UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Statistical Models II

2122-2-F8203B042-F8203B013M

Obiettivi formativi

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

Conoscenza e comprensione

Questo insegnamento permette allo studente di:

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

Capacità di applicare conoscenza e comprensione

Questo insegnamento permette allo studente di:

- sviluppare l'inferenza statistica tramite le moderne tecniche del bootstrap;
- stimare, selezionare ed interpretare i modelli di miscugli di distribuzioni per popolazioni eterogenee;
- adattare i modelli a variabili latenti;

- applicare le conoscenze teoriche per l’analisi dei dati raccolti in svariati ambiti tra cui l'epidemiologia, la medicina, la biologia, la genetica e la salute pubblica;

- sviluppare del codice in ambiente R.

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

L'insegnamento fornisce i concetti essenziali per lo sviluppo dei metodi statistici parametrici e non parametrici sia in ambito teorico che applicato per i contesti lavorativi di sbocco degli studenti del corso di laurea in Biostatistica (biostatistico/statistico/demografico e affini). L’insegnamento risulta pertanto indispensabile per il successivo percorso universitario.

Contenuti sintetici

Nella prima parte del corso vengono impartiti i concetti di base per simulare delle realizzazioni da variabili casuali. Vengono introdotte le principali procedure di ricampionamento: bootstrap e Jackknife e la loro applicazione nell’ambito dell’inferenza statistica.

L’algoritmo Expectation-Maximization (EM) viene introdotto come metodo di imputazione dei dati mancanti attraverso la stima dei parametri di modelli lineari generalizzati. Viene illustrato il suo utilizzo per la stima dei parametri dei modelli miscuglio (finite mixture models) e a variabili latenti. Lo studente approfondisce le competenze nell'utilizzo della semantica del software R.

Programma esteso

La prima parte del corso riguarda i metodi di simulazione e concerne anche 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 tra cui metodo della trasformata inversa. 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 tramite gli intervalli di confidenza ottenuti con il metodo del percentile e con la correzione per la distorsione. 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 riferita ai conteggi tramite un modello lineare generalizzato. Si introducono i modelli miscuglio per variabili risposta sia quantitative che categoriali, considerando in particolare i miscugli con componenti Gaussiane. Si illustra la stima dei modelli miscuglio con l'algoritmo Expectation-Maximization e l'interpretazione dei risultati comprendente la stima della densità e la classificazione delle unità statistiche con le probabilità a posteriori.

La teoria è affiancata da numerose applicazioni per l’analisi di dati reali e simulati riguardanti gli ambiti della biostatistica in modo da facilitare anche l’apprendimento dell’ambiente R con l’ausilio del marcatore di testo RMarkdown. Le principali librerie utilizzate sono boot, bootstrap e mclust. Lo studente è incoraggiato ad elaborare documenti riproducibili e sviluppare i commenti ai risultati delle analisi in modo critico anche tramite apprendimento cooperativo.

Prerequisiti

Per una più agevole comprensione dei contenuti del corso è consigliato conoscere le nozioni di Probabilità e di Inferenza Statistica. Lo studente deve inoltre conoscere 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 capacità di analisi dei dati. Le lezioni si svolgono in laboratorio informatico. Vengono assegnati ogni settimana degli esercizi di riepilogo basati sull'applicazione dei modelli proposti a dati reali o simulati relativi alla parte di 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.

Durante il periodo di emergenza Covid-19 le lezioni si svolgeranno in modalità da remoto (lezioni videoregistate) con incontri periodici in videoconferenza tramite piattaforma webex e/o in presenza secondo le calendarizzazioni previste dall'ateneo e che verranno rese note nella pagina del corso.

Modalità di verifica dell'apprendimento

Le seguenti modalità di verifica dell'apprendimento riguardano sia gli studenti frequentanti che non frequentanti. L'esame è in forma scritta con orale facoltativo, non sono previste prove intermedie. L'esame scritto ha durata complessiva massima di due ore e si svolge presso il laboratorio informatico. Durante la prova occorre risolvere gli esercizi applicati alla luce degli argomenti teorici sviluppati durante il corso e rispondere ad alcune domande di teoria. Le analisi sono condotte tramite l'ambiente R, Rstudio e RMarkdown. Gli esercizi 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 in cui si descrive il procedimento e si illustrano i risultati. Le domande di teoria permettono di verificare l'apprendimento delle nozioni teoriche impartite durante il corso.

Durante il periodo di emergenza Covid-19 a seconda delle disposizioni di ateneo l'esame si svolgerà in laboratorio informatico oppure in videoconferenza tramite piattaforma Webex.

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. Tutti i files sono resi disponibili 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 pubblicati anche alcuni precedenti testi d'esame.

Durante il periodo di emergenza Covid-19 nella pagina del corso vengono anche pubblicate le videoregistrazioni delle lezioni.

I principali testi di riferimento sono elencati nella bibliografia delle dispense, tra gli altri si segnalano i seguenti. Alcuni tra questi disponibili anche in ebook i seguenti:

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. and 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. (2021). *Dispensa di Modelli Statistici II, parte di teoria e applicazioni con R*. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Rizzo M. L. (2008). *Statistical Computing with R*, Chapman & Hall/CRC, New York.

R Core Team (2021). 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

1° Semestre, Ciclo I, Ottobre-Novembre 2021

Lingua di insegnamento

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