UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Bayesian Inference

2122-2-F8203B042-F8203B042M

Obiettivi formativi

Il corso fornisce le conoscenze dei principi di base per l'inferenza statistica in ambito Bayesiano. Il ragionamento Bayesiano viene presentato in modo integrato con l'approccio tradizionale dell'inferenza statistica.

Conoscenza e comprensione

Questo insegnamento permette allo studente di apprendere:

- la regola di Bayes e l'utilizzo della probabilità per aggiornare l'informazione fornita dai dati osservati;
- gli elementi fondamentali dell'inferenza Bayesiana: il calcolo delle distribuzioni a priori, della verosimiglianza e della distribuzione a posteriori;
- il metodo Monte Carlo per la simulazione della distribuzione a posteriori;
- il calcolo della distribuzione predittiva;
- gli algoritmi Markov Chain Monte Carlo: Metropolis-Hastings e Gibbs sampler;
- il modello di regressione lineare multipla ed il modello di regressione logistica multipla in termini Bayesiani;
- i modelli di Markov per l'analisi dei dati longitudinali.

Capacità di applicare conoscenza e comprensione

Questo insegnamento permette allo studente di:

- applicare i metodi di inferenza Bayesiana a casi di studio rilevanti nell’ambito della biostatistica, dell'epidemiologia, della medicina, della biologia, dell’ambiente, della genetica e della salute pubblica;

- applicare i modelli statistici utilizzando dati ripetuti nel tempo per le stesse unità;

- applicare metodi di classificazione basati su modelli statistici;

- sviluppare del codice in ambiente R e SAS;

- Creare report riproducibili come strumento di presentazione dei risultati delle analisi.

Le lezioni teoriche sono affiancate da esercitazioni pratiche su dati reali e simulati in cui si utilizza sia l’ambiente R, Rstudio e Rmarkdown che il software SAS. In tal modo lo studente impara ad utilizzare due diversi linguaggi di programmazione.

L'insegnamento fornisce i concetti principali dell'inferenza Bayesiana, un metodo statistico essenziale nell'ambito teorico e dell'analisi dei dati per i contesti lavorativi di sbocco (biostatistico/statistico/demografico e affini) degli studenti del corso di laurea in Biostatistica. Il corso risulta indispensabile per il successivo percorso universitario.

Contenuti sintetici

Introduzione all’inferenza Bayesiana e alla regola di Bayes. Metodi di specificazione del modello e delle distribuzioni a priori.

Determinazione della distribuzione a posteriori con metodi esatti, famiglie coniugate: Gaussiana, Poisson-Gamma, Beta-Binomiale, Multinomiale-Dirichelet.

Inferenza Bayesiana non parametrica.

Metodi di sintesi della distribuzione a posteriori, intervalli di credibilità e intervalli con la massima densità a posteriori.

Introduzione ai processi stocastici di Markov e proprietà. Modello passeggiata casuale.

Modello di transizione per dati longitudinali.

Modello di Markov a variabili latenti per dati longitudinali con covariate.

Metodi Markov Chain Monte Carlo: Algorithmo Metropolis-Hastings e campionamento Gibbs.

Ambiente R e Rstudio, utilizzando principalmente le seguenti librerie: probBayes, learnBayes, LMest. RMarkdown attraverso la libreria knitr per integrare codice e output. Software SAS: proc MCMC.

Programma esteso

Il corso comprende un’introduzione all’inferenza Bayesiana e il confronto con l’inferenza classica. Viene ripresa la regola di Bayes e la regola delle probabilità totali attraverso l’esempio del Bayes'billiard. Vengono sviluppati gli

aspetti di specificazione delle distribuzioni a priori, la stima esatta delle distribuzioni a posteriori e l'interpretazione dei modelli Bayesiani. Viene introdotto il modello Beta-Binomiale ed illustrato anche l'approccio Bayesiano non parametrico. Enfasi viene posta anche sulla distribuzione predittiva. Vengono illustrate le caratteristiche di scelta e di determinazione delle distribuzioni a priori: esempi e convenienza della famiglia coniugata. Si considera anche la scelta delle distribuzioni a priori non informative e la nozione di scambiabilità è illustrata attraverso il teorema di rappresentazione di De Finetti.

Vengono trattati i metodi di sintesi della distribuzione a posteriori: intervalli di credibilità, intervalli con la massima densità a posteriori. Famiglie coniugate: Beta-Binomiale e Gaussiana, modello Poisson-Gamma. Introduzione alla distribuzione multinomiale e di Dirichlet. La teoria viene affiancata da svariati esempi di applicazione dei modelli Bayesiani nell'ambito della biostatistica attraverso dati reali e simulati riguardanti l'epidemiologia, la farmaco epidemiologia, la medicina e la biologia oltre che l'ecologia e le scienze ambientali.

Vengono introdotti i processi stocastici Markoviani le proprietà e le caratteristiche delle catene di Markov. Si mostra il processo passeggiata casuale anche attraverso la simulazione delle sue traiettorie. Viene introdotto il modello di transizione per dati longitudinali ed il modello latente di Markov. Vengono illustrati gli algoritmi di stima maggiormente utilizzati nell'ambito del metodo Markov Chain Monte Carlo (MCMC): l'algoritmo Metropolis-Hastings e l'algoritmo Gibbs sampling. Vengono discusse diverse misure che permettono la valutazione diagnostica della loro convergenza.

La teoria è affiancata da numerose applicazioni a dati reali e simulati riguardanti gli ambiti della biostatistica in modo da facilitare anche lo sviluppo della conoscenza della semantica in ambiente R e del software SAS. Gli esempi sono svolti in Rstudio con l'ausilio di RMarkdown. Lo studente durante le esercitazioni è incoraggiato anche tramite l'apprendimento cooperativo, ad elaborare documenti riproducibili e a sviluppare il commento ai risultati delle analisi in modo critico. Nelle ultime settimane viene spiegato l'utilizzo delle procedure SAS per la stima Bayesiana dei modelli di regressione lineare a logistica.

Prerequisiti

Si consiglia di riprendere le nozioni impartite nei seguenti corsi: Statistica, Probabilità e Inferenza Statistica, Modelli Statistici II.

Metodi didattici

Sono previste lezioni frontali riguardanti la parte teorica sui concetti di base dell'inferenza Bayesiana e dei modelli di Markov per dati longitudinali. Le lezioni di teoria sono affiancate da esercitazioni pratiche che permettono allo studente di sviluppare l'aspetto della scienza dei dati. Le lezioni si tengono presso il laboratorio informatico. Settimanalmente vengono assegnati esercizi di riepilogo da svolgere con dati reali o simulati relativi alla parte di programma svolto. Durante il corso con l'ausilio di R nell'ambiente RStudio e del marcatore di testo RMarkdown oppure del software SAS, gli studenti imparano ad elaborare documenti riproducibili per le analisi dei dati e la stima dei modelli proposti. 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

L’esame è in forma scritta con orale facoltativo, non sono previste prove intermedie. Le seguenti modalità di verifica dell’apprendimento riguardano sia gli studenti frequentanti che non frequentanti. 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 e SAS. Gli esercizi permettono di verificare la capacità di comprensione del problema, la sua risoluzione tramite l'applicazione dei modelli Bayesiani e di modelli per dati longitudinali 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 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. 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 anche i file riferiti al materiale didattico che comprende: 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 dovuto alla pandemia 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 e-book i seguenti:

Albert, J. (2009). *Bayesian computation with R*. Springer Science & Business Media.

Albet, J., Hu, J. (2019). *Probability and Bayesian modeling*. Chapman and Hall/CRC.

Bartolucci, F., Farcomeni, A., Pennoni, F. (2013). *Latent Markov Models for longitudinal data*, Chapman and Hall/CRC, Boca Raton.

Migon, H. S., Gamerman, D., Louzada, F. (2014). *Statistical inference: an integrated approach*. Chapman & Hall.

Pennoni, F. (2021). *Dispensa di Inferenza Bayesiana -parte di teoria e applicazioni con R e SAS*. Dipartimento di Statistica e Metodi Quantitativi, Università degli Studi di Milano-Bicocca.

Robert, C., Casella, G. (2004). *Monte Carlo Statistical Methods* (Second edition). Springer–Verlag, New York.

Dipak, D. K., Ghosh, S. K., Mallick, B. K. (2000). *Generalized linear models: A Bayesian perspective*.CRC press.

SAS/STAT PROC MCMC, *User’s guide*, SAS Institute, 2012.

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 II, Novembre 2020- Gennaio 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