UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining

2021-2-F8204B018-F8204B034M

Obiettivi formativi

Il corso si pone come obiettivo l'acquisizione delle principali tecniche per l'esplorazione dei dati (*data mining*) e di apprendimento supervisionato (*supervised learning*) e la loro implementazione nell'ambiente di programmazione R. Durante il corso verrà data particolare enfasi al processo di modellazione dei dati per la previsione (*predictive modelling*).

Alla fine del corso lo studente sarà in grado di affrontare l'analisi di dati complessi a fini previsivi attraverso il processo di esplorazione, manipolazione e modellazione dei dati.

Contenuti sintetici

Il corso integra considerazioni di carattere teorico con aspetti pratico-applicativi di analisi dei dati e di programmazione in R.

- Aspetti teorici: il compromesso tra distorsione e varianza, stime vincolate/penalizzate, splines e modelli additivi generalizzati, quantificazione dell'incertezza delle previsioni
- Aspetti applicativi: esplorazione, manipolazione e modellizzazione dei dati in R per la previsione

Programma esteso

- Errore di previsione: il compromesso tra distorsione e varianza
- Il modello e il processo di modellizzazione
- Stime vincolate e penalizzate: i metodi
- Splines e modelli additivi generalizzati
- Quantificare l'incertezza delle previsioni

- Aspetti computazionali

Prerequisiti

Si consiglia la conoscenza degli argomenti trattati nei corsi *Probabilità e Statistica Computazionale M* e *Statistica Avanzata M*.

Metodi didattici

Le lezioni si svolgono sia in aula che in laboratorio, integrando aspetti di carattere teorico con quelli pratico-applicativi di analisi dei dati e di programmazione in R.

Nel periodo di emergenza Covid-19 le lezioni si svolgeranno in modalità da remoto asincrono, eventualmente con eventi in videoconferenza sincrona e/o in presenza fisica.

Modalità di verifica dell'apprendimento

La modalità di verifica si basa su una prova finale con orale facoltativo. La prova finale è composta da due parti:

1. Prova scritta: domande di teoria ed esercizi
2. Homework

Il voto della prova finale è dato dalla media pesata delle parti 1. e 2. Qualora lo studente (oppure i docenti) richiedano la prova orale, il voto finale è una media dei voti della prova finale e della prova orale.

La prova scritta (parte 1.)

Testi di riferimento

- Archivio del corso: https://github.com/aldosolari/DM
- Arnold (2019) *A Computational Approach to Statistical Learning,* Chapman & Hall
- Azzalini, Scarpa (2004). *Analisi dei dati e data mining*. Springer-Verlag Italia

- Gareth, Witten, Hastie, Tibshirani (2013). *Introduction to Statistical Learning with applications in R*. Springer
- Hastie, Tibshirani, Friedman (2009).
- Kuhn, Johnson (2013). *Applied Predictive Modelling*. Springer
- Kuhn, Johnson (2019). *Feature Engineering and Selection*. Chapman and Hall/CRC
- Wickham, Grolemund (2015) *R for Data Science*. O'Reilly Cookbooks

Periodo di erogazione dell’insegnamento

Primo semestre, primo ciclo.

Lingua di insegnamento

Le lezioni si svolgono in italiano, tuttavia la maggior parte dei libri di testo è in lingua inglese.