UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Streaming Data Management and Time Series Analysis

2627-2-FDS02Q023

Obiettivi

Il corso si propone di fornire agli studenti metodologie statistiche e di machine learning avanzate per la gestione, l'analisi e la previsione di serie storiche, con particolare attenzione ai contesti di dati in streaming.

- **Conoscenza e capacità di comprensione** (DdD 1): Gli studenti acquisiranno una profonda comprensione teorica dei modelli lineari per serie storiche (ARIMA, UCM, VAR), delle formulazioni in spazio degli stati, del filtro di Kalman e degli approcci non parametrici e di machine learning applicati alla previsione.
- **Capacità di applicare conoscenza e comprensione** (DdD 2): Gli studenti impareranno a gestire e modellare dati in streaming e serie storiche attraverso l'utilizzo di R o Python, implementando flussi di lavoro accurati di feature engineering e procedure di selezione dei modelli per risolvere problemi empirici di previsione.
- **Autonomia di giudizio** (DdD 3): Gli studenti svilupperanno la capacità critica di valutare le prestazioni dei modelli predittivi, eseguire una rigorosa selezione delle variabili e dei modelli, e giudicare quale metodologia sia più adatta in base alla natura specifica e ai limiti dei dati sotto analisi.
- **Abilità comunicative** (DdD 4): Attraverso la redazione e l'illustrazione orale di una relazione scritta (paper) incentrata su un esercizio di previsione empirica, gli studenti affineranno la capacità di comunicare risultati statistici complessi, giustificare le scelte metodologiche adottate e presentare chiaramente gli insight derivanti dai dati.
- **Capacità di apprendimento** (DdD 5): Il corso fornisce gli strumenti concettuali e l'autonomia di programmazione necessari affinché gli studenti possano esplorare, studiare e applicare in modo indipendente metodi avanzati o di recente sviluppo per l'analisi delle serie storiche nel corso della loro futura carriera professionale o accademica.

Contenuti sintetici

Gestione dei dati in tempo reale, modelli basati su filtri lineari (ARIMA), modelli basati su componenti non

osservabili (state-space/Kalman filter), modelli non parametrici (reti neurali, support vector machine, vicini più vicini, ecc.).

Programma esteso

Prima parte

- Teoria della previsione statistica (miglior previsore e miglior previsore lineare).
- Processi stazionari e integrati
- Modelli ARIMA
- Modelli VAR e cointegrazione (accenni)
- Modelli a componenti non nosservabili (UCM)
- Forma state-space
- Kalman filter e stime di massima verosimiglianza di modelli in forma state-space
- Smoothing delle variabili di stato e dei disturbi (estrazione comonenti e identificazione di anomalie).
- Applicazioni a dati reali usando R (o Python)

Seconda parte

- Principali task di time-series mining
- Classificazione, regressione e previsione
- Approcci non-parametrici statistici
- Approcci non-parametrici basati su Machine Learning
- Reti Neurali

Prerequisiti

Per seguire l'insegnamento con successo è necessario conoscere l'inferenza statistica, l'algebra matriciale ed R o Python (useremo R nelle lezioni, ma potete utilizzare anche Python se preferite).

Modalità didattica

Lezioni teoriche e pratiche in laboratorio informatico. Ogni lezione sarà un mix di didattica erogativa e interattiva. Ogni volta che un nuovo concetto o strumento teorico verrà spiegato, un'applicazione pratica fatta insieme agli studenti illustrerà la relativa implementazione. Infatti, la teoria serve alla pratica e la pratica aiuta a capire la teoria.

Materiale didattico

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

Altro materiale sarà reso disponibile sulle pagine elearning.

Periodo di erogazione dell'insegnamento

Primo semestre

Modalità di verifica del profitto e valutazione

L'esame è organizzato in due parti. Entro la data dell'esame ciascuno studente dovrà produrre e inviare al docente un elaborato dove una o più serie storiche concordate con il docente dovranno essere analizzate e previste per mezzo di modelli ARIMA, UCM e machine learning. Lo studente illustrerà l'elaborato durante l'orale in una quindicina di minuti e il docente potrà fare domande sul contenuto. Nel medesimo giorno dell'orale vi è anche uno scritto dalla durata di un'ora che prevede la risposta a cinque tra domande teoriche aperte ed esercizi su modelli ARIMA e UCM.

Per superare l'esame entrambe le parti dovranno essere sufficienti e il voto finale sarà calcolato come media aritmetica semplice delle votazioni delle due parti.

La valutazione della parte teorica sarà basata sull'esattezza e la completezza delle risposte alle domande proposte (ogni risposta ha lo stesso peso nel voto). La valutazione dell'elaborato sarà basata sulla qualità dei modelli costruiti e, in particolare, sulle features costruite e sulla selezione dei modelli finali.

Non sono previste prove in itinere.

Orario di ricevimento

Pelagatti: su appuntamento (matteo.pelagatti@unimib.it).

Sustainable Development Goals