UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Mining

2122-3-E4102B085-E4102B086M

Obiettivi formativi

Data mining

Il corso intende fornire un'introduzione alle principali tecniche statistiche di Data Mining attraverso le più moderne tecniche e strategie per l'analisi di grandi moli di dati, illustrando le problematiche connesse.

Alla fine del corso lo studente ha la possibiltà di proporre i principali algoritmi , discernendo pregi e difetti, essendo in grado di sperimentare ed applicare le conoscenze acquisite su dati reali.

Contenuti sintetici

Il corso affronta lo studio di tecniche modellistiche algoritmiche e le principali problematiche e tecniche statistiche di Data Mining

Programma esteso

Data mining

Il Data mining, robustezza, overfitting e problematiche di validazione dei risultati, Regole associative, Modelli statistici per la classificazione supervisionata (modello lineare, analisi discriminante parametrica, modello logistico binario e multinomiale), Algoritmi per la classificazione supervisionata (Naive Bayes, Nearest Neighbour, neural network, regressioni lasso, Alberi decisionali e Classificativi, PLS, Bagging, Boosting and Random forest)

Prerequisiti

Superamento esame di Analisi statistica Multivariata

Metodi didattici

Nel periodo di emergenza Covid-19 le lezioni si svolgeranno in modalità mista: parziale presenza e lezioni sincrone (streeming) via piattaforme web.

Modalità di verifica dell'apprendimento

PROVA SCRITTA

PROJECT WORK (Sviluppo di un progetto originale a partire da una semplice idea o dall'analisi di un caso esistente)

Lavoro applicativo da svolgere autonomamente o in gruppo di max 3 persone su dataset scelti dallo studente (R o SAS) su cui applicare i principali argomenti svolti a lezione .

Di seguito le analisi da svolgere per i due moduli in ogni project work (Sas base o R):

Data mining (Sas Enterprise Miner o R)

1 PROJECT WORK, analisi con con target binario (classificazione)

(ANALISI DA SVOLGERE: analisi descrittive, proposta diversi modelli, validation strategies, preprocessing, tuning modelli, confronto modelli, score di nuovi dati)

In totale per superare l'esame da 15 cfu è necessario completare due project work (1 di statistica computazionale + 1 di Data mining) su due dataset differenti

Portali per la scelta dei dataset:

https://archive.ics.uci.edu/ml/datasets

www.kaggle.com

PROVA ORALE

I principali output del PROJECT WORK (svolto nelle settimane precedenti la data dell'orale) vanno stampati e portati all'orale, se in presenza.

Altrimenti il COLLOQUIO avviene via WEB DI DISCUSSIONE SUL project work (Nel periodo di emergenza Covid-19 gli esami orali saranno solo telematici. Verranno svolti utilizzando la piattaforma WebEx e nella pagina e-learning dell'insegnamento verrà riportato un link pubblico per l'accesso all'esame di possibili spettatori virtuali).

L'esame orale, per ciascun modulo, consta di domande sulla TEORIA affrontata a lezione e sul commento degli output del lavoro applicativo per verificare la comprensione dei principali strumenti adottati e il conseguente "modus operandi" dell'analisi statistica svolta.

Lo studente deve dimostrare di aver appreso il funzionamento dei principali algoritmi, essendo in grado di comprenderne pregi e difetti e di applicare tali strumenti su dati reali.

Non sono previste prove in itinere

Testi di riferimento

Data mining

Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R

http://www-bcf.usc.edu/~gareth/ISL/

Chapter 2-3-4-5- 8

Lucidi sul moodle

Periodo di erogazione dell’insegnamento

I semestre, ciclo II

Lingua di insegnamento

ITA