UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Modeling

2324-1-FDS01Q004

Obiettivi

Il corso permette allo studente di apprendere le procedure analitiche ed inferenziali riguardanti seguenti modelli statistici: la regressione lineare multipla e alcune sue estensioni, i modelli lineari generalizzati ed alcuni modelli mistura Gaussiani univariati e multivariati.

Conoscenza e comprensione

Lo studente viene introdotto alla conoscenza di modelli statistici avanzati per l'analisi di dati con diverse tipologie di variabili risposta. Si illustrano anche le relative ipotesi alla base della teoria considerando il metodo di stima della massima verosimiglianza e dei minimi quadrati per i 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 come l'economia, la finanza, e le scienze sociali. Lo studente è incoraggiato a fornire anche una valutazione critica circa i risultati ottenuti con le analisi empiriche.

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 gli studenti imparano ad impostare in modo organico il ragionamento statistico attraverso l'analisi dei dati e la redazione di relazioni che illustrino il codice, le analisi ed i risultati. La teoria viene affiancata da applicazioni pratiche. Il corso consente agli studenti di acquisire solide basi teoriche 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 sull’impianto concettuale dell’inferenza statistica, viene presentato il procedimento di ricampionamento noto come bootstrap per ottenere misure di precisione in ambito non parametrico per alcuni stimatori di interesse. In seguito 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. Vengono inoltre trattati i modelli lineari generalizzati, con particolare riferimento al modello di regressione logistica multipla e al modello logistico multinomiale. 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, utilizzando anche le librerie RMarkdown tramite 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 causale. Vengono richiamati i concetti di verosimiglianza e di inferenza Bayesiana.

La prima parte del corso riguarda l’introduzione al metodo di ricampionamento noto come bootstrap per la determinazione dell’errore standard come misura di accuratezza. Il metodo viene applicato a diversi stimatori utilizzando dati di interesse.

La seconda 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 sulla base alle ipotesi del modello.

Viene introdotta la distribuzione Gaussiana bivariata e multivariata con simulazioni di realizzazioni casuali da tali distribuzioni. La distribuzione viene espressa graficamente anche attraverso l’utilizzo delle curve di livello.

Sono considerati diversi strumenti diagnostici per la valutazione del modello in base ai residui di regressione con particolare enfasi per la determinazione degli outliers, valori anomali e punti di leva. Viene affrontato il problema della selezione delle variabili esplicative più rilevanti attraverso l’utilizzo dei criteri informativi quali il criterio di Akaike. Si impara a valutare il modello anche in relazione alla sua capacità predittiva.

Vengono introdotti i modelli lineari generalizzati per l'analisi di variabili risposta categoriali con due o più categorie. Si illustrano il modello di regressione logistica multipla ed il modello multinomiale, enfatizzando in particolare l’interpretazione dei coefficienti di regressione.

Tra i metodi di ottimizzazione, l'algoritmo expectation-maximization viene spiegato come strumento computazionale per massimizzare la funzione di verosimiglianza.
I modelli mistura sono introdotti come metodi di clustering. Vengono illustrati in modelli mistura 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. Queste applicazioni sono realizzate utilizzando diverse 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.

Settimanalmente vengono assegnati degli esercizi e gli studenti nello svolgimento sono incoraggiati a scrivere report in cui commentano il codice, ed offrono al lettore una spiegazione del procedimento di analisi svolto oltre ad una descrizione critica rispetto ai risultati ottenuti.
Gli studenti sono invitati a svolgere gli esercizi assegnati anche in gruppo, allo scopo di promuovere l'apprendimento cooperativo.

Prerequisiti

Per una più facile comprensione dei contenuti del corso, è utile avere conoscenze di base in 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 di probabilità, e di densità, densità cumulata, 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 la distribuzione Gaussiana, di Bernoulli, binomiale, Poisson, geometrica, uniforme, ed esponenziale. Occorre conoscere i principi di base dell'analisi statistica multivariata e dell'algebra lineare nonché una conoscenza elementare del linguaggio di programmazione R.

Modalità didattica

Le lezioni teoriche sono integrate da esercitazioni pratiche che consentono agli studenti di apprendere la teoria applicando modelli per l'analisi di dati reali e simulati. Le lezioni si svolgono presso il laboratorio informatico. Ogni settimana vengono assegnati esercizi di riepilogo per favorire l'apprendimento della teoria e delle applicazioni. Durante il corso, utilizzando R nell'ambiente RStudio e l'interfaccia RMarkdown, gli studenti imparano a creare report riproducibili. Sono incoraggiati a collaborare nella risoluzione dei problemi applicativi, al fine di promuovere l'apprendimento cooperativo. Durante le sessioni di tutoraggio, gli studenti confrontano le soluzioni degli esercizi assegnati settimanalmente.

Materiale didattico

Il materiale didattico principale consiste nelle dispense preparate dal docente, che coprono sia gli argomenti teorici che le applicazioni sviluppate con il software R. Queste dispense saranno rese disponibili sulla pagina della piattaforma e-learning dell'università dedicata al corso. Inoltre, il docente pubblica alla fine di ogni lezione le slides, i programmi di calcolo e i dataset utilizzati. Settimanalmente vengono assegnati esercizi, alcuni dei quali verranno accompagnati dalle relative soluzioni. Sulla stessa pagina web sono disponibili 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 (2023). 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 2024 - Maggio 2024

Modalità di verifica del profitto e valutazione

Le seguenti modalità di verifica dell'apprendimento si applicano sia agli studenti frequentanti che a quelli non frequentanti le lezioni che si svolgono in presenza. L'esame è composto da una parte scritta con domande aperte e da una parte orale facoltativa. Durante il corso sono previste anche prove intermedie. Gli studenti frequentanti avranno l'opportunità di ricevere un bonus se consegneranno alcuni degli esercizi assegnati nelle date indicate. L'esame scritto ha una durata massima di un'ora e mezza e si svolge in laboratorio informatico. Durante l'esame, gli studenti devono rispondere a domande aperte di teoria e risolvere gli esercizi basandosi sugli argomenti teorici trattati e sulle esercitazioni pratiche assegnate settimanalmente durante il corso. Le domande di teoria valutano l'apprendimento dei concetti teorici insegnati. Le analisi empiriche sono condotte utilizzando l'ambiente R, RStudio e RMarkdown e permettono di verificare la capacità degli studenti di applicare modelli statistici avanzati a dati reali o simulati e di elaborare report riproducibili che descrivano i dati, le procedure e i risultati ottenuti. Durante l'esame è consentito l'utilizzo del materiale di studio e del codice R implementato durante il corso. Ogni domanda avrà un punteggio di circa 3 o 4 punti. Lo studente supera l'esame con una votazione di almeno 18/30.

Orario di ricevimento

Settimanalmente, secondo gli orari indicati nella pagina elearning del corso.

Sustainable Development Goals

SALUTE E BENESSERE | ISTRUZIONE DI QUALITÁ | RIDURRE LE DISUGUAGLIANZE | PARTNERSHIP PER GLI OBIETTIVI