UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models II

2324-2-F8203B042-F8203B013M

Obiettivi formativi

Il corso permette allo studente di apprendere le procedure analitiche ed inferenziali riguardanti modelli statistici avanzati e di sviluppare una conoscenza critica delle assunzioni dei modelli alla base della teoria attraverso le applicazioni a dati reali.

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.
• Apprendere l'implementazione di studi di simulazione.
• Utilizzare la semantica del software R, anche attraverso l'ambiente RMarkdown, per creare documenti riproducibili che includono il codice, i risultati e i commenti delle analisi, con l'obiettivo di garantire la replicabilità dei risultati.
• Interpretare i risultati delle elaborazioni in modo rigoroso, fornendo una descrizione completa degli stessi, anche per scopi divulgativi rivolti a un pubblico non accademico.

Capacità di applicare conoscenza e comprensione

L'insegnamento permette allo studente di:

Condurre l'inferenza statistica tramite tecniche di ricampionamento (bootstrap);
Stimare, selezionare ed interpretare i modelli di miscugli di distribuzioni per popolazioni eterogenee; stimare modelli a variabili latenti e interpretare i risultati;
Applicare le conoscenze teoriche per analizzare dati derivanti dagli ambiti applicativi del corso di studio quali l'epidemiologia, la medicina, la biologia, la genetica e la salute pubblica.
Implementare codice per analisi descrittive ed inferenziali con il software R.

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 Biostatistica.

Contenuti sintetici

Nella prima parte del corso 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 del corso 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 introdotto i modelli miscuglio Gaussiani, vengono descritti i passi dell'algoritmo EM per la stima di massima verosimiglianza dei parametri di questi modelli e dei modelli a variabili latenti con distribuzione discreta. Le lezioni di teoria sono affiancate da esercitazioni pratiche. Il corso 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 del corso riguarda i metodi di simulazione come i metodi lineari congruenziali per la generazione di numeri pseudo-casuali, i test grafici e statistici, tra cui il test Kolmogorov-Smirnov e il test Chi- Quadrato per la verifica della pseudo-casualità. La teoria è affiancata da esempi applicativi utilizzando diversi modelli distributivi dai quali vengono simulate realizzazioni quali: la distribuzione esponenziale, la distribuzione binomiale e di la distribuzione Gaussiana.
Nella seconda parte del corso si introducono i principali metodi di ricampionamento: Jackknife e bootstrap, gli intervalli di confidenza bootstrap 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 dei soggetti affetti da COVID-19 in base alle serie dei dati nazionali settimanali forniti ufficialmente in Italia dall'inizio della pandemia. L'algoritmo Expectation-Maximization viene illustrato dettagliatamente sia come algoritmo di stima dei parametri dei modelli a variabili latenti discrete sia come metodo per l'imputazione dei valori mancanti in una tabella a doppia entrata in relazione un modello lineare generalizzato.
Si illustrano i modelli miscuglio (finite mixture models) per variabili risposta sia quantitative che categoriali assumendo una distribuzione di Gauss per le componenti del miscuglio. In particolare si considera la stima della densità e alla classificazione delle unità statistiche con il metodo della massima probabilità a posteriori.

La teoria è affiancata da esercitazioni 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, dplyr, tscount, boot, bootstrap, mclust, MultiLCIRT. Lo studente è incoraggiato ad elaborare documenti riproducibili in cui commenta il codice ed i risultati delle analisi in modo critico anche tramite apprendimento cooperativo.

Prerequisiti

Per una più agevole comprensione dei contenuti del corso 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 riguardanti la parte teorica sui concetti di base dei modelli statistici. Le lezioni di teoria sono affiancate da esercitazioni pratiche che consentono agli studenti di apprendere la teoria applicando i modelli per l'analisi di dati reali e simulati. Le lezioni si svolgono in laboratorio informatico. Settimanalmente vengono assegnati degli esercizi di riepilogo relativi al programma svolto. Durante il corso 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.

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 che si svolgono in presenza. L’esame è in forma scritta con orale facoltativo, non sono previste prove intermedie. Durante il corso non sono previste 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 due ore 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 2 o 3 punti. Lo studente supera l'esame con una votazione di almeno 18/30.

Testi di riferimento

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.

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. (2023). 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, Ottobre-Novembre 2023

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