Statistical methods in data mining

Academic Year: (2023 / 2024) **Review date: 28-04-2023**

Department assigned to the subject: Statistics Department

Coordinating teacher: MUÑOZ GARCIA, ALBERTO

Type: Compulsory ECTS Credits : 6.0

Year : 4 Semester : 1

REQUIREMENTS (SUBJECTS THAT ARE ASSUMED TO BE KNOWN)

Regression Methods and Multivariate Analysis, third course. Knowledge of R statistical software.

OBJECTIVES

1. To know and use advanced statistical techniques, with last generation software support.
2. To extract and analyze information from large data sets.

1. Ability of information analysis and synthesis.
2. Modelization and resolution of practical problems in Data Mining.
3. Oral and written communication skills.

DESCRIPTION OF CONTENTS: PROGRAMME

1. Introduction Tidyverse
   1.1 Data wrangling
   2.2 Data Visualization: ggplot2
   2.3 Grouping and summarizing.

2. Text Mining.
   2.1 Main concepts.
   2.2 Word clouds.
   2.3 Term by document matrix.
   2.4 R implementations and applications.

3. Data visualization. Metric Multidimensional Scaling, Correspondence Analysis, Biplots.
   3.1 Metric Multidimensional Scaling.
   3.2 Biplots.
   3.2 Perceptual Mappings.

4. Cluster Analysis. Hierarchical Methods, k-means and mixture models.
   4.1 Bottom up hierarchical clustering algorithms.
   4.2 k-means and related algorithms.

5. Information Theory and classification trees.
   5.1 Information theory.
   5.2 Classification trees algorithms.
   5.3 Real case: credit scoring.
   5.4 Case studies.

6. Association Rules.
   6.1 Main concepts and algorithms.
   6.2 Complete example with R code.
   6.3 Case studies.

7. Deep Learning.
   7.1 Support Vector Machines.
   7.2 Neural Networks for classification.
   7.3 Neural Networks for regression.

8. Case Studies.

8.1 Comprehensive real cases involving all the studied techniques.

LEARNING ACTIVITIES AND METHODOLOGY

Theory (4 ECTS). Theory clases with lessons available in Web. Practice (2 ECTS). Problem and case studies solving. Computational practices in computer rooms. Oral presentations and debates.

ASSESSMENT SYSTEM

50%: Final exam oriented to practice, consisting in a practical individual data analysis case.
10%: Continuous evaluation (*).
40%: Handing a final project.

(*) Continuous evaluation consists of handing several case studies (homework) along the course.

% end-of-term-examination:	50
% of continuous assessment (assigments, laboratory, practicals…):	50

BASIC BIBLIOGRAPHY

- A.J. Izenman Modern Multivariate Statistical Techniques, Springer, 2008
- E. Alpaydin Introduction to Machine Learning, 2nd Edition, MIT Press, 2010
- X. Wu The top ten algorithms in data mining, Chapman &Hall /CRC, 2009

ADDITIONAL BIBLIOGRAPHY

- I.H. Witten , E. Frank, M.A. Hall Data Mining. Practical Machine Learning Tools and Techniques, 3d Edition, Morgan Kaufmann, 2011
- John M. Chambers Software for Data Analysis. Programming with R., Springer, 2008
- Luis Torgo Data Mining with R, Chapman & Hall/CRC, 2001
- W.J. Braun, D.J. Murdoch A first course in statistical programming with R, Cambridge University Press, 2007

BASIC ELECTRONIC RESOURCES

- Rbloggers . Blogs de R: https://www.r-bloggers.com/