INTRODUZIONE AL DATA MINING

INF/01 - 9 CFU - 1° Semester

Teaching Staff

GIOVANNI MICALE
Email: gmicale@dmi.unict.it
Office: Dipartimento di Matematica e Informatica, Blocco III, Stanza 40
Phone: 0957383071
Office Hours: Mercoledì - 11-13

LEARNING OBJECTIVES

General teaching training objectives in terms of expected learning outcomes.

Knowledge and understanding: The course aims to give the knowledge and basic and advanced skills to the analysis of data.

Applying knowledge and understanding: the student will acquire knowledge about the models and algorithms for analyzing data such as: mining high support, recommendation systems, search for similarities in high dimension, networks analysis, neural networks, classification and clustering.

Making judgments: Through concrete examples and case studies, the student will be able to independently develop solutions to specific problems related to data analysis.

Communication skills: the student will acquire the necessary communication skills and expressive appropriateness in the use of technical language in the general area of data analysis.

Learning skills: The course aims to provide students with the necessary theoretical and practical methods to deal independently and solve new problems that may arise during a work activity. For this purpose, different topics will be covered in class by involving students in the search for possible solutions to real problems, using benchmarks available in the literature.

COURSE STRUCTURE

Lectures.

Should teaching be carried out in mixed mode or remotely, it may be necessary to introduce changes with respect to previous statements, in line with the programme planned and outlined in the syllabus.

DETAILED COURSE CONTENT

The course includes a theoretical part, in which the main Data Mining problems will be explained, and a practical part, in which we will introduce the R programming language and we will show how to solve the illustrated data mining problems using R. The two parts of the course will be carried on in parallel.

The following topics will be covered:

- Introduction to Data Mining
- Mention on probability theory
- R programming language
- High support Data Mining (apriori algorithm, frequent itemsets, association rules)
- Classification (decision trees, SVM, bayesian classifiers, lazy classifiers, rules extractors)
- Clustering (hierarchical, k-means, density-based)
- Recommendation systems
- Markov chains and HMM
- Introduction to Networks (Centrality measures, Clustering coefficient)
- Network random models
- Graph matching
- Graph mining
- Neural Networks (Feed-Forward, Convolutional, Recurrent, Long-Short Term Memory)

Concerning R language, we will show base functions for data analysis as well as several packages for data mining, such as "caret" (for classification), "igraph" (for network analysis and visualization) and "keras" (for building neural networks).

TEXTBOOK INFORMATION

For the theoretical description of data mining problems, we will mainly refer to different chapters of the following book:

- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman (http://www.mmds.org)

Other suggested textbooks for Data Mining are:

- Data Mining: Concepts and Techniques, Jiawei Han and Micheline Kamber, The Morgan Kaufmann Series in Data Management Systems
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie,Robert Tibshirani, Jerome Friedman, Springer

Concerning probability theory, the suggested textbook is:

- Statistical methods in Bioinformatics: an Introduction (Second Edition), Warren J. Ewens e Gregory R. Grant, Springer.

A suggested book for learning the R programming language (available online) is:

- The R book (Second Edition), Michael J. Crawley, Wiley (https://www.cs.upc.edu/~robert/teaching/estadistica/TheRBook.pdf).