Course guide
200644 - APE - Statistical Learning

Last modified: 15/06/2023

Unit in charge: School of Mathematics and Statistics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.
1004 - UB - (ENG)Universitat de Barcelona.

Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Optional subject).

Academic year: 2023 **ECTS Credits:** 5.0 **Languages:** Spanish

LECTURER

Coordinating lecturer: PEDRO FRANCISCO DELICADO USEROS

Others: Segon quadrimestre:
PEDRO FRANCISCO DELICADO USEROS - A
ÀLEX SÁNCHEZ PLA - 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.

REQUIREMENTS

"Fundamentos de Inferencia Estadística" o "Inferencia Estadística Avanzada"
"Statistical Software: R and SAS"

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-CE9. CE-9. Ability to implement statistical and operations research algorithms.
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 available.

TEACHING METHODOLOGY

Learning is organized into theoretical-practical sessions with the instructors. All the sessions combine 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 students.
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).

Autonomous learning will consist of the study and resolution of theoretical and practical problems that the student should turn in throughout the course.
Specifocally, the planned activities are:
- Study of the learning materials, before and/or after each session with the instructors.
- Detailed analysis of diverse data sets. It will be attempted that each data set serves as a basis for a case study in several 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 fundamentals of the of Artificial Neural Networks (including deep-learning models and convolutional neural networks), and to acquire the necessary abilities to apply them.
To know the principal cross-validation procedures for assessing model accuracy.

STUDY LOAD

Type	Hours	Percentage
Self study	80,0	64.00
Hours large group	30,0	24.00
Hours small group	15,0	12.00

Total learning time: 125 h

CONTENTS

Introduction to statistical learning

Description:
1. Supervised and unsupervised learning.
2. Machine learning and statistical learning.

Full-or-part-time: 1h 30m
Theory classes: 1h
Laboratory classes: 0h 30m

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.
5. Comparing classification rules. ROC curve.

Full-or-part-time: 6h
Theory classes: 4h
Laboratory classes: 2h

Non-parametric regression. Generalized Additive Models

Description:
1. Introduction to nonparametric modeling.
2. Local polynomial regression. The bias-variance trade-off. Linear smoothers. Choosing the smoothing parameter.
3. Nonparametric regression with binary response. Generalized nonparametric regression model. Estimation by maximum local likelihood.
4. Spline smoothing. Penalized least squares nonparametric regression. Cubic splines, interpolation and smoothing. B-splines. Fitting generalized nonparametric regression models with splines.
5. Generalized Additive Models (GAM). Multiple nonparametric regression. The curse of dimensionality. Additive models. Generalized additive models.

Full-or-part-time: 15h
Theory classes: 10h
Laboratory classes: 5h

Tree-based Methods

Description:
1. The Basics of Decision Trees. Regression Trees. Classification Trees.
2. Ensemble Learning. Bagging. Random Forests. Boosting.

Full-or-part-time: 10h 30m
Theory classes: 7h
Laboratory classes: 3h 30m

Artificial Neural Networks

Description:
1. Feed-Forward Network Functions.
2. Network Training.
3. Error Backpropagation.
4. Deep Learning models.
5. Convolutional Neural Networks.

Full-or-part-time: 12h
Theory classes: 8h
Laboratory classes: 4h

GRADING SYSTEM

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

BIBLIOGRAPHY

Basic:
- Hastie, Trevor; Tibshirani, Robert; Wainwright, Martin. Statistical learning with sparsity: the lasso and generalizations [on line]. Boca Raton, FL: Chapman & Hall/CRC, 2015 [Consultation: 04/07/2023]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=4087701. ISBN 9781498712163.
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome. The Elements of statistical learning : data mining, inference, and prediction [on line]. 2nd ed. New York [etc.]: Springer, cop. 2009 [Consultation: 04/07/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-0-387-84858-7. ISBN 9780387848570.
- Lantz, Brett. Machine learning with R : discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R [on line]. 2nd ed. Birmingham: Packt Pub, 2015 [Consultation: 04/07/2023]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=2122139. ISBN 9781784393908.
- James, Gareth. An Introduction to statistical learning : with applications in R [on line]. New York: Springer, 2013 [Consultation: 04/07/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/978-1-4614-7138-7. ISBN 9781461471370.
- Bowman, A. W; Azzalini, Adelchi. Applied smoothing techniques for data analysis : the Kernel approach with S-Plus illustrations. Oxford: Clarendon Press, 1997. ISBN 0198523963.
- Wood, Simon N. Generalized additive models : an introduction with R. Boca Raton, Fla. [etc.]: Chapman & Hall/CRC, 2006. ISBN 9781584884743.
- Chollet, François; Allaire, J. J. Deep Learning with R. Manning Publications, 2018. ISBN 9781617295546.

Complementary:
- Wasserman, Larry. All of nonparametric statistics [on line]. New York: Springer, 2006 [Consultation: 04/07/2023]. Available on: https://link-springer-com.recursos.biblioteca.upc.edu/book/10.1007/0-387-30623-4. ISBN 9780387251455.
- Haykin, Simon S. Neural networks and learning machines. 3rd. Upper Saddle River: Prentice Hall, cop. 2009. ISBN 9780131471399.
- Bishop, Christopher M. Pattern recognition and machine learning. New York: Springer, cop. 2006. ISBN 9780387310732.

RESOURCES

Other resources:
ATENEA