UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Bayesian Inference

2223-2-F8203B042-F8203B042M

Obiettivi formativi

Il corso fornisce la conoscenza dei principi di base per l'inferenza statistica in ambito Bayesiano. Il ragionamento Bayesiano viene presentato in modo integrato con l'approccio all'inferenza statistica sviluppato in senso classico.

Conoscenza e comprensione

Questo insegnamento permette allo studente di apprendere:

- la regola di Bayes e l'utilizzo delle regole probabilistiche per aggiornare l'informazione fornita dai dati osservati;
- gli elementi fondamentali dell'inferenza Bayesiana quali il calcolo delle distribuzioni a priori, della funzione di 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 e logistica multipla Bayesiano;
- i modelli di Markov per dati longitudinali.

Capacità di applicare conoscenza e comprensione

Questo insegnamento permette allo studente:

- applicare i metodi di inferenza Bayesiana a casi di studio rilevanti nei seguenti ambiti: biostatistica, epidemiologia, medicina, biologia, ambiente, genetica e 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 basate su dati reali e simulati in cui si utilizzano

sia l’ambiente R, Rstudio e Rmarkdown che il software SAS.
L’inferenza Bayesiana si caratterizza quale metodo statistico imprescindibile 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.
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 delle catene di Markov. Modello passeggiata casuale.
Modello di transizione per dati longitudinali.
Modello di Markov a variabili latenti per dati longitudinali e estensioni del modello con covariate nella parte manifesta e nella parte latente.
Metodi Markov Chain Monte Carlo: Algoritmo Metropolis-Hastings e campionamento Gibbs.
Ambiente R e Rstudio, utilizzando principalmente le seguenti librerie: probBayes, learnBayes, LMest, LaplaceDemon, RMarkdown attraverso la libreria knitr per integrare il codice, i risultati delle analisi ed i commenti.
Software SAS: proc MCMC.

Programma esteso

Durante il corso viene ripresa la regola di Bayes e la regola delle probabilità totali attraverso l’esempio del Bayes'billard. Vengono sviluppati gli aspetti riguardanti la specificazione delle distribuzioni a priori, la stima esatta delle distribuzioni a posteriori e l’interpretazione dei modelli. Viene introdotto il modello beta-binomiale e le altre famiglie coniugate: Gaussiana, modello Poisson-gamma, multinomiale-Dirichlet e enfasi viene posta anche sulla distribuzione predittiva. L’inferenza viene confrontata con quella derivante dall’approccio classico. Vengono illustrate le caratteristiche di scelta e di determinazione delle distribuzioni a priori considerando distribuzioni a priori informative e non informative. La nozione di scambiabilità viene 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.
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 enunciando le proprietà e le caratteristiche delle catene di Markov. Si mostrano le caratteristiche del processo passeggiata casuale, attraverso le simulazioni delle traiettorie per matrici stocastiche con diverse dimensioni. Viene introdotto il modello di transizione per dati longitudinali, ed il modello latente di Markov. Vengono illustrati anche da un punto di vista computazionale gli algoritmi di stima 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 applicativi del corso di laurea 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 concernenti anche il commento critico ai risultati delle analisi.

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 di teoria e queste sono affiancate da esercitazioni pratiche che permettono allo studente di sviluppare l'aspetto della scienza dei dati. Le lezioni sono impartire presso il laboratorio informatico. 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 analizzare i dati e stimare i modelli Bayesiani elaborando documenti riproducibili. Settimanalmente vengono assegnati esercizi di riepilogo da svolgere con dati reali o simulati relativi alla parte di programma svolto dove gli studenti 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 e SAS e 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

Le dispense di teoria e di applicazioni redatte dal docente costituiscono il materiale didattico principale che è reso disponibile nella pagina della piattaforma e-learning dell'ateneo dedicata al corso al termine di ogni lezione. Il materiale didattico comprende: le slides, i programmi di calcolo, gli esercizi, alcuni testi testi d'esame riferiti agli appelli precedenti.

I principali testi di riferimento sono elencati nella bibliografia delle dispense. Alcuni tra questi anche disponibili in e-book sono i seguenti:

Albert, J. (2009). Bayesian computation with R. Springer Science & Business Media.
Albert, 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. (2022). Dispensa di Inferenza Bayesiana: 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 (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

1° semestre, Ciclo II, Novembre 2022-Gennaio 2023

Lingua di insegnamento

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

Sustainable Development Goals

SALUTE E BENESSERE