UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Modelli Statistici II

2021-2-F8203B011-F8203B013M

Obiettivi formativi

Il corso introduce alle procedure analitiche ed inferenziali condotte tramite modelli statistici avanzati e simulazioni con l'intento di sviluppare una conoscenza critica delle assunzioni alla base della teoria. Argomenti principali sono il bootstrap ed i modelli di miscugli di distribuzioni.

Conoscenza e comprensione

Questo insegnamento permette allo studente:

- di analizzare i dati con modelli statistici sviluppati sia per variabili risposta categoriali che continue
- di sviluppare i metodi di simulazione;
- di servirsi della semantica di R anche tramite l'ambiente RMarkdown per creare codice e documenti che permettono di riprodurre i risultati delle analisi;
- d'interpretare i risultati delle elaborazioni in modo rigoroso e di fornire una descrizione chiara degli stessi con finalità divulgative.

Capacità di applicare conoscenza e comprensione

Lo studente sarà in grado di:

- Sviluppare l'inferenza statistica tramite tecniche di bootstrap;
- Stimare, selezionare ed interpretare i modelli di miscugli di distribuzioni per popolazioni eterogenee;
- Trattare modelli con variabili latenti;
- Applicare le conoscenze teoriche a dati riguardanti gli ambiti dell'epidemiologia, della medicina, della biologia, della 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 sia a livello scritto che nell'esposizione orale.

L'insegnamento fornisce i concetti principali per lo sviluppo di metodi statistici parametrici e non parametrici essenziali nell'ambito teorico e dell'analisi dei dati 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 del modello lineare generalizzato. 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 riguarda i metodi di simulazione e concerne anche i metodi lineari congruenziali per la generazione di numeri pseudo-casuali, i test grafici e statistici (test Kolmogorov-Smirnov e test Chi-Quadrato) per la verifica della pseudo-casualità. Vengono esaminati alcuni metodi per la generazione di realizzazioni da variabili casuali: metodo della trasformata inversa, metodo di accettazione/rifiuto. La teoria è affiancata da esempi applicativi utilizzando diversi modelli distributivi tra cui 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 Vengono illustrati alcuni metodi di ottimizzazione tra cui l'algoritmo Expectation-Maximization. Applicazione del metodo per l'imputazione dei valori mancanti in una tabella a doppia entrata tramite un modello lineare generalizzato. Si introducono i modelli miscuglio per variabili quantitative e categoriali in particolare con componenti assunti con distribuzione di Gauss. 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.

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

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

Durante il periodo di emergenza Covid-19 le lezioni si svolgeranno in modalità da remoto (lezioni videoregistate) con incontri periodici (ogni 2 settimane) in videoconferenza tramite piattaforma webex secondo le calendarizzazioni previste che verranno rese note nella pagina del corso nella sezione ORARI.

Modalità di verifica dell'apprendimento

L’esame è in forma scritta con orale obbligatorio, non sono previste prove intermedie. Le seguenti modalità di verifica dell’apprendimento riguardano sia gli studenti che non frequentanti. L'esame scritto ha durata complessiva di un’ora e trenta minuti e si svolge presso il laboratorio informatico. Durante la prova occorre risolvere gli esercizi applicati alla luce degli argomenti teorici sviluppati durante il corso. 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.

Con esito positivo (ovvero con votazione di almeno 18/30) lo studente accede alla prova orale in cui discute la prova scritta in riferimento agli aspetti teorici trattati nel corso. Entrambe le prove devono essere sostenute nella stessa sessione d’esame. La prova orale permette di verificare la comprensione della teoria e la capacità argomentativa dello studente nonché l’apprendimento delle nozioni teoriche impartite durante il corso.

Durante il periodo di emergenza Covid-19 la modalità di esame sarà la stessa e a seconda delle disposizioni di ateneo si svolgerà in laboratorio informatico oppure in videoconferenza tramite piattaforma webex.

Testi di riferimento

Il materiale didattico è composto principalmente dalle dispense redatte dal docente riguardanti sia la parte teorica

che le applicazioni. Questo è reso disponibile nella pagina della piattaforma e-learning dell'ateneo dedicata al corso. Il docente pubblica al termine di ogni lezione anche le slides, i programmi di calcolo, i dati, gli esercizi e le soluzioni. Nella stessa pagina sono pubblicati alcuni testi d'esame.

I principali testi di riferimento sono elencati nella bibliografia delle dispense, tra gli altri si segnalano i seguenti:

Il materiale didattico è costituito principalmente dalle dispense redatte dal docente riguardanti sia la parte teorica che le applicazioni. Tutto il materiale è disponibile nella pagina della piattaforma e-learning dell'ateneo dedicata al corso. Il docente pubblica al termine di ogni lezione anche le slides, i programmi di calcolo, gli esercizi, i dati, e le soluzioni di ogni lezione. Nella stessa pagina sono pubblicati alcuni 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. Alcuni tra questi anche disponibili 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. (2020). *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 (2020). 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 2020

Lingua di insegnamento

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