UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Statistical Modeling

2223-1-FDS01Q004

Obiettivi formativi

Il corso introduce alle procedure analitiche ed inferenziali condotte tramite modelli statistici avanzati: la regressione lineare multipla e alcune sue estensioni, i modelli lineari generalizzati ampiamente utilizzati per l'analisi dei dati in molti ambiti. Il corso introduce lo studente anche all'analisi di classificazione effettuata attraverso i modelli statistici. Oltre alla modellizzazione il corso introduce anche alcuni metodi computazionali.

Conoscenza e comprensione
Lo studente viene introdotto ai modelli statistici avanzati per diverse tipologie di variabili risposta e alle relative ipotesi alla base della teoria considerando il metodo di stima della massima verosimiglianza dei parametri dei modelli. L'analisi dei dati viene condotta utilizzando il software R e l'ambiente RMarkdown che permette di creare documenti riproducibili contenenti il codice, i risultati ed i commenti. Gli esempi applicativi riguardano dati reali e simulati provenienti da diversi ambiti quali l'economia, la finanza, e le scienze sociali. Lo studente è incoraggiato ad una valutazione critica dei modelli utilizzati.

Capacità di applicare conoscenza e comprensione
Il corso fornisce competenze nell'utilizzo della semantica del software open-source R per l'analisi descrittiva dei dati multivariati e per la stima dei parametri di modelli univariati e multivariati. Attraverso R e RStudio lo studente impara a sviluppare in modo organico il ragionamento statistico attraverso l'analisi dei dati e la redazione di relazioni che illustrino il codice, le analisi e commentino i risultati. La teoria viene affiancata da applicazioni pratiche. Il corso permette agli studenti di acquisire solidi elementi di teoria e capacità di applicare i modelli statistici proposti a dati reali. L'insegnamento è indispensabile per il successivo percorso universitario in quanto fornisce i concetti essenziali per lo sviluppo dei metodi statistici parametrici e non parametrici sia in ambito teorico che applicativo per i contesti lavorativi di sbocco degli studenti del corso di laurea in Data Science.

Contenuti sintetici

Nella prima parte del corso, dopo una breve introduzione, viene presentato il modello di regressione lineare multipla con le sue ipotesi, i minimi quadrati ordinari e la stima della massima verosimiglianza, le proprietà statistiche degli stimatori dei minimi quadrati, la stima della varianza, le misure di adattamento, la diagnostica della regressione e la previsione. Successivamente, lo studente viene introdotto al metodo di ricampionamento bootstrap e ai suoi scopi inferenziali. Vengono inoltre trattati i modelli lineari generalizzati. L'algoritmo expectation-maximization viene introdotto come strumento per la stima di massima verosimiglianza dei parametri dei modelli di classificazione. Il corso fornisce competenze nell'uso della semantica del software R anche utilizzando le librerie RMarkdown attraverso la libreria knitr per integrare il codice, i risultati delle analisi ed i commenti.

Programma esteso

Nell'introduzione al corso vengono richiamati alcuni concetti dell'inferenza statistica e dell'inferenza intesa in senso causale.

- La prima parte del corso riguarda il modello di regressione lineare multipla, i metodi di stima a minimi quadrati e della massima verosimiglianza. Le proprietà degli stimatori dei minimi quadrati vengono discusse in base alle ipotesi del modello.
- Viene introdotta la distribuzione di Gauss bivariata e multivariata e vengono simulate delle realizzazioni casuali da queste distribuzioni.
- Si considerano vari strumenti diagnostici per la valutazione del modello in base ai residui di regressione e vengono introdotti alcuni criteri per la selezione delle variabili, come il criterio di informazione di Akaike.
- Lo studente viene introdotto al metodo di ricampionamento bootstrap e ai suoi scopi inferenziali. Vengono introdotti i modelli lineari generalizzati per l'analisi dei dati categoriali sia binari che con più di due categorie.
- Lo studente impara a valutare il modello statitico anche il base alla capacità predittiva con relativi intervalli di previsione.
- Tra i metodi di ottimizzazione, l'algoritmo expectation-maximization viene spiegato come strumento computazionale per massimizzare la funzione di verosimiglianza.
- I modelli miscuglio sono introdotti come metodi di clustering e si considerando in particolare i miscugli di distribuzioni Gaussiane.

Le spiegazioni teoriche sono affiancate dalle applicazioni empiriche, basate su dati simulati e reali riferiti a diversi ambiti applicativi: l'economia, la finanza, la biologia, l'ecologia e le scienze ambientali. Le applicazioni sono svolte utilizzando numerose librerie del software statistico open-source R, RStudio e l'interfaccia RMarkdown attraverso la libreria knitr. Questo permette di introdurre lo studente ai principi della riproducibilità della ricerca.

Lo studente svolgendo gli esercizi assegnati è incoraggiato a scrivere report in cui commenta il codice ed offre al lettore una spiegazione del procedimento di analisi condotto oltre ad una descrizione critica dei risultati ottenuti.
Lo studente è invitato a svolgere gli esercizi assegnati in gruppo in modo da sviluppare l'apprendimento cooperativo.

Prerequisiti

Per una più facile comprensione dei contenuti del corso, si raccomanda di conoscere le nozioni di probabilità e di inferenza statistica ed i contenuti del corso Fondamenti di Probabilità e Statistica. Il corso presuppone una conoscenza preliminare dei seguenti argomenti: probabilità di un evento, funzione di distribuzione della probabilità, densità, funzioni di distribuzione cumulativa, legge della probabilità totale, indipendenza degli eventi, teorema di Bayes, aspettativa e varianza di una variabile casuale, standardizzazione e percentili di una variabile casuale, variabili casuali continue e discrete quali Bernoulli, binomiale, Poisson, geometrica, uniforme, esponenziale, Gaussiana, Student-t, chi-quadrato, grafici e misure numeriche per descrivere i dati. Teoria dell'inferenza statistica,

e della stima di massima verosimiglianza. Analisi multivariata dei dati e dell'algebra lineare. Lo studente deve inoltre conoscere la semantica di base del linguaggio di programmazione in ambiente R.

Metodi didattici

Le lezioni teoriche sono integrate da esercitazioni pratiche che consentono allo studente di apprendere la teoria e di applicare i modelli per l'analisi di dati reali e simulati. Le lezioni si svolgono presso il laboratorio informatico. Settimanalmente vengono assegnati degli esercizi di riepilogo per favorire l'apprendimento della teoria e delle applicazioni. Durante il corso, con l'aiuto di R nell'ambiente RStudio e dell'interfaccia RMarkdown, gli studenti imparano anche a elaborare documenti riproducibili. Sono incoraggiati ad affrontare i problemi applicativi con l'ulteriore obiettivo di sviluppare l'apprendimento cooperativo. Verranno anche impartite delle lezioni di tutoraggio per aiutare gli studenti a sviluppare gli esercizi assegnati.

Modalità di verifica dell'apprendimento

Le seguenti modalità di verifica dell'apprendimento si applicano sia agli studenti frequentanti che a quelli non frequentanti le lezioni in presenza. L'esame è scritto con domande aperte e orale facoltativo; sono previste prove intermedie. L'esame scritto ha una durata massima complessiva di un'ora e mezza e si svolge in laboratorio informatico. Durante l'esame è necessario rispondere a domande aperte di teoria e risolvere gli esercizi alla luce degli argomenti teorici trattati e delle esercitazioni pratiche assegnate settimanalmente durante il corso. Le domande di teoria verificano l'apprendimento dei concetti teorici insegnati. Le analisi empiriche sono condotte utilizzando l'ambiente R, RStudio e RMarkdown e consentono di verificare la capacità dello studente di applicare modelli statistici avanzati a dati reali o simulati e elaborare report riproducibili di descrizione dai dati e delle procedure e di illustrazione dei risultati ottenuti. L'esame si svolge a libro aperto e gli studenti possono consultare il codice R implementato durante il corso e le esercitazioni. Lo studente supera l'esame con una votazione di almeno 18/30.

Testi di riferimento

Il materiale didattico è costituito principalmente dalle dispense preparate dal docente riguardanti sia gli argomenti teorici sia le applicazioni sviluppate con il software R. Queste saranno rese disponibili presso la pagina della piattaforma e-learning dell'università dedicata al corso. Inoltre, il docente pubblica alla fine di ogni lezione: le slide, i programmi di calcolo, i dataset. Settimanalmente vengono assegnati esercizi, e alcune soluzioni. Nella stessa pagina sono pubblicati degli esempi del testo d'esame.
I riferimenti primari saranno elencati nella bibliografia delle dispense; tra gli altri, si segnalano i seguenti disponibili presso la biblioteca o in ebook:

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). Latent Markov models for longitudinal data, Chapman and Hall/CRC, Boca Raton.

Bishop, Y. M., Fienberg, S. E., Holland, P. W. (2007). Discrete multivariate analysis: theory and practice. Springer Science & Business Media, New York.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). Model-based clustering and classification for data science: With applications in R. Cambridge University Press.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. D. (2021). Regression: Models, methods and applications. Springer Berlin, Heidelberg.

Faraway, J. J. (2014). Extending the Linear models with R, 2nd Edition, Chapman & Hall, CRC Press.

Hastie, T., D. and Tibshirani, R. (2013). An introduction to statistical learning, New York, Springer.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, 2nd Edition. Chapman and Hall/CRC, London.

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Xie, Y., Dervieux, C. and Riederer E. (2020). R Markdown Cookbook. Chapman & Hall, CRC

Periodo di erogazione dell'insegnamento

2° semestre, Marzo 2023-Maggio 2023

Lingua di insegnamento

Il corso viene erogato in lingua inglese e l’esame si svolgerà in lingua inglese.

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | RIDURRE LE DISUGUAGLIANZE | PARTNERSHIP PER GLI OBIETTIVI