270107 - MD - Data Mining

Coordinating unit: 270 - FIB - Barcelona School of Informatics
Teaching unit: 723 - CS - Department of Computer Science
715 - EIO - Department of Statistics and Operations Research

Academic year: 2019
Degree: BACHELOR'S DEGREE IN INFORMATICS ENGINEERING (Syllabus 2010). (Teaching unit Optional)
ECTS credits: 6  Teaching languages: English

Teaching staff
Coordinator: - Karina Gibert Oliveras (karina.gibert@upc.edu)
- Mario Martín Muñoz (mmartin@cs.upc.edu)

Others: - Beatriz Sevilla Villanueva (beatriz.sevilla@upc.edu)

Prior skills
Foundations of probability and statistics. Basic Programming in R

Requirements
- Prerequisite PE
- Prerequisite PRO2

Degree competences to which the subject contributes

Specific:
CS12.2. To conceive, deploy, organize and manage computer systems and services, in business or institutional contexts, to improve the business processes; to take responsibility and lead the start-up and the continuous improvement; to evaluate its economic and social impact.
CS12.3. To demonstrate knowledge and application capacity of extraction and knowledge management systems .
CS12.6. To demonstrate knowledge and capacity to apply decision support and business intelligence systems.

Generical:
G3. THIRD LANGUAGE: to know the English language in a correct oral and written level, and accordingly to the needs of the graduates in Informatics Engineering. Capacity to work in a multidisciplinary group and in a multi-language environment and to communicate, orally and in a written way, knowledge, procedures, results and ideas related to the technical informatics engineer profession.
G9. PROPER THINKING HABITS: capacity of critical, logical and mathematical reasoning. Capacity to solve problems in her study area. Abstraction capacity; capacity to create and use models that reflect real situations. Capacity to design and perform simple experiments and analyse and interpret its results. Analysis, synthesis and evaluation capacity.
Teaching methodology

The learning methodology will consist in the analysis of case studies concerning complex data sets from real problems. From these problems the body of necessary scientific knowledge will be introduced. The theoretical and practical lessons are interleaved such that programming and/or integration of data mining functions enhance the assimilation of the various concepts explained. The open programming environment R will be used in the laboratory.

The laboratory classes will be devoted to solving problems related to the knowledge provided in the theory classes and to the resolution by the students of a similar problem. This problem may include the resolution of very brief conceptual questions and will be delivered for its evaluation. Finally, the students must complete two full practical works, a statistical modeling problem and a modelling problem of the "scientific", "transaction" or "marketing" kind (only one of them must be chosen by the student). This last practical work will be presented orally to the whole class.

Learning objectives of the subject

1. Knowing the types of the main problems of Data Mining
2. Data quality assessment and preprocessing
3. Problem solving: identify the statistical and/or machine learning techniques more appropriate to solve the problem
4. Implement simple learning algorithms
5. Validation of results
6. Presentation of results in a professional environment for decision making

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 medium group: 0h</td>
<td>0.00%</td>
</tr>
<tr>
<td></td>
<td>Hours small group: 30h</td>
<td>20.00%</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 6h</td>
<td>4.00%</td>
</tr>
<tr>
<td></td>
<td>Self study: 84h</td>
<td>56.00%</td>
</tr>
</tbody>
</table>
# Content

## Introduction to Data Mining

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

**Description:**
Statistical modeling and types of problems: analysis of binary data ("transactions"), analysis of scientific data and analysis of data from enterprises

## Visualization and dimensionality reduction

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

**Description:**
Feature selection and extraction. Visualization of multivariate data.

## Clustering

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

**Description:**
Direct partitioning methods, hierarchical methods and expectation maximization

## Predictive Methods

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

**Description:**
Regressió lineal múltiple i generalitzada. Regressió Logística. Xarxes Neuronals

## Decision Trees

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

**Description:**
Classification and regression trees (CART).

## Validation protocols and data resampling

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

**Description:**
Holdout, cross-validation and the bootstrap
## Generation of association rules

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

**Description:**
A-priori and Eclat algorithms.

## Discriminant Analysis

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

**Description:**
Bayesian decision theory. LDA and QDA Discriminant Analysis and Naïve Bayes

## Non parametric discrimination

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

**Description:**
Nearest neighbours

## Regression Shrinkage and Variable Selection

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

**Description:**
Regularized linear regression. LASSO and the Elastic Net methods.

## Formal concept analysis

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

**Description:**
Formal method for pattern finding

## Preprocessing

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

**Description:**
A
**Bagging i ensemble methods**

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

**Description:**
Bagging i ensemble methods
### Planning of activities

<table>
<thead>
<tr>
<th>Development Unit 1</th>
<th>Hours: 2h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theory classes: 2h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 0h</td>
</tr>
</tbody>
</table>

**Specific objectives:**

1

<table>
<thead>
<tr>
<th>A review of R language</th>
<th>Hours: 6h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theory classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 6h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 0h</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Development of item 2</th>
<th>Hours: 16h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theory classes: 4h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 4h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 8h</td>
</tr>
</tbody>
</table>

**Specific objectives:**

2

<table>
<thead>
<tr>
<th>Development of item 3</th>
<th>Hours: 9h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theory classes: 3h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 2h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 4h</td>
</tr>
</tbody>
</table>

**Specific objectives:**

2

<table>
<thead>
<tr>
<th>Development of Item 4</th>
<th>Hours: 11h</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Theory classes: 3h</td>
</tr>
<tr>
<td></td>
<td>Practical classes: 0h</td>
</tr>
<tr>
<td></td>
<td>Laboratory classes: 4h</td>
</tr>
<tr>
<td></td>
<td>Guided activities: 0h</td>
</tr>
<tr>
<td></td>
<td>Self study: 4h</td>
</tr>
</tbody>
</table>
## Specific objectives:

1. **Development of item 5**
   - **Hours:** 9h
   - Theory classes: 3h
   - Practical classes: 0h
   - Laboratory classes: 2h
   - Guided activities: 0h
   - Self study: 4h

2. **Development of Item 6**
   - **Hours:** 7h
   - Theory classes: 3h
   - Practical classes: 0h
   - Laboratory classes: 0h
   - Guided activities: 0h
   - Self study: 4h

3. **Development of Item 7**
   - **Hours:** 9h
   - Theory classes: 3h
   - Practical classes: 0h
   - Laboratory classes: 2h
   - Guided activities: 0h
   - Self study: 4h

4. **Development of Item 8**
   - **Hours:** 11h
   - Theory classes: 3h
   - Practical classes: 0h
   - Laboratory classes: 4h
   - Guided activities: 0h
   - Self study: 4h
### Development of Item 9

**Specific objectives:**
- 2

**Hours:** 11h  
- Theory classes: 3h  
- Practical classes: 0h  
- Laboratory classes: 2h  
- Guided activities: 0h  
- Self study: 6h

### Development of Item 10

**Specific objectives:**
- 6

**Hours:** 13h  
- Theory classes: 3h  
- Practical classes: 0h  
- Laboratory classes: 4h  
- Guided activities: 0h  
- Self study: 6h

### Practice 1

**Specific objectives:**
- 2, 3, 5, 6

**Hours:** 23h  
- Guided activities: 3h  
- Self study: 20h

### Practice 2

**Specific objectives:**
- 3, 5, 6, 7

**Hours:** 23h  
- Guided activities: 3h  
- Self study: 20h
The evaluation of the course will be based on the grade obtained in the exercises developed during the lab sessions. On the other hand, there will be two practical works. For each practical work, the student will deliver the corresponding written report. Finally, at the end of the course, the students must present orally the second practical work.

The student will be required to show the necessary reasoning as well as English skills. These skills will be evaluated using the corresponding rubrics.

The overall laboratory grade is the average of the grades obtained for the exercises developed out of the laboratory sessions.

The final mark will be obtained as follows:

\[
\text{Lab} = \text{overall laboratory grade} \\
\text{PR1} = \text{grade for the first practical work} \\
\text{PR2} = \text{grade for the second practical work} \\
\text{Final grade} = 0.2*\text{Labo} + 0.4*\text{Pr1} + 0.4*\text{Pr2}
\]

In both practical works (counting 40% each), 35% corresponds to the technical correction and 5% corresponds to the 'reasoning' generic competence, so that this competence gets an overall weight of 10% of the final grade.

**Bibliography**

**Others resources:**

Hyperlink
- [https://cran.r-project.org/](https://cran.r-project.org/)
- [http://www.kdnuggets.com/](http://www.kdnuggets.com/)
- [https://www.cs.waikato.ac.nz/ml/weka/](https://www.cs.waikato.ac.nz/ml/weka/)