UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici II

2425-2-F8203B042-F8203B013M

Obiettivi formativi

Il corso rientra nelle aree di apprendimento delle scienze statistiche, dell'informatica e delle scienze sociali. Il corso mira a fornire agli studenti una preparazione circa le procedure analitiche ed inferenziali riguardanti: il bootstrap non parametrico, la distribuzione Gaussiana multivariata, i modelli lineari generalizzati per dati di conteggio, e i modelli mistura Gaussiani univariati e multivariati, nonché modelli predittivi. Il corso mira a sviluppare una conoscenza critica delle assunzioni dei modelli alla base della teoria attraverso applicazioni empiriche con dati reali e simulati.

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 metodi di simulazione.
• Utilizzare la semantica del software R, anche attraverso l'ambiente RMarkdown, per apprendere 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 anche per scopi divulgativi rivolti a un pubblico non accademico.

Capacità di applicare conoscenza e comprensione

• 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 software open source R per le analisi descrittive ed inferenziali.

Il corso permette agli studenti di acquisire solidi elementi di teoria e di sviluppare le applicazioni attraverso un approccio di “problem solving”. Il corso è inerente alla scienza dei dati, conoscenza oggi essenziale 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 di simulazioni di dati da diverse distribuzioni probabilistiche quali: la distribuzione esponenziale, la distribuzione binomiale e di la distribuzione Gaussiana.
Nella seconda 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. Si considerano 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 due 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 con il metodo della massima verosimiglianza dei parametri dei modelli a variabili latenti con distribuzione discreta 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 assumendo una distribuzione di Gauss per le componenti del miscuglio, sia categoriali. 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 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.

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 il corso vengono discusse le soluzioni agli esercizi assegnati.

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 in presenza; 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. 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. Le ore previste di didattica erogativa sono 30 e quelle di didattica interattiva sono 24 e comprendono le lezioni di esercitazione.

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. L'esame scritto ha una durata di circa 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 applicativi basandosi sugli argomenti teorici trattati e sulle esercitazioni pratiche assegnate settimanalmente durante il corso. Le domande di teoria valutano l'apprendimento dei concetti essenziali dell’inferenza statistica con metodi avanzati. Le analisi empiriche devono essere condotte utilizzando l'ambiente R, RStudio e RMarkdown e 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. Durante l'esame è consentito l'utilizzo del materiale di studio e del codice R implementato durante il corso. Ogni domanda avrà un punteggio variabile da 2 a 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 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, 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:

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, Teoria e Applicazioni con R. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

R Core Team (2022). 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 2024

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