UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Streaming Data Management and Time Series Analysis

2627-2-FDS02Q023

Aims

The course aims to provide students with advanced statistical and machine learning methodologies to handle and model time series data, with a focus on streaming contexts.

- **Knowledge and understanding** (DdD 1): Students will acquire a deep theoretical understanding of linear time series models (ARIMA, UCM, VAR), state-space formulations, the Kalman filter, and non-parametric/machine learning approaches for forecasting.
- **Applying knowledge and understanding** (DdD 2): Students will learn how to clean, manage, and model streaming and time series data using R (or Python), implementing precise feature engineering and model selection workflows to solve empirical prediction problems.
- **Making judgements** (DdD 3): Students will develop the critical capacity to evaluate model performance, execute rigorous model selection, and judge which methodology (linear vs. non-parametric) is most suited to the specific nature and limits of the data under analysis.
- **Communication skills** (DdD 4): Through the preparation and oral defense of an empirical project report, students will refine their ability to communicate complex statistical results, justify methodological choices, and present data insights clearly.
- **Learning skills** (DdD 5): The course provides students with the foundational toolkit and programming autonomy necessary to explore, study, and apply advanced or newly developed time series methods independently in their future professional or academic career.

Contents

Streaming data management, linear-filter-based models (ARIMA, VAR), unobserved component models (state-space form/Kalman filter), nonparametric methods (nonparametric regression, tree-based methods, neural networks, support vector machines, nearest neighbors, etc.).

Detailed program

First part

- Theory of statistical prediction (best predictor, best linear predictors).
- Stationary and integrated processes
- ARIMA models
- VAR models and cointegration (basic concepts)
- Unobserved Component Models (UCM)
- State-space form
- Kalman filter and maximum likelihood estimation of the model in state-space form
- State and disturbance smoothing
- Many applications to actual data using R (or Python)

Second part

- Main time series mining tasks
- Similarity and Clustering
- Classification, regression, and forecasting
- Non-parametric approaches based on statistical methods
- Non-parametric approaches based on machine Learning
- Artificial Neural Networks

Prerequisites

Attending students should know statistical inference, matrix algebra, and R, or Python (we will adopt R in class, but you can use Python if you prefer).

Teaching form

Theoretical lectures and hands-on sessions in the computer lab. Each lesson will be a mix of lecture-based and interactive teaching. Every time a new theoretical concept or tool is introduced, a practical application carried out together with the students will demonstrate its implementation. In fact, theory serves practice, and practice helps to understand theory.

Textbook and teaching resource

Rob J Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice* (2nd ed): https://otexts.com/fpp2/

Pelagatti M. (2015) *Time Series Modelling with Unobserved Component Models*. Chapman and Hall/CRC (il libro è scaricabile gratuitamente sotto indirizzo IP di Bicocca).

Abhijit Ghatak (2019) Deep Learning with R. Springer

Further material will be available in the elearning platform.

Semester

First semester

Assessment method

The examination is organized in two parts. First, by the date of the examination, each student must produce and send to the lecturer a paper in which they have to analyze and predict one or more time series (in agreement with the lecturers) using linear (ARIMA, UCM) and non-linear methods (RNN, SVM, etc.). The student will illustrate the paper during the oral examination in ca. 15 minutes, and the lecturers will ask questions about its content. On the same day of the oral exam, there will also be a one-hour written assessment, which consists in answering five theoretical questions on ARIMA and UCM models.

To pass the exam, both parts must have a positive valuation, and the final grade will be computed as the arithmetic mean of the grades of the two parts.

The evaluation of the theoretical part is based on the exactness and completeness of the answers (each answer is equally weighted). The assessment of the prediction exercise is based on the quality of the modeling. We will pay particular attention to feature engineering and model selection procedures.

No in-course assessments will take place.

Office hours

Pelagatti: by appointment (matteo.pelagatti@unimib.it).

Sustainable Development Goals