270652 - MVA - Multivariate Analysis

Coordinating unit: 270 - FIB - Barcelona School of Informatics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research
Academic year: 2019
Degree: MASTER'S DEGREE IN INNOVATION AND RESEARCH IN INFORMATICS (Syllabus 2012). (Teaching unit Optional)
ECTS credits: 6  Teaching languages: English

Prior skills
The course implies having previously done a basic course in statistics, programming and mathematics; in particular having acquired the following concepts:
- Average, covariance and correlation matrix.
- Hypothesis Test
- Matrix algebra, eigenvalues and eigenvectors.
- Programming algorithms.
- Multiple linear-regression

Degree competences to which the subject contributes

Specific:
CEC1. Ability to apply scientific methodologies in the study and analysis of phenomena and systems in any field of Information Technology as well as in the conception, design and implementation of innovative and original computing solutions.
CEC2. Capacity for mathematical modelling, calculation and experimental design in engineering technology centres and business, particularly in research and innovation in all areas of Computer Science.

General:
CG1. Capability to apply the scientific method to study and analyse of phenomena and systems in any area of Computer Science, and in the conception, design and implementation of innovative and original solutions.
CG3. Capacity for mathematical modeling, calculation and experimental designing in technology and companies engineering centers, particularly in research and innovation in all areas of Computer Science.

Transversal:
CTR4. INFORMATION LITERACY: Capability to manage the acquisition, structuring, analysis and visualization of data and information in the area of informatics engineering, and critically assess the results of this effort.
CTR6. REASONING: Capacity for critical, logical and mathematical reasoning. Capability to solve problems in their area of study. Capacity for abstraction: the capability to create and use models that reflect real situations. Capability to design and implement simple experiments, and analyze and interpret their results. Capacity for analysis, synthesis and evaluation.

Teaching methodology
The aim of the course is to give the statistical foundations for data mining. Learning is done through a combination of theoretical explanation and its application to a real case. The lectures will develop the necessary scientific knowledge, while lab classes will be its application to solving problems of data mining. These problems constitute the practices of the subject, which will be developed in part during laboratory classes. The implementation of practices foster generic skills related to teamwork and presentation of results and serve to integrate different knowledge of the subject. The software used will be primarily R.

Learning objectives of the subject
1. Visual representation of the data
2. Multivariate description of data
3. Multivariate inference
4. Classification of new individuals

**Study load**

<table>
<thead>
<tr>
<th>Total learning time: 150h</th>
<th>Hours large group: 26h</th>
<th>17.33%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours medium group: 0h</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Hours small group: 26h</td>
<td>17.33%</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 2h</td>
<td>1.33%</td>
</tr>
<tr>
<td></td>
<td>Self study: 96h</td>
<td>64.00%</td>
</tr>
</tbody>
</table>
## Content

### Multivariate Data Analysis

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

**Description:**

### Principal Component Analysis

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

**Description:**

### Singular Value Decomposition. Biplots

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

**Description:**
Simultaneous representation of the rows and columns of a data table.

### Factor Analysis

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

**Description:**
Latents constructs. Measurement model.

### Multidimensional Scaling

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

**Description:**
Visualisation of link matrices.

### Correspondence Analysis

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

**Description:**
Analysis of frequency data.
## Multiple Correspondence Analysis

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

**Description:**
Analysis of categorical data

## Hierarchical clustering

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

**Description:**
Synthesis of the represented information. Consolidation of the partition

## Multivariate normal distribution

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

**Description:**
Definition and properties

## Sampling distributions of the normal multivariate distribution

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

**Description:**
Inferences respect to the covariance matrix. Inferences respect to the centroid of the distribution. Wishart distribution. T2 of Hotelling, Wilks lambda.

## Discriminant Analysis

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

**Description:**
With the assumption of multivariate normal distribution. Linear discriminant analysis. Quadratic discriminant analysis.

## Naive Bayes

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

**Description:**
Simplifying the linear discriminant analysis
## Discriminant analysis without probabilistic assumptions

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

**Description:**
K nearest neighbor classifier

## Decision trees

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

**Description:**
Classification and regression trees

## Association rules

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

**Description:**
Apriori algorithm
### Planning of activities

| Multivariate Data Analysis | Hours: 6h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 2h |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific objectives:</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

| Principal Component Analysis | Hours: 11h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 7h |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific objectives:</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

| Singular Value Decomposition. Biplots | Hours: 7h  
Theory classes: 1h  
Practical classes: 0h  
Laboratory classes: 1h  
Guided activities: 0h  
Self study: 5h |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific objectives:</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

| Factor Analysis | Hours: 7h  
Theory classes: 1h  
Practical classes: 0h  
Laboratory classes: 1h  
Guided activities: 0h  
Self study: 5h |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific objectives:</td>
<td>1, 2</td>
</tr>
</tbody>
</table>
## Multidimensional Scaling

**Specific objectives:**
1, 2

**Hours:** 9h  
Theory classes: 2h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

## Correspondence Analysis

**Specific objectives:**
1, 2

**Hours:** 9h  
Theory classes: 2h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

## Multiple Correspondence Analysis

**Specific objectives:**
1, 2

**Hours:** 9h  
Theory classes: 2h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

## Hierarchical Clustering

**Specific objectives:**
1, 2

**Hours:** 9h  
Theory classes: 2h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

---

**Theory classes:** 2h  
**Practical classes:** 0h  
**Laboratory classes:** 2h  
**Guided activities:** 0h  
**Self study:** 5h  
**Total Hours:** 9h
### Multivariate normal distribution

**Hours:** 9h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

**Specific objectives:**  
3

### Sampling distributions.

**Hours:** 9h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

**Specific objectives:**  
3

### Multivariate Statistical Tests

**Hours:** 9h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

**Specific objectives:**  
4

### Hotelling's T2

**Hours:** 9h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

**Specific objectives:**  
4
## Decision trees

**Hours:** 9h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

**Specific objectives:**  
4

## Association rules

**Hours:** 9h  
Theory classes: 2h  
Practical classes: 0h  
Laboratory classes: 2h  
Guided activities: 0h  
Self study: 5h

**Specific objectives:**  
4

## Practices on Exploratory Multivariate Analysis

**Hours:** 10h  
Guided activities: 1h  
Self study: 9h

## Final Practical Work

**Hours:** 9h  
Guided activities: 1h  
Self study: 8h

## Quiz

**Hours:** 10h  
Guided activities: 0h  
Self study: 10h
The course evaluation will be based on the marks obtained in practical exercises conducted during the course, an examination grade and the grade obtained in the final practice. Each practice will lead to the drafting of the relevant report writing and may be made jointly, up to a maximum of two students per group.

The exercises conducted throughout the course aim to consolidate the learning of multivariate techniques. The final practice is that students show their maturity to solve a real problem using multivariate visualization techniques, "clustering" interpretation and prediction. Students will choose between different alternatives to solve the problem. This practice will be presented and publicly defended the student must answer any questions about the theoretical models and methods used in the solution. Practices are conducted using the software R.

The written test will be held the last day of class and evaluate the assimilation of the basic concepts of the subject. While the presentation of the second practice will be done during the examination period.

The exercises are weighted 30%, examination 40% respectively and practice final 30%.
Bibliography

Basic:


Complementary:


Others resources:

Hyperlink

https://cran.r-project.org/

http://www.dtmvic.com/05_SoftwareE.html