UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici II

2223-2-F8203B042-F8203B013M

Obiettivi formativi

Il corso introduce alle procedure analitiche ed inferenziali condotte tramite modelli statistici avanzati e alle simulazioni utili per l'inferenza statistica con l'intento di sviluppare una conoscenza critica delle assunzioni dei modelli alla base della teoria.

Conoscenza e comprensione
L'insegnamento permette allo studente di:

- analizzare i dati con modelli statistici sviluppati per variabili risposta sia categoriali che continue;
- implementare le simulazioni;
- servirsi della semantica del software R anche tramite l'ambiente RMarkdown per creare dei documenti riproducibili contenenti il codice, risultati ed i commenti delle analisi;
- interpretare i risultati delle elaborazioni in modo rigoroso, fornendo una descrizione esaustiva degli stessi anche per finalità divulgative.

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 interpretarne i risultati;
- applicare le conoscenze teoriche per l'analisi dei dati derivanti dagli ambiti applicativi del corso di studio quali l'epidemiologia, la medicina, la biologia, la genetica e la salute pubblica.
- implementare il codice riferito al software R.

Lo studente viene incoraggiato a presentare la teoria ed i risultati delle applicazioni in modo organico.

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 (biostatistico/statistico/demografico e affini).

Contenuti sintetici

Nella prima parte del corso vengono trattati i concetti di probabilità utili per simulare delle realizzazioni da variabili casuali. Vengono introdotte le principali procedure di ricampionamento utilizzate per l'inferenza statistica: bootstrap e Jackknife.
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. I passi dell'algoritmo sono anche illustrati in relazione alla stima dei parametri dei modelli miscuglio (finite mixture models) e dei modelli a variabili latenti. Le lezioni di teoria sono affiancate da esercitazioni pratiche in cui lo studente approfondisce anche la conoscenza della semantica del software R.

Programma esteso

La prima parte del corso riguarda i metodi di simulazione e comprende 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. Vengono esaminati alcuni metodi per la generazione di realizzazioni da variabili casuali. La teoria è affiancata da esempi applicativi utilizzando diversi modelli distributivi: la distribuzione esponenziale, binomiale e di Gauss. Si considera la convoluzione di variabili casuali e la generazione di realizzazioni dalla stessa.
Nella seconda parte si introducono i principali metodi di ricampionamento: Jackknife e bootstrap. Si illustra l'utilizzo del bootstrap per l'inferenza statistica, in particolare per il calcolo degli intervalli di confidenza ottenuti con il metodo del percentile e con la correzione 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 giornalieri. L'algoritmo Expectation-Maximization viene illustrato dettagliatamente sia come algoritmo di stima dei parametri dei modelli a variabili latenti sia come metodo per l'imputazione dei valori mancanti in una tabella a doppia entrata utilizzando le stime di massima verosimiglianza dei parametri di un modello lineare generalizzato. Si introducono i modelli miscuglio per variabili risposta sia quantitative che categoriali assumendo una distribuzione di Gauss per le componenti del miscuglio. Si presta particolare attenzione all'interpretazione dei risultati rispetto alla 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 durante le esercitazioni sono skimr, MASS, dplyr, tscount, boot, bootstrap e mclust. 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 permettono allo studente di sviluppare la tecniche di analisi dei dati. Le lezioni si svolgono in laboratorio informatico. Settimanalmente vengono assegnati degli esercizi di riepilogo basati sull'applicazione dei modelli proposti a dati reali o simulati 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. Gli stessi vengono incoraggiati ad affrontare il problema applicativo con lo scopo ulteriore di sviluppare l'apprendimento cooperativo.

Modalità di verifica dell'apprendimento

Le seguenti modalità di verifica dell'apprendimento riguardano sia gli studenti che non frequentanti. L'esame è in forma scritta con orale facoltativo, non sono previste prove intermedie. L'esame scritto ha durata di circa due ore e si svolge presso il laboratorio informatico. Durante la prova occorre rispondere a domande aperte di teoria e risolvere gli esercizi alla luce degli argomenti teorici sviluppati durante il corso. Le domande di teoria permettono di verificare l'apprendimento delle nozioni teoriche impartite durante il corso. Le analisi empiriche condotte tramite l'ambiente R, Rstudio e RMarkdown permettono di verificare la capacità di comprensione del problema, la sua risoluzione tramite l'applicazione di modelli statistici avanzati a dati reali o simulati e l'elaborazione di report con la descrizione del procedimento e l'illustrazione dei risultati. L'esame è a libro aperto e gli studenti possono consultare il codice R utilizzato durante il corso. Lo studente supera la prova con una votazione almeno pari a 18/30.

Testi di riferimento

Il materiale didattico è costituito principalmente dalle dispense redatte dal docente. Queste riguardano sia gli argomenti di teoria che le applicazioni sviluppate con il software R oppure SAS, gli esercizi e le soluzioni. Il materiale predisposto dal docente viene reso disponibile nella pagina della piattaforma e-learning dell'ateneo dedicata al corso. Il docente pubblica al termine di ogni lezione: le slides, i programmi di calcolo, gli esercizi, i dataset, e le soluzioni degli esercizi. Nella stessa pagina vengono anche pubblicati alcuni testi delle precedenti prove d'esame.

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. (2022). 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 (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 2022

Lingua di insegnamento

Il corso viene erogato in lingua italiana. Gli studenti Erasmus possono utilizzare il materiale didattico predisposto in Inglese e richiedere al docente di svolgere la prova d’esame in lingua inglese.

Sustainable Development Goals

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