</td>
</tr>
</tbody>
</table>
## Content

### Data preprocessing

**Degree competences to which the content contributes:**

**Description:**
Outliers, missing data and transformations

### Principal component analysis

**Degree competences to which the content contributes:**

**Description:**
Multivariate description of a table of continuous variables. Regression with principal components.

### Factor analysis

**Degree competences to which the content contributes:**

**Description:**
The singular value decomposition, biplots, factor analysis

### Multidimensional scaling (MDS)

**Degree competences to which the content contributes:**

**Description:**
Distance measures. Metric multidimensional scaling. Algorithms.

### Cluster analysis

**Degree competences to which the content contributes:**

**Description:**

### Correspondence analysis

**Degree competences to which the content contributes:**

**Description:**
# Discriminant analysis

**Degree competences to which the content contributes:**

**Description:**
- Multivariate normal distribution. Fisher's linear discriminant analysis.

## Univariate time series models

**Degree competences to which the content contributes:**

**Description:**
- Exponential smoothing, ARIMA models

## Intervention analysis

**Degree competences to which the content contributes:**

**Description:**
- Outliers, seasonal effects, intervention analysis.
### Planning of activities

<table>
<thead>
<tr>
<th></th>
<th>Hours: 12h</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data preprocessing</strong></td>
<td></td>
</tr>
<tr>
<td>Description:</td>
<td>Practical on data preprocessing</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hours: 14h</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Principal component analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Description:</td>
<td>Application of principal component analysis in practical data analysis</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hours: 9h</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor analysis</strong></td>
<td></td>
</tr>
<tr>
<td>Description:</td>
<td>Practical data analysis using the method</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hours: 8h</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multidimensional scaling</strong></td>
<td></td>
</tr>
<tr>
<td>Description:</td>
<td></td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>1</td>
</tr>
</tbody>
</table>
**270215 - AD - Data Analysis**

<table>
<thead>
<tr>
<th>Description:</th>
<th>Clustering</th>
<th>Hours: 12h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Analysis of distance matrices with this method</td>
<td>Theory classes: 4h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Practical classes: 0h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Laboratory classes: 4h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self study: 4h</td>
</tr>
</tbody>
</table>

| Specific objectives: | 1 |

<table>
<thead>
<tr>
<th>Description:</th>
<th>Correspondence Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Application of the method to quantitative data matrices</td>
</tr>
</tbody>
</table>

| Specific objectives: | 2 |

<table>
<thead>
<tr>
<th>Description:</th>
<th>Discriminant Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Application of the method to empirical data sets</td>
</tr>
</tbody>
</table>

| Specific objectives: | 2 |
### Univariate time series models

**Description:**
Fitting time series models to data sets on the computer

**Specific objectives:**
- 4

**Hours:** 14h
- Theory classes: 4h
- Practical classes: 0h
- Laboratory classes: 4h
- Guided activities: 0h
- Self study: 6h

### Intervention analysis

**Description:**
Application of intervention analysis to real data sets

**Specific objectives:**
- 4

**Hours:** 9h
- Theory classes: 2h
- Practical classes: 0h
- Laboratory classes: 3h
- Guided activities: 0h
- Self study: 4h

### Practical on exploratory data analysis

**Description:**
Student do an exploratory analysis of a data set and hand in a questionnaire about it.

**Specific objectives:**
- 1, 2, 3, 4

**Hours:** 18h
- Guided activities: 3h
- Self study: 15h

### Project

**Description:**
Students realize, in couples, a complete multivariate study of a certain dataset using the techniques they studied during the course, and hand in a written report about it.

**Specific objectives:**
- 1, 2, 3, 4

**Hours:** 16h
- Guided activities: 3h
- Self study: 13h
Exam concerning basic concepts

<table>
<thead>
<tr>
<th>Description:</th>
<th>Hours: 16h 30m</th>
</tr>
</thead>
<tbody>
<tr>
<td>There are two exams related to the theoretical concepts of the course.</td>
<td>Guided activities: 2h</td>
</tr>
<tr>
<td>Specific objectives:</td>
<td>Self study: 14h 30m</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td></td>
</tr>
</tbody>
</table>

Qualification system

The student's final grade for the course is based on grades obtained for weekly homework assignments (25%), a partial exam half-way the course (25%), a final exam covering the second half of the course (25%) and a project (25%).

Each weekly assignments consists of resolving a questionnaire. These assignments aim at consolidating knowledge of the techniques exposed in the theoretical sessions. The assignments require analysis of datasets in the statistical environment R.

A project is carried out by a group of two students, and students have to show they can resolve problems with the techniques they have learned during the course. Each group hands in a written report about their project at the end of the course.

The two exams will be programmed according to the calendar of the faculty, and evaluate if students have assimilated the basic concepts of the material of the course.

For the resit exam, the student can choose to do a re-examination of only the first partial (25%), or of only the second partial (25%), or of both partials (50%). The re-evaluation thus represents at most 50% of the final course grade.
Bibliography

Basic:


Complementary:


