UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Science and Statistical Models For Unstructured Data

2122-3-E4102B076

Learning objectives

To introduce students to modern tools for non-supervised classification and dimensionality reduction, with a particular focus on:

1. The conceptual unity underlying the problem and its connection of the statistical tools
2. The logic and mathematical structure underlying the algorithms
3. The differences among the algorithms and their consequences in data analysis
4. The intrinsic limits of statistical algorithms

In summary, the goal of the course is to provide students with state-of-the-art knowledge and competencies on non-supervised classification tools, together with a deep comprehension of the structure of the statistical techniques and a criticism capability towards their use, in terms of problem conceptualization, algorithm selection, implementation and validation, analysis of the results and their interpretation.

This way, the course provides a sound basis for practical applications within the demographic, socio-economic and biostatistical field and, in general, in all of those areas where complex data systems are to be addressed.

Contents

The course introduces the problem of non-supervised classification and dimensionality reduction, shows its application to real context, provides basic mathematical results (mainly from linear algebra), illustrates the main statistical algorithms for linear and non-linear reduction, as well as for numerical, non-numerical and partially ordered data. The illustration of the statistical tools is complemented with and supported by examples of analysis and software coding.

Detailed program

1. The problem of non-supervised classification and dimensionality reduction: examples of data analysis from socio-economics and humanities

2. Elements of linear algebra: vector spaces, scalar products, orthogonal projections, matrix norms

3. Linear techniques

    - Singular Vaue Decomposition (SVD) and its link with Principal Component Analysis
    - Non-negative Matrix Factorization and its comparison to SVD.
    - The Johnson-Lindestrauss Lemma: boundend distortion dimensionality reduction: Random Projections
    - Multidimensional Scaling.

4. Non-linear techniques

    - Data in differentiable manifolds: Isomap
    - Self-organizing map (SOM)
    - Entropy, Kullback-Liebler divergence and dimensionality reduction: SNE e t-SNE

5. Categorical and Partially ordered data

    - Correspondence Analysis
    - Ranking extraction from multidimensional datasets

6. The limits of statistical algorithms: the no-free lunch theorem.

Prerequisites

There are no formal prerequisites, but basic competencies on linear algebra, descriptive statistics and data analysis are necessary

Teaching methods

Frontal lessons with exercises and simulations, using the R language, held by the teacher. Social networks will be used to ease discussions among the students and with the teacher.

Assessment methods

Oral exam, so as to check for the comprehension of the logic structure behind the addressed methodologies and the knowledge of their analytical form.

Such a choice is motivated by:

1. The content of the course, which is of a methodological nature.
2. The relevance for the students to become capable to argue and to organize their thought, being able to perform analogical connections among the different parts of the program.

There are no special exams for students not attending the course, nor partial exams

Textbooks and Reading Materials

Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Wang J. - Springer 2012.

Methods of Multivariate Analysis, Rencher A. C., Wiley 2002

Introduction to Lattices and Order (second edition), Davey B.A:, Priestley H. A., CUP 2002 (chapter 1).

Papers and notes provided by the teacher online.

Semester

I semester II cycle

Teaching language

Italian