UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

High Dimensional Data Analysis

2627-2-FDS02Q022

Learning objectives

This is an advanced statistics course focused on the analysis of high-dimensional data, where the number of variables may be large relative to (or even exceed) the number of observations. The course aims to introduce modern statistical methods for dimension reduction, regularized regression, and variable selection in high-dimensional settings. Emphasis is placed on both the theoretical foundations and computational aspects of the methods. In line with the educational objectives of the Degree Programme, this course is part of the Statistical-Mathematical learning area. It contributes to developing the advanced analytical and methodological skills essential to defining the quantitative profile of the graduate, ensuring the coherence of individual study plans. Throughout the course, students will be encouraged to develop autonomous learning skills and the ability to critically assess new techniques, in line with the objectives of the Master's Degree Programme in Data Science.

Learning Objectives according to Dublin Descriptors (DdD)

1. Knowledge and understanding
Understand the challenges and theoretical foundations underlying statistical analysis in high-dimensional settings.
Learn core techniques such as penalized regression (e.g., Lasso, Ridge), subset selection, dimension reduction, and model selection criteria.
Familiarize with concepts such as sparsity, bias-variance trade-off, overfitting, and the curse of dimensionality.

2. Applied knowledge and understanding
Apply high-dimensional modeling techniques to real-world problems in fields such as genomics and image analysis.
Use R statistical software to implement and evaluate high-dimensional models.
Interpret model outputs and validate models using simulation studies and cross-validation techniques.

3. Autonomy of judgment
Develop the ability to choose appropriate statistical methods in complex, high-dimensional scenarios.
Critically assess the limitations and assumptions of models, and recognize potential sources of bias or instability.
Evaluate competing models using empirical evidence and theoretical criteria.

4. Communication skills

- Communicate complex statistical concepts and results clearly and rigorously, both in written form and during in-class discussions.
- Present and discuss analytical outcomes, assumptions, and limitations of high-dimensional techniques in a precise and structured manner.
- Develop the ability to justify methodological choices and interpret statistical findings in light of theoretical and empirical considerations.

5. Ability to learn
Develop the capacity to keep up with advances in the fast-evolving field of statistical learning and high-dimensional data analysis.
Engage with recent literature and critically evaluate new methodologies or applications.

Contents

This course covers methods for regression which can be applied to high-dimensional data.

Detailed program

1. Linear regression, bias/variance trade-off
2. Regularization, ridge and lasso regression
3. Model selection, cross-validation
4. Nonparametric Regression. k-*nearest neighbors* (k-NN). Kernel smoothing. Regression splines, Smoothing splines, Local regression
5. High-dimensional inference
6. Selective Inference

Prerequisites

The course requires prior knowledge of probability, statistical inference, linear algebra, and programming. Particular emphasis is placed on familiarity with the linear regression model—both in its descriptive and inferential aspects—and with asymptotic theory based on normal approximations.

Teaching methods

The course is delivered entirely in a computer-equipped classroom, combining theoretical instruction with computational practice using the R software.
All lectures will be held in English and will include:

- frontal lectures, aimed at introducing the main theoretical and methodological concepts;
- tutorial sessions, focused on practical implementation of open-source tools and their application to real-world problems in high-dimensional data analysis.

The course consists of 42 hours of in-person instruction, organized in 2- or 3-hour sessions.

Assessment methods

Assessment will be based on a final individual written exam consisting of two open-ended questions, each subdivided into sections covering both theoretical and applied aspects of the course content. Evaluation will consider the correctness, completeness, clarity, and appropriateness of language used in the responses.

No midterm exams are planned.

The final grade is expressed on a scale up to 31/30, with honors (cum laude) awarded in the case of outstanding performance. While no rigid scoring rubric is applied to each question, grades are assigned based on the overall quality of the answers and the student's command of the material, ensuring both transparency and consistency in grading.

Textbooks and Reading Materials

- Lecture notes provided by the instructor
- Azzalini, Scarpa (2012) Data analysis and data mining, an introduction . New York: Oxford University Press
- Gareth, Witten, Hastie, Tibshirani (2014) An Introduction to Statistical Learning, with Applications in R . Springer
- Hastie, Tibshirani, Friedman (2009) The Elements of Statistical Learning. Data Mining, Inference and Prediction . Springer
- Hastie, Tibshirani and Wainwright (2015) Statistical Learning with Sparsity: The Lasso and Generalizations . CRC Press

Semester

First semester

Teaching language

English

Sustainable Development Goals

QUALITY EDUCATION