UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining e Statistica Computazionale

1920-3-E4102B085

Obiettivi formativi

Data mining e Statistica computazionale (insegnamento in due moduli)

Statistica computazionale

L'obiettivo principale del corso è introdurre strumenti software avanzati e di alta complessità computazionale per disegnare ed eseguire analisi di dati e modellazione statistica complessa.

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

Statistica computazionale

- **(1) SAS language and R (overiview)**
- **(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) Logistic Regression**

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 politomico e ordinale), Algoritmi per la classificazione supervisionata (Naive Bayes, Nearest Neighbour, neural network, Alberi decisionali e Classificativi, PLS, Bagging, Boosting and Random forest)

Prerequisiti

Superamento esame di Analisi statistica Multivariata

Metodi didattici

Lezione frontale e sessioni di laboratorio

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):

Statistica computazionale

1 PROJECT WORK completo con con target quantitativo

(analisi descrittive, trasformazioni, diagnostiche, model selection, heteroskedasticità,..... fino alla costruzione di un modello robusto)

comprendente infine un' analisi di regressione logistica con target binario (discretizzare il target precedente)

utilizzando le covariate di interesse controllando

collinearity e separation (opzionale model selection)

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 due project work (stat computazionale+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.

COLLOQUIO DI DISCUSSIONE SULLO SCRITTO

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

Statistica computazionale

Gareth, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning with Applications in R
http://www-bcf.usc.edu/~gareth/ISL/

Chapter 3 (no section 3.5), Chapter 4, 6,7

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

Lucidi sul moodle

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, cicli I e II

Lingua di insegnamento

ITA