UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining

2627-3-E4101B026

Obiettivi formativi

Il corso intende fornire una visione completa del Data Mining, dal pre-processamento del dato fino alla selezione del miglior modello statistico per l'analisi e la comprensione del problema. Durante il corso verranno affrontate le principali tecniche per il trattamento dei dati e presentati metodi statistici di tipo supervisionato. Inoltre, verranno introdotti concetti relativi al Text Mining.
Alla fine del corso, lo studente sarà in grado di confrontare e selezionare il miglior metodo di Data Mining per il problema oggetto di analisi. Saprà trattare le principali problematiche relative al dato e, autonomamente, affrontare un problema reale nel miglior modo.

Il corso contribuisce al raggiungimento degli obiettivi formativi nell'area di apprendimento del Corso di Laurea Triennale: "Statistica".
.

Contenuti sintetici

Trattamento dei missing values.
Metodi supervisionati di classificazione/regressione.
Trade-off bias varianza.
Text mining.
Market basket analysis.

Programma esteso

1. Introduzione al Data mining.

2. Pre-processing: trattamento dei missing values. Metodi di imputazione singola e multipla.
3. Introduzione alla classificazione con esempi e concetti introduttivi. Metodi di classificazione: discriminante lineare, discriminante quadratico, k-nn e alberi decisionali.
4. Trade off bias varianza. Definizione di overfitting e relative tecniche di mitigazione.
5. Text mining con esempi e concetti di base: pre-processing (ad esempio eliminazione stop words) e rappresentazioni grafiche per il Text Mining.
6. Market Basket Analysis e algoritmo aPriori.

Prerequisiti

Analisi Statistica Multivariata e programmazione in R.

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.

Le 42 ore di didattica saranno così suddivise:

- 30 ore di lezione svolte in modalità erogativa;
- 12 ore di attività di laboratorio.

Modalità di verifica dell'apprendimento

Scritto
(20 su 32) Prova scritta mirata a verificare gli argomenti presentati in aula.

Progetto
(12 su 32) Progetto applicativo da svolgere autonomamente o in gruppo (max. 3 persone) su un dataset assegnato dal docente o scelto dagli studenti. Il progetto è realizzato in R e deve dimostrare la capacità di affrontare un problema reale in ogni suo aspetto utilizzando quanto visto a lezione. Il progetto si compone sia del codice R sia di un report di presentazione realizzato attraverso Rmarkdown.

Testi di riferimento

Fonte principale:
Gareth J., Witten D., Hastie T., Tibshirani R., *An Introduction to statistical learning with application in R*, springer (2013).

Fonti utili per approfondire R:
W. N. Venables, D. M. Smith and the R Core Team, An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics.
https://cran.r-project.org/doc/manuals/R-intro.pdf

C. Agostinelli, Introduzione a R. https://cran.r-project.org/doc/contrib/manuale.0.3.pdf

Altro materiale utile:
http://www.feat.engineering

Altro materiale verrà indicato a lezione.

Periodo di erogazione dell'insegnamento

II Semestre - III periodo

Lingua di insegnamento

Italiano

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ