UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining e Machine Learning

2627-3-E4102B087

Obiettivi formativi

Data Mining e Machine Learning (Esame Unico 15 cfu, 120 ore: 8 ore per CFU)

Materiale: https://elearning.unimib.it/course/view.php?id=67389

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.

In particolare si perseguono i seguenti obiettivi specifici:

1. Conoscenza e capacità di comprensione (teoria del data mining e del machine learning);
2. Conoscenza e capacità di comprensione applicate (attraverso le lezioni in laboratorio e il project work in gruppo);
3. Autonomia di giudizio (attraverso il project work da presentare all'esame, quiz in classe e data analysis da svolgere autonomamente);
4. Abilita comunicative (attraverso l'esposizione del project work svolto per l'esame durante l'orale);
5. Capacita di apprendere (attraverso materiali di approfondimento forniti sul moodle che lo studente può affrontare in autonomia sulla base delle conoscenze acquisite durante la parte comune).

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
Costruzione di un modello (stimatore) robusto a partire dalla stima di un modello base.

- (1) R and dplyr (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.
- (6) multilevel models for longitudinal data and hierachcal data (cenni)

Machine Learning section

Problematiche connesse a grandi moli 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

SOLO lezioni in presenza

VECCHIE VIDEOLEZIONI https://drive.google.com/drive/folders/10zCiGvEVf0xo18GwLU9W1NA6rj-g93aX?usp=sharing
SI ENTRA IN GOOGLE DRIVE E SI ACCEDE AI FILES CON LE SOLE CREDENZIALI UNIMIB.

Modalità di verifica dell'apprendimento

PROVA ORALE SU UN ELABORATO SVOLTO da portare all'orale (PROJECT WORK) e SUGLI ARGOMENTI SVOLTI A LEZIONE

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 nei due project work (composto da due parti, sezione Data Mining e sezione Machine Learning):

Project work Data Mining
Analisi target quantitativo:

1. 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).
2. Oppure usare un mdellopiù complesso per dati logitudinali o gerarchici e usare mutitlvel models o sample selection models in casi di campione troncato

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.
LO SVOLGIMENTO DEL PROJECT WORK, ANCHE SE RITENUTO DI OTTIMA FATTURA, NON COMPORTA IL SUPERAMENTO DELL'ESAME, QUALORA EMERGESSERO CARENZE SUGLI ARGOMENTI TEORICI

**PROVA IN ITINERE DI DM
**E' prevista una prova-esame in itinere a novembre alla fine del modulo di DM (nella settimana dedicata agli appelli del I periodo).
(Project work da svolgere in gruppo e prova orale individuale)
Lo studente deve dimostrare di aver appreso COME RENDERE UN MODELLO ROBUSTO empriricamente (PW) e di conoscere la teoria statistica sottostante.

Le prove orali sono individuali, sebbene per comodità il docente tenderà ad interrogare congiuntamente tutti i componenti del gruppo che han svolto il project work (e che si presenteranno all'orale).

Testi di riferimento

Data Mining

Carter Hill, William E. Griffiths, Guay C. Lim.
Principles of Econometrics (chapters 2, 4 ,6 ,8 9, 12, 13)

An Introduction to Statistical Learning with Applications in R (Chapter 3 (no section 3.5), Chapter/section 4.1, 4.2, 4.3 , 6.1, 6.2, chapter 7)
https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html

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
https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.ehu.eus/ccwintco/uploads/9/93/A_Handbook_of_Statistical_Analyses_Using_R_Second_Edition.pdf&ved=2ahUKEwjLzbfZq-WGAxWvgv0HHRx7AzAQFnoECBEQAQ&usg=AOvVaw0P4Jf6CnMmRwFth4y5zQsh

Machine Learning
Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R (Chapter 2-4-5-6-8-10.1, 10.2, 12(parte PCA))
https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf.download.html
Lucidi sul moodle

Periodo di erogazione dell'insegnamento

I semestre

Lingua di insegnamento

ITA

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ