UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining e Machine Learning

2223-3-E4102B087

Obiettivi formativi

Data Mining e Machine Learning

Sezione Data Mining
L'obiettivo principale è introdurre metodologie avanzate anche di tipo non analitico/algoritmico ad alta complessità computazionale per disegnare ed eseguire analisi di dati .

Sezione Machine learning
Tale sezione intende fornire un'introduzione alle principali tecniche statistiche di Machine learning 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 avrà la possibiltà di conoscere i principali algoritmi di DM e ML, discernendo pregi e difetti, essendo in grado di sperimentare ed applicare le conoscenze acquisite su dati reali con R studio.

Contenuti sintetici

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

Programma esteso

Data mining section

- (1) SAS language and R (overview)

- (2) Interpretazione di Modelli lineari complessi (Anova, Ancova, GLM) con interazioni, trasformate,
- (3) Robust methods (Bootstrap, Jacknife, Robust Regression, IRLS, WLS, nonparametric regression, loess smoothing and splines)
- (4) Passi per costruzione di un modello Robusto
- (5) missing data mechanism, missing imputation, (y, X)-transformation, misure di Influenza, diagnostiche, heteroschedaticità, model selection.

Machine Learning section

Problematiche connesse a grandi mli di dati, robustezza, overfitting e problematiche di validazione dei risultati, Regole associative, Modelli statistici per la classificazione supervisionata (modello lineare, analisi discriminante parametrica, modello logistico politomico e ordinale), Algoritmi per la classificazione supervisionata (Naive Bayes, Nearest Neighbour, Neural Network, Alberi decisionali e Classificativi, PLS, Bagging, Boosting e Random forest)

Prerequisiti

Superamento esame di Analisi statistica Multivariata

Metodi didattici

Lezioni in presenza

Modalità di verifica dell'apprendimento

PROVA ORALE SU UN ELABORATO SVOLTO DA PORTARE ALL'ORALE (PROJECT WORK)
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 nel project work (composto da due parti, sezione Data Mining e sezione Machine Learning):

Project work Data Mining
Analisi target quantitativo: costruzione di un modello robusto (analisi descrittive, trasformazioni, diagnostiche, model selection, heteroskedasticità, inferenza robusta) e una breve analisi con target binario (stampare output di una regressione logistica).

Project work Machine learning
Analisi con con target binario (classificazione)
(Analisi descrittive, preprocessing, proposta diversi modelli, validation strategies, tuning modelli, confronto modelli, studio della soglia, score di nuovi dati)

Il dataset delle due parti può essere lo stesso (nel PW di Machine learning potete binarizzare il target quantitativo del PW di Data mining o scegliere un'altra variabile) SOLO SE DI ADEGUATA COMPLESSITA'

Portali per la scelta dei dataset:

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

www.kaggle.com

SVOLGIMENTO PROVA ORALE

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

L'esame orale, per ciascuna sezione (DM, ML) 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.

E' prevista una prova-esame in itinere a novembre alla fine del modulo di DM.

Testi di riferimento

Data Mining

Principles of Econometrics (chapters 2, 4 ,6 ,8 9, 12, 13) Carter Hill, William E. Griffiths, Guay C. Lim. http://www-bcf.usc.edu/~gareth/ISL/

An Introduction to Statistical Learning with Applications in R (Chapter 3 (no section 3.5), Chapter 4, 6, until 6.1, chapter 7) Carter Hill, William E. Griffiths, Guay C. Lim.

Lucidi del docente

Consigliati
Principles of Econometrics associate R book https://bookdown.org/ccolonescu/RPoE4/

A Handbook of Statistical Analyses Using R (2nd Edition) Chapters 5,6,7,8,10

Machine Learning
Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R (Chapter 2-3-4-5- 8)

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

Lucidi sul moodle

Periodo di erogazione dell'insegnamento

I semestre

Lingua di insegnamento

ITA

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ