200644 - APE - Statistical Learning

Coordinating unit: 200 - FME - School of Mathematics and Statistics
Teaching unit: 1004 - UB - (ENG)Universitat de Barcelona
Academic year: 2017
Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Teaching unit Optional)
ECTS credits: 5
Teaching languages: Spanish

Teaching staff
Coordinator: PEDRO FRANCISCO DELICADO USEROS
Others: Segon quadrimestre:
PEDRO FRANCISCO DELICADO USEROS - A
FERRAN REVERTER COMES - A
ESTEBAN VEGAS LOZANO - A

Prior skills
Familiarity with the foundations of calculus in one and more variables. Intermediate studies in probability and inference. Skills using the R environment for statistical computing and programming. Any good online R course may help, like http://www.ub.edu/stat/docencia/EADB/Curso%20basico%20de%20R.htm.

Requirements
"Fundamentos de Inferencia Estadística" o "Inferencia Estadística Avanzada"
"Computación en Estadística y en Optimización"

Degree competences to which the subject contributes
Specific:
MESIO-CE2. CE-2. Ability to master the proper terminology in a field that is necessary to apply statistical or operations research models and methods to solve real problems.
MESIO-CE3. CE-3. Ability to formulate, analyze and validate models applicable to practical problems. Ability to select the method and / or statistical or operations research technique more appropriate to apply this model to the situation or problem.
MESIO-CE6. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.
MESIO-CE8. CE-8. Ability to discuss the validity, scope and relevance of these solutions and be able to present and defend their conclusions.
MESIO-CE4. CE-4. Ability to use different inference procedures to answer questions, identifying the properties of different estimation methods and their advantages and disadvantages, tailored to a specific situation and a specific context.

Transversal:
CT1a. ENTREPRENEURSHIP AND INNOVATION: Being aware of and understanding how companies are organised and the principles that govern their activity, and being able to understand employment regulations and the relationships between planning, industrial and commercial strategies, quality and profit.
CT3. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are
200644 - APE - Statistical Learning

To know the structure of supervised and unsupervised learning problems.
To be able to fit a multiple linear regression model, and also a glm, using penalized version of the standard ordinary least squares (OLS) and maximum likelihood estimators.
To know the essential common characteristics of non-parametric regression estimators (bias-variance trade-off, smoothing parameter choice, effective number of parameters, etc.) and the details of three of them: local polynomial regression, spline smoothing, generalized additive models (GAM).
To know the principal Tree-based Methods and be able to apply these methods in real data sets.
To understand the essentials of Support Vector Machines, and to acquire the necessary abilities for applying these methods in diverse situations with a practical interest.
To understand the fundamentals of the of Artificial Neural Networks, and to acquire the necessary abilities to apply them
To know the principal cross-validation procedures for assessing model accuracy.

Teaching methodology

On-Site Learning
On-site learning is organized into theoretical-practical sessions. All the sessions will be held in the computer lab, combining a 50% of expository classes and other 50% of guided practice and workshops.
In the expository part of the sessions, the theoretical aspects are presented and discussed, accompanied by practical examples using slides that will be provided previously to the student.
The fundamental work environment of the practical sessions will be R, of which an intermediate knowledge is presumed (use of the environment and basic programming).

Off-Site Learning
Off-site learning will consist of the study and resolution of theoretical and practical problems that the student should turn in throughout the course.
Concretely, the planned activities are:
- Study of the learning materials, before and/or after each on-site session.
- Detailed analysis of diverse data sets. It will be attempted that each data set serves as a basis for a case study in diverse methods.
- The completion of theoretical and practical exercises on the studied methods. The practical exercises will require completion of programming tasks in R.

Learning objectives of the subject

To know the structure of supervised and unsupervised learning problems.
To be able to fit a multiple linear regression model, and also a glm, using penalized version of the standard ordinary least squares (OLS) and maximum likelihood estimators.
To know the essential common characteristics of non-parametric regression estimators (bias-variance trade-off, smoothing parameter choice, effective number of parameters, etc.) and the details of three of them: local polynomial regression, spline smoothing, generalized additive models (GAM).
To know the principal Tree-based Methods and be able to apply these methods in real data sets.
To understand the essentials of Support Vector Machines, and to acquire the necessary abilities for applying these methods in diverse situations with a practical interest.
To understand the fundamentals of the of Artificial Neural Networks, and to acquire the necessary abilities to apply them
To know the principal cross-validation procedures for assessing model accuracy.

Study load

<table>
<thead>
<tr>
<th>Total learning time: 125h</th>
<th>Hours large group: 30h</th>
<th>24.00%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hours small group: 15h</td>
<td>12.00%</td>
</tr>
<tr>
<td></td>
<td>Self study: 80h</td>
<td>64.00%</td>
</tr>
</tbody>
</table>
## Content

<table>
<thead>
<tr>
<th>Topic</th>
<th>Learning time:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction to statistical learning</strong></td>
<td>1h 30m</td>
</tr>
<tr>
<td><strong>Penalized regression estimators: Ridge regression and Lasso</strong></td>
<td>6h</td>
</tr>
<tr>
<td><strong>Generalized Additive Models</strong></td>
<td>13h 30m</td>
</tr>
<tr>
<td><strong>Tree-based Methods</strong></td>
<td>6h</td>
</tr>
</tbody>
</table>

### Introduction to statistical learning

**Description:**
1. Supervised and unsupervised learning.

### Penalized regression estimators: Ridge regression and Lasso

**Description:**
1. Ridge regression.
2. Cross-validation.
3. Lasso estimator in the multiple linear regression model. Cyclical coordinate optimization.
4. Lasso estimator in the GLM.

### Generalized Additive Models

**Description:**
1. Introduction to nonparametric modeling.
2. Local polynomial regression. The bias-variance trade-off. Linear smoothers. Choosing the smoothing parameter.

### Tree-based Methods

**Description:**
## Artificial Neural Networks

<table>
<thead>
<tr>
<th>Learning time: 9h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory classes: 9h</td>
</tr>
</tbody>
</table>

**Description:**
1. Feed-Forward Network Functions.
2. Network Training.
3. Error Backpropagation.

## Support Vector Machine

<table>
<thead>
<tr>
<th>Learning time: 6h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theory classes: 6h</td>
</tr>
</tbody>
</table>

**Description:**
1. Maximum Margin Classifier.

## Qualification system

It is based on two parts:
1) Practical exercises done through the course: 50%
2) Final exam: 50%
Bibliography

Basic:


Complementary:


Others resources:

ATENEA