UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Modelling

2526-1-FDS02Q040

Obiettivi

L'insegnamento rientra nelle aree di apprendimento delle scienze statistiche, informatica e delle scienze sociali. L'attività didattica mira a fornire agli studenti una preparazione circa le procedure analitiche ed inferenziali riguardanti modelli statistici avanzati quali: la regressione lineare multipla e alcune sue estensioni, alcuni modelli lineari generalizzati e approcci basati sui modelli per l'analisi dei gruppi, quali i modelli mistura Gaussiani univariati e multivariati, nonché modelli predittivi. Lo studente acquisisce anche competenze relative alla messa in atto di ricerche riproducibili e replicabili. Inoltre, sviluppa abilità comunicative scritte, poiché è richiesta la redazione di testi che accompagnino i risultati delle analisi svolte.

Durante l'insegnamento, lo studente svilupperà i seguenti aspetti definiti nei descrittori di Dublino:

(DdD 1) Conoscenza e capacità di comprensione

- Descrivere i principali fondamenti teorici e metodologici dei modelli statistici avanzati, inclusi la regressione lineare multipla, i modelli lineari generalizzati e gli approcci basati su modelli per il clustering tramite modelli miscuglio.
- Comprendere e spiegare le assunzioni alla base di questi modelli, insieme alle tecniche di stima come i minimi quadrati e la massima verosimiglianza.
- Riconoscere l'importanza della riproducibilità e replicabilità nella ricerca statistica, soprattutto attraverso l'integrazione di R, RStudio e RMarkdown nei flussi di lavoro per l'analisi dei dati.

(DdD 2) Applicazione della conoscenza e della comprensione

- Condurre analisi statistiche utilizzando R e RMarkdown, applicando modelli statistici appropriati a dati reali e simulati provenienti da diversi ambiti (ad esempio, economia, finanza, scienze sociali e scienze ambientali).
- Implementare tecniche avanzate come il bootstrap, la diagnostica per i modelli statistici, la selezione dei modelli e la classificazione.
- Produrre report riproducibili che includano codice, analisi, interpretazioni e commenti critici.

(DdD 3) Capacità di giudizio

- Valutare criticamente la validità delle assunzioni del modello e le implicazioni dei risultati empirici.
- Valutare la performance dei modelli statistici utilizzando diagnostiche e metriche appropriate (ad esempio, curva ROC, AUC, errore di classificazione).
- Scrivere rapporti tecnici o report divulgativi esprimendo giudizi ben motivati riguardo alla scelta, implementazione e limiti dei metodi statistici in contesti pratici.

(DdD 4) Capacità comunicative

- Comunicare efficacemente i risultati statistici attraverso report ben strutturati, annotati e riproducibili in RMarkdown.
- Presentare ragionamenti analitici e interpretazioni in modo chiaro e scritto, soprattutto nel descrivere le fasi e le motivazioni delle analisi empiriche.

(DdD 5) Capacità di apprendimento

- Continuare lo studio di argomenti avanzati di statistica in modo autonomo, utilizzando come base i materiali didattici, la bibliografia e gli esempi di codice forniti.
- Approfondire la comprensione dei metodi di data science e applicarli in contesti interdisciplinari, supportato dalla capacità di gestire e strutturare autonomamente le proprie attività di apprendimento e ricerca.

In particolare:

Conoscenza e comprensione

Lo studente viene introdotto ai 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 i metodi 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. L'insegnamento permette agli studenti di acquisire solidi elementi di teoria e di sviluppare le applicazioni attraverso un approccio di "problem solving". L'insegnamento è inerente alla scienza dei dati, conoscenza oggi essenziale in ogni ambiente lavorativo. L'insegnamento è obbligatorio per il successivo percorso di studio.

Capacità di applicare conoscenza e comprensione

L'insegnamento 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 anche attraverso tutoraggi. L'insegnamento consente agli studenti di acquisire solide basi teoriche e capacità di applicare modelli statistici moderni per l'analisi dei dati, oltre che a sviluppare capacità per condurre ricerche in modo riproducibile e replicabile. 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 dell’insegnamento, 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 di 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. L'algoritmo expectation-maximization viene introdotto come strumento per la stima di massima verosimiglianza dei parametri dei modelli di classificazione. Questi ultimi vengono anche introdotti come modelli probabilistici che permettono “supervised learning”. L’insegnamento 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 condotte su dati reali e simulati ed i commenti al codice e ai risultati ottenuti.

Programma esteso

Nell’introduzione all’insegnamento vengono richiamati alcuni concetti dell’inferenza statistica e dell’inferenza causale. Vengono richiamati i concetti di verosimiglianza e di inferenza Bayesiana.

La prima parte dell’insegnamento 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 dell’insegnamento 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 delle 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 di 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 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 miscuglio con componenti Gaussiane sono introdotti come metodi di classificazione e di clustering per l’apprendimento supervisionato. Questi permettono di individuare pattern (attraverso l’analisi discriminante) ottenuti con approcci probabilistici basati sulle stime delle probabilità a posteriori. I risultati dei modelli generativi vengono valutati utilizzando set di training e validation. Gli strumenti diagnostici basati sull’errore di classificazione, sul punteggio Brier, sulla curva ROC (Receiver Operating Characteristic) e sull’area sotto alla curva (AUC) vengono discussi e presentati nei contesti applicativi.

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. Durante l'attività didattica vengono discusse le soluzioni agli esercizi assegnati.

Prerequisiti

Per una più facile comprensione dei contenuti dell'insegnamento, è utile avere conoscenze di base in probabilità e di inferenza statistica ed i contenuti dell'insegnamento Fondamenti di Probabilità e Statistica. L'insegnamento 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

Tutte le lezioni si svolgono in presenza. Le spiegazioni teoriche sono integrate da applicazioni empiriche basate su dati reali e simulati provenienti da diversi ambiti quali l'economia, la finanza, la biologia, l'ecologia e le science ambientali. Le lezioni si svolgono presso il laboratorio informatico. Queste applicazioni sono sviluppate utilizzando diverse librerie del software statistico open-source R e RStudio, insieme all'interfaccia RMarkdown tramite la libreria knitr. Quest'approccio permette agli studenti di acquisire familiarità con i principi della riproducibilità della ricerca. Durante le lezioni vengono utilizzati i laboratori virtuali di Ateneo. Le ore previste di didattica erogativa sono 30 e quelle di didattica interattiva sono 12.

Ogni settimana vengono assegnati esercizi, incoraggiando gli studenti a redigere report in cui commentano il codice e forniscono una spiegazione del procedimento di analisi dei dati svolto, insieme a una valutazione critica dei risultati ottenuti. Gli studenti sono inoltre incoraggiati a svolgere gli esercizi assegnati anche in gruppo, promuovendo così l'apprendimento cooperativo. Le soluzioni vengono discusse durante le sessioni di tutoraggio, nelle quali gli studenti sono invitati ad una maggiore interazione tra loro e con il docente.

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 all'insegnamento. 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. Le video registrazioni delle lezioni e delle esercitazioni sono disponibili al termine della lezione nella pagina del corso.
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.

Pennoni, F. (2025). Statistical Modeling, Teaching Notes Theory and Applications in R. Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy.

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 2025 - Maggio 2025

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. L'esame è composto da una parte scritta con domande aperte e da una parte orale facoltativa. Durante l’insegnamento non sono previste prove intermedie ma è prevista l’acquisizione di 2 punti bonus con la consegna di alcuni esercizi che concorrono al punteggio finale. 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 l’insegnamento. 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. La prova mira anche a promuovere la capacità degli studenti di pianificare e gestire in modo efficace il tempo necessario per la stesura dell’elaborato. Durante l'esame è consentito l'utilizzo del materiale di studio e del codice R implementato durante l’insegnamento. 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 dell'insegnamento in modalità sia in presenza che on-line.

Sustainable Development Goals

SALUTE E BENESSERE | PARITÁ DI GENERE | CITTÀ E COMUNITÀ SOSTENIBILI | LOTTA CONTRO IL CAMBIAMENTO CLIMATICO