UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models II

2526-2-F8203B042-F8203B013M

Obiettivi formativi

L'insegnamento rientra nelle aree di apprendimento delle scienze statistiche, dell'informatica e delle scienze sociali. Mira a fornire agli studenti una preparazione riguardanti i seguenti approcci inferenziali: bootstrap non parametrico, distribuzione Gaussiana multivariata, modelli lineari generalizzati per dati di conteggio, modelli mistura Gaussiani univariati e multivariati, nonché modelli predittivi.
Durante l'attività didattica lo studente sviluppa una comprensione critica delle assunzioni alla base dei modelli teorici, attraverso applicazioni empiriche su dati reali e simulati. 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.

Conoscenza e comprensione

L'insegnamento consente agli studenti di:

• Analizzare i dati utilizzando modelli statistici avanzati sviluppati per variabili risposta univariate e multivariate, sia di natura categoriale che continua.
• Sviluppare la conoscenza dei metodi di simulazione.
• Utilizzare la semantica del software R, anche attraverso l'ambiente RMarkdown, per sviluppare un metodo di ricerca replicabile e riproducibile. I documenti generati includono il codice, i risultati e i commenti al codice e alle analisi svolte.
• Interpretare i risultati delle elaborazioni in modo rigoroso sviluppando capacità espressive e di sintesi testuale anche per scopi divulgativi rivolti a un pubblico non accademico. In questo modo sviluppa autonomia di giudizio e affina le proprie abilità comunicative.

Capacità di applicare conoscenza e comprensione

L'insegnamento consente agli studenti di:

• Condurre l'inferenza statistica tramite tecniche di ricampionamento (bootstrap);

• Stimare, selezionare ed interpretare i modelli di miscugli di distribuzioni per popolazioni eterogenee;
• Concettualizzare i modelli a variabili latenti, stimare i parametri con il principio di massima verosimiglianza e interpretare i risultati;
• Applicare le conoscenze teoriche per analizzare dati di diverse tipologie derivanti dagli ambiti applicativi del corso di studio quali l'epidemiologia, la medicina, la biologia, la genetica e la salute pubblica.
• Implementare codice con il linguaggio del software open source R per le analisi descrittive ed inferenziali adottando un approccio open source che garantisca la riproducibilità e la replicabilità delle analisi.

L'insegnamento permette agli studenti di acquisire solidi elementi di teoria e di sviluppare le applicazioni pratiche attraverso un approccio di "problem solving". L'insegnamento si inserisce nell'ambito della scienza dei dati, conoscenza oggi essenziale per i contesti lavorativi di sbocco degli studenti del corso di laurea in Biostatistica. Al termine dell'insegnamento, grazie al materiale fornito (le dispense del docente corredate da un'ampia bibliografia, i codici per i software R e l'interfaccia RMarkdown), lo studente è in grado di proseguire in modo autonomo nell'approfondimento di questa disciplina.

Contenuti sintetici

Nella prima parte dell'insegnamento vengono richiamate le principali distribuzioni probabilistiche che si utilizzano per simulare delle realizzazioni da variabili casuali. Viene presentato il procedimento di ricampionamento noto come bootstrap per ottenere misure di precisione in ambito non parametrico per alcuni stimatori di interesse.
Nella seconda parte dell'insegnamento viene presentato il modello lineare generalizzato basato sulla distribuzione binomiale negativa e viene introdotto l'algoritmo Expectation-Maximization (EM) come metodo di imputazione dei dati mancanti utilizzando le stime di massima verosimiglianza dei parametri di un modello lineare generalizzato. Dopo aver presentato la distribuzione Gaussiana multivariata si illustrano i modelli miscuglio Gaussiani. Vengono descritti i passi dell'algoritmo EM per la stima di massima verosimiglianza dei parametri dei modelli e dei modelli a variabili latenti con distribuzione discreta. Le lezioni di teoria sono affiancate da esercitazioni pratiche. L'insegnamento fornisce competenze nell'uso della semantica del software R, utilizzando anche la libreria RMarkdown tramite la libreria knitr per integrare il codice, i risultati delle analisi ed i commenti.

Programma esteso

La prima parte dell'attività didattica riguarda i metodi lineari congruenziali per la generazione di numeri pseudo-casuali ed i test grafici per la verifica della pseudo-casualità. La teoria è affiancata da esempi di simulazioni di dati da alcune distribuzioni probabilistiche, tra cui la distribuzione Gaussiana, beta e di Poisson.
Nella seconda parte dell'attività didattica, 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. Si illustrano gli intervalli di confidenza ottenuti sia con il metodo del percentile che con il metodo BCA che permette di correggere per la distorsione.
Viene introdotto il modello autoregressivo di Poisson per dati di conteggio e l'analogo modello basato sulla distribuzione Binomiale Negativa per tener conto dell'overdispersion. I modelli vengono applicati all'analisi dei conteggi relativi alle capacità riproduttive delle specie animali.
L'algoritmo Expectation-Maximization viene illustrato negli step E e M sia come algoritmo di stima di massima verosimiglianza dei parametri dei modelli a variabili latenti con distribuzione discreta, sia come metodo di imputazione dei valori mancanti di una tabella a doppia entrata utilizzando i parametri un modello lineare generalizzato. Si introducono sia la formulazione che le caratteristiche della distribuzione normale multivariata. Si simulano i valori da questa distribuzione utilizzando diverse matrici di varianza-covarianza.
Si illustrano i modelli miscuglio (finite mixture models) univariati e multivariati per variabili risposta quantitative assumendo una distribuzione di Gauss per le componenti del miscuglio. In particolare si considera la stima della densità e la classificazione delle unità statistiche con il metodo della massima probabilità a posteriori.

La teoria è affiancata da esercitazioni pratiche in cui vengono sviluppate, nell'ambiente R e con l'ausilio del marcatore di testo RMarkdown, numerose applicazioni volte all'analisi e all'adattamento dei modelli statistici per dati reali e simulati riguardanti gli ambiti della biostatistica. Le principali librerie del software R utilizzate sono skimr, MASS, boot, bootstrap, mclust, MultiLCIRT. Lo studente è incoraggiato ad elaborare documenti riproducibili in cui commenta in forma testuale il codice ed i risultati delle analisi in modo critico anche tramite apprendimento cooperativo. Settimanalmente vengono assegnati degli esercizi e gli studenti nello svolgimento sono incoraggiati a scrivere reports in cui commentano il codice, ed offrono una spiegazione del procedimento di analisi svolto oltre ad una descrizione critica dei risultati ottenuti. Durante l'attività didattica vengono discusse le soluzioni agli esercizi assegnati.

Prerequisiti

Per una più agevole comprensione dei contenuti dell'insegnamento è necessario conoscere le nozioni di Probabilità e di Inferenza Statistica e la semantica di base del linguaggio di programmazione in ambiente R.

Metodi didattici

Sono previste lezioni frontali, le lezioni di teoria sono affiancate da esercitazioni pratiche che consentono agli studenti di apprendere tramite problem solving analizzando dati reali e simulati. Le lezioni si svolgono in laboratorio informatico. Settimanalmente vengono assegnati degli esercizi di riepilogo relativi al programma svolto. Durante l'insegnamento con l'ausilio di R nell'ambiente RStudio e l'interfaccia di RMarkdown, gli studenti imparano ad elaborare documenti riproducibili che contengono codice, descrizioni e commenti ai risultati delle analisi. Sono incoraggiati a collaborare tra di loro nella risoluzione dei problemi applicativi, al fine di promuovere l'apprendimento cooperativo. Le ore previste di didattica erogativa sono 30 e quelle di didattica interattiva sono 17 e queste ultime che vengono prevalentemente svolte. Nella seconda parte delle lezioni che constano di 3 ore vengono si tende a coinvolgere gli studenti in modo interattivo. Le esercitazioni sono svolte in modalità interattiva in presenza presso il laboratorio informatico. Vengono rese disponibili nella pagina di e-elearning le video-registrazioni in asincrono sia delle lezioni che delle esercitazioni.

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 frontali. L'esame è in forma scritta con orale facoltativo, non sono previste prove intermedie ma durante lo svolgimento delle lezioni è prevista l'acquisizione di 2 punti bonus con la consegna di alcuni esercizi che concorrono al punteggio finale. L'esame scritto ha una durata massima di due ore e si svolge in laboratorio informatico. Le domande aperte di teoria a cui gli studenti devono rispondere mirano a valutare la comprensione dei concetti essenziali dell'inferenza statistica condotta con metodi avanzati, mentre gli esercizi applicativi condotti utilizzando l'ambiente R, RStudio e RMarkdown, permettono di verificare la capacità degli studenti di applicare le metodologie proposte nonché 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 e personalmente dallo studente. Ogni punto di ogni esercizio ha una valutazione di circa 3 punti. Lo studente supera l'esame con una votazione non inferiore a 18/30.

Testi di riferimento

Il materiale didattico principale consiste nelle dispense preparate dal docente, che coprono, gli argomenti teorici, le applicazioni sviluppate con il software R, gli esercizi e le soluzioni. 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, e le relative soluzioni. Sulla stessa pagina web sono disponibili degli esempi del testo d'esame.
I riferimenti bibliografici principali sono elencati nella bibliografia delle dispense alcuni dei quali sono i seguenti che risultano disponibili presso la biblioteca di Ateneo anche in formato ebook:

I principali testi di riferimento sono elencati nella bibliografia delle dispense alcuni dei quali sono i seguenti che sono anche disponibili in ebook presso la biblioteca dell'Ateneo:

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.

Blitzstein, J. K., Hwang, J. (2014). Introduction to probability, Chapman & Hall/CRC.

Gentle, J. E., Hardle W., Mori Y. (2004). Handbook of computational statistics. Springer-Berlin.

Lange, K. (2010). Numerical analysis for statisticians, 2nd Edition, Springer, New York.

Pennoni, F. (2025). Dispensa di Modelli Statistici II, parte di teoria e applicazioni con R. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Periodo di erogazione dell'insegnamento

Semestre I, ciclo I, Settembre-Novembre 2025

Lingua di insegnamento

Il corso viene erogato in lingua italiana. Gli studenti Erasmus possono utilizzare il materiale didattico predisposto in lingua inglese e fornito dal docente su richiesta. Possono inoltre richiedere di svolgere la prova d'esame in lingua inglese.

Sustainable Development Goals

SALUTE E BENESSERE | RIDURRE LE DISUGUAGLIANZE | LOTTA CONTRO IL CAMBIAMENTO CLIMATICO