



UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Statistical Modelling

2627-1-FDS02Q040

Obiettivi

L'insegnamento rientra nelle aree di apprendimento delle scienze statistiche, dell'informatica e delle scienze sociali. Mira a fornire agli studenti conoscenze metodologiche e applicative relative a modelli statistici avanzati, tra cui la regressione lineare multipla e alcune sue estensioni, i modelli di regressione logistica, gli approcci basati su modelli per il clustering mediante modelli a mistura finita di distribuzioni Gaussiane univariate e multivariate, nonché modelli predittivi.

Durante le attività didattiche gli studenti sviluppano una comprensione critica delle assunzioni alla base dei modelli teorici attraverso applicazioni empiriche su dati reali e simulati. Acquisiscono inoltre competenze per condurre ricerche riproducibili e replicabili. In aggiunta, sviluppano abilità comunicative scritte necessarie per produrre testi che descrivano le procedure utilizzate nella risoluzione degli esercizi e per presentare e discutere i risultati delle analisi.

Al termine dell'insegnamento lo studente sarà in grado di:

(DdD 1) Conoscenza e capacità di comprensione

- Descrivere i principali fondamenti teorici dei modelli statistici avanzati, inclusi la regressione lineare multipla, la regressione logistica multipla e gli approcci basati su modelli per il clustering mediante modelli a mistura finita.
- Comprendere e spiegare le assunzioni alla base di tali modelli, insieme alle procedure di stima, quali i minimi quadrati e la massima verosimiglianza.
- Riconoscere la rilevanza della riproducibilità e della replicabilità nella ricerca attraverso l'integrazione di R, RStudio e RMarkdown nei flussi di lavoro per l'analisi dei dati.

(DdD 2) Capacità di applicare conoscenza e comprensione

- Sviluppare simulazioni da distribuzioni univariate e multivariate.
- Condurre analisi statistiche con R e RMarkdown su dati reali provenienti da diversi ambiti, quali economia, finanza, scienze sociali e scienze ambientali.
- Implementare tecniche statistiche avanzate, quali il ricampionamento bootstrap, la diagnostica dei modelli, la selezione dei modelli e metodi di previsione.
- Produrre report riproducibili che includano codice, analisi, interpretazioni e commenti critici.

(DdD 3) Autonomia di giudizio

Nel corso dell'insegnamento lo studente svilupperà la capacità di:

- 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 anche per la previsione, quali curva ROC, AUC ed errore di classificazione.
- Redigere brevi saggi tecnici o report esprimendo giudizi ben motivati sulla scelta, sull'implementazione e sui limiti dei metodi statistici in contesti applicativi.

(DdD 4) Abilità comunicative

Attraverso le attività dell'insegnamento lo studente sarà in grado di:

- Comunicare efficacemente i risultati statistici mediante report ben strutturati, annotati e riproducibili in RMarkdown.
- Presentare in forma scritta ragionamenti analitici e interpretazioni in modo chiaro, in particolare nella descrizione delle fasi e delle motivazioni delle analisi empiriche.

(DdD 5) Capacità di apprendimento

Al termine dell'insegnamento lo studente avrà sviluppato la capacità di:

- Proseguire autonomamente lo studio di argomenti avanzati di statistica, 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 di ricerca.

Conoscenza e comprensione

Lo studente viene introdotto ai modelli statistici avanzati per l'analisi di dati con diverse tipologie di variabili risposta. Vengono illustrate le assunzioni teoriche rilevanti, considerando i metodi di stima della massima verosimiglianza e dei minimi quadrati per i parametri dei modelli. L'analisi dei dati è condotta utilizzando il software R e l'ambiente RMarkdown, che consente di creare documenti riproducibili contenenti codice, risultati e commenti. Le applicazioni riguardano dati reali e simulati provenienti da diversi ambiti, quali economia, finanza e scienze sociali. Lo studente è inoltre incoraggiato a fornire una valutazione critica dei risultati ottenuti dalle analisi empiriche. In questo modo sviluppa autonomia di giudizio e affina le abilità comunicative. L'insegnamento permette agli studenti di acquisire solide basi teoriche e applicative; riguarda la scienza dei dati, conoscenza oggi essenziale in ogni ambiente lavorativo, ed è propedeutico agli insegnamenti successivi previsti dal piano di studio. Al termine dell'insegnamento, grazie ai materiali forniti, alle dispense del docente corredate da una bibliografia completa, al codice R e all'interfaccia RMarkdown, lo studente è in grado di proseguire autonomamente l'approfondimento della disciplina.

Capacità di applicare conoscenza e comprensione

L'insegnamento fornisce competenze nell'utilizzo della semantica del software open-source R per l'analisi descrittiva di dati multivariati e per la stima dei parametri di modelli univariati e multivariati. Attraverso R e RStudio, gli studenti imparano a impostare in modo sistematico il ragionamento statistico mediante l'analisi dei dati e la redazione di report che illustrano codice, analisi e risultati.

La teoria è affiancata da applicazioni pratiche, sviluppate anche durante le lezioni di tutoraggio. L'insegnamento consente agli studenti di acquisire solide basi teoriche e capacità di applicare metodi statistici moderni per l'analisi dei dati, oltre a sviluppare la capacità di condurre ricerche riproducibili e replicabili.

Contenuti sintetici

Nella prima parte dell'insegnamento, dopo una breve introduzione all'impianto concettuale dell'inferenza statistica e ad alcuni aspetti legati alla causalità, viene illustrata la procedura di ricampionamento nota come bootstrap, finalizzata a ottenere misure di accuratezza per stimatori di interesse. Successivamente viene presentato il modello di regressione lineare multipla con le relative assunzioni. Sono introdotti i metodi dei minimi quadrati ordinari e della stima di massima verosimiglianza, insieme alle loro proprietà statistiche. Vengono inoltre trattate le misure di adattamento, la diagnostica della regressione e la previsione.

Sono discussi i modelli lineari generalizzati, con particolare riferimento al modello di regressione logistica multipla. Sono inoltre introdotti modelli probabilistici di classificazione per l'apprendimento supervisionato, quali i modelli a mistura finita di distribuzioni Gaussiane. L'insegnamento fornisce competenze nell'uso della semantica del software R, utilizzando le librerie RMarkdown tramite il pacchetto knitr per integrare codice, risultati delle analisi condotte su dati reali e simulati e commenti al codice e ai risultati ottenuti.

Programma esteso

L'insegnamento si apre con un'introduzione al quadro dell'inferenza statistica e ad alcuni concetti connessi all'inferenza causale.

- La prima parte introduce il metodo di ricampionamento noto come bootstrap per la determinazione dell'errore standard come misura di accuratezza. Il metodo è applicato a diversi stimatori utilizzando dati rilevanti provenienti da diverse fonti, ad esempio psicologia, ambiente e altri ambiti applicativi.
- La prima parte dell'insegnamento riguarda inoltre il modello di regressione lineare multipla e i metodi di stima a minimi quadrati e di massima verosimiglianza. Le proprietà degli stimatori dei minimi quadrati e degli stimatori di massima verosimiglianza sono discusse sulla base delle assunzioni del modello.
- Durante l'insegnamento, le conoscenze dello studente sulle distribuzioni univariate vengono estese alle distribuzioni Gaussiane bivariate e multivariate. Da tali distribuzioni sono simulate realizzazioni casuali, che vengono rappresentate graficamente anche utilizzando le curve di livello.
- Sono considerati diversi strumenti diagnostici per la valutazione del modello basati sui residui di regressione, con particolare enfasi sulla multicollinearità mediante il Variance Inflation Factor, sugli outlier, sui punti influenti e sui punti di leva. Viene affrontato il problema della selezione delle variabili esplicative più rilevanti mediante criteri informativi, tra cui il criterio di informazione di Akaike (AIC). Gli studenti imparano inoltre a valutare le previsioni del modello.
- Vengono richiamate le distribuzioni di Bernoulli e Binomiale e le definizioni di odds, log-odds e odds ratio. È introdotto il modello di regressione logistica multipla per una variabile risposta binaria, con particolare enfasi sull'interpretazione dei coefficienti di regressione. Viene discusso l'uso del modello logistico per l'apprendimento supervisionato e la validazione dei risultati in termini di sensibilità e specificità. È inoltre presentata la rappresentazione della curva ROC (Receiver Operating Characteristic).

Gli aspetti teorici sono introdotti insieme ad applicazioni empiriche basate su dati provenienti da diversi ambiti, quali economia, finanza, biologia, ecologia e scienze ambientali. Le applicazioni sono sviluppate con il software statistico R e con RStudio, utilizzando diverse librerie insieme all'interfaccia RMarkdown e alla libreria knitr. Tale approccio mira a familiarizzare gli studenti con i principi della ricerca riproducibile.

Agli studenti è richiesto di scrivere report riproducibili nei quali commentino criticamente il codice e i risultati delle analisi empiriche. L'apprendimento cooperativo è incoraggiato attraverso esercizi assegnati. Ogni settimana vengono proposti esercizi e gli studenti sono invitati a redigere report in cui commentano il codice, forniscono una spiegazione del processo di analisi svolto e descrivono criticamente i risultati ottenuti. Durante le attività didattiche vengono discusse le soluzioni degli esercizi assegnati.

Prerequisiti

Per una più agevole comprensione dei contenuti dell'insegnamento, è consigliata la conoscenza dei contenuti dell'insegnamento di Fondamenti di Probabilità e Statistica. L'insegnamento presuppone una conoscenza preliminare dei seguenti argomenti: probabilità di un evento; funzione di distribuzione di probabilità; funzione di densità; funzione di ripartizione; legge della probabilità totale; indipendenza degli eventi; teorema di Bayes; valore atteso 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, Gaussian, t di

Student e chi-quadrato; grafici e misure numeriche per la descrizione dei dati; inferenza statistica; inferenza di massima verosimiglianza; conoscenze di base di analisi dei dati multivariati e di algebra lineare. È inoltre richiesta la conoscenza della semantica di base del linguaggio di programmazione nell'ambiente R.

Modalità didattica

Le lezioni erogate in presenza coprono gli aspetti teorici e sono integrate da esercitazioni pratiche che permettono agli studenti di apprendere la teoria e di applicare i modelli per analizzare dati reali e simulati. Le lezioni si svolgono presso il laboratorio informatico. Settimanalmente vengono assegnati esercizi riepilogativi da svolgere come compito a casa, con l'obiettivo di consolidare l'apprendimento della teoria e delle sue applicazioni.

Durante l'insegnamento, con il supporto di R nell'ambiente RStudio e dell'interfaccia RMarkdown, gli studenti imparano inoltre a creare documenti riproducibili. Sono incoraggiati ad affrontare problemi applicativi anche con l'obiettivo di sviluppare forme di apprendimento cooperativo. Sono previste sessioni di tutoraggio per aiutare gli studenti nello svolgimento degli esercizi e nel confronto delle soluzioni. Le ore previste di didattica erogativa sono 30, mentre quelle di didattica interattiva sono 12. Nella seconda parte di ciascuna lezione di tre ore si promuove il coinvolgimento interattivo degli studenti. Le esercitazioni pratiche sono svolte in modalità interattiva e in presenza nel laboratorio informatico. Le videoregistrazioni asincrone delle lezioni e delle esercitazioni sono rese disponibili sulla piattaforma e-learning.

Materiale didattico

Il materiale didattico consiste principalmente nelle dispense preparate dal docente, che coprono sia gli argomenti teorici sia le applicazioni sviluppate con il software R. Tutti i file sono disponibili sulla pagina dell'insegnamento nella piattaforma e-learning dell'Università. Inoltre, al termine di ciascuna lezione il docente pubblica le slide, i programmi di calcolo e i dataset utilizzati. Settimanalmente vengono assegnati esercizi e alcune soluzioni sono fornite e discusse. Sulla stessa pagina sono pubblicati anche esempi di testi d'esame.

I riferimenti principali saranno elencati nella bibliografia delle dispense. Tra gli altri, si segnalano i seguenti testi, alcuni dei quali disponibili anche presso la biblioteca o come 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.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R, 2nd Edition, New York, Springer.

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

Pennoni, F. (2026). Handout for Statistical Models II: Theory and Applications with R. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2026). 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

Secondo semestre, marzo-maggio 2027. L'insegnamento è erogato in lingua inglese e l'esame si svolge in lingua inglese.

Modalità di verifica del profitto e valutazione

Le seguenti modalità di verifica dell'apprendimento si applicano sia agli studenti frequentanti sia agli studenti non frequentanti le lezioni tenute in laboratorio. L'esame consiste in una prova scritta con domande aperte e in una prova orale facoltativa. La prova scritta ha una durata massima complessiva di un'ora e mezza e si svolge nel laboratorio informatico. Durante l'esame gli studenti devono rispondere a domande aperte di teoria e risolvere esercizi ed esercitazioni pratiche con dati analoghi a quelli assegnati settimanalmente durante l'insegnamento. Le domande di teoria valutano la comprensione dei concetti teorici trattati durante il corso. Le analisi empiriche sono condotte utilizzando l'ambiente R, RStudio e RMarkdown. Tali analisi consentono agli studenti di dimostrare la capacità di comprendere e risolvere problemi applicando modelli statistici avanzati a dati reali o simulati e di produrre report riproducibili che descrivano il codice e illustrino i risultati ottenuti.

La prova mira inoltre a promuovere la capacità degli studenti di pianificare e gestire in modo efficace il tempo necessario per completare l'elaborato scritto. Durante l'esame è consentito l'utilizzo del materiale di studio e del codice R implementato durante l'insegnamento. Ciascuna domanda ha un punteggio indicativo di circa 3 punti. Per superare la prova lo studente deve conseguire una valutazione di almeno 18/30.

Orario di ricevimento

Settimanalmente, secondo gli orari indicati nella pagina e-learning dell'insegnamento, in modalità sia in presenza sia online.

Sustainable Development Goals

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