UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Bayesian Inference

2425-2-F8203B042-F8203B042M

Obiettivi formativi

Il corso rientra nelle aree di apprendimento delle scienze statistiche, dell'informatica e delle scienze sociali. Il corso permette allo studente di apprendere le procedure analitiche ed inferenziali in ambito Bayesiano. Il ragionamento Bayesiano viene presentato in modo integrato con l'approccio all'inferenza statistica sviluppato in senso classico. Il corso permette agli studenti di acquisire solidi elementi di teoria Bayesiana e di sviluppare le applicazioni attraverso un approccio di "problem solving" con dati reali e simulati in riferimento a problemi applicativi inerenti alla biostatistica.

Conoscenza e comprensione

Lo studente viene introdotto alla conoscenza dei principali modelli statistici Bayesiani per l'analisi di dati con diverse tipologie di variabili risposta. Si illustrano anche le relative ipotesi alla base della teoria considerando i modelli coniugati e gli algoritmi di stima per ottenere tramite simulazioni la distribuzione a posteriori. Al fine di introdurre la conoscenza delle procedure Markov Chain Monte Carlo vengono presentati i concetti essenziali dei processi stocastici Markoviani. Lo studente impara anche a trattare i dati longitudinali e a stimare alcuni modelli statistici che tengono conto delle misure ripetute. L'analisi dei dati viene condotta sia utilizzando il software R nell'ambiente RMarkdown che permette di creare documenti riproducibili contenenti il codice, i risultati ed i commenti, sia utilizzando specifiche procedure per l'analisi Bayesiana con il software SAS. Gli esempi applicativi riguardano dati reali e simulati provenienti da diversi ambiti di riferimento per il corso di studi. Lo studente è incoraggiato a fornire anche una valutazione critica circa i risultati ottenuti con le analisi empiriche.

Capacità di applicare conoscenza e comprensione

Il corso fornisce competenze nell'utilizzo dei modelli Bayesiani per applicazioni a casi di studio rilevanti nei seguenti ambiti: biostatistica, epidemiologia, medicina, biologia, ambiente, genetica e salute pubblica. Attraverso R ed RStudio gli studenti imparano ad impostare in modo organico il ragionamento statistico attraverso l'analisi dei dati e la redazione di relazioni che illustrino il codice, le analisi ed i risultati. Attraverso l'utilizzo del software SAS gli studenti imparano a stimare tramite simulazioni i modelli Bayesiani e ad impostare gli input richiesti dagli algoritmi di stima. La teoria viene affiancata da applicazioni pratiche. Il corso consente agli studenti di acquisire solide basi

teoriche e capacità di applicare i modelli statistici proposti a dati reali. Lo studente è incoraggiato a fornire anche una valutazione critica circa i risultati ottenuti con le analisi empiriche.

L'insegnamento è indispensabile per il successivo percorso universitario in quanto fornisce i concetti essenziali per lo sviluppo dei metodi Bayesiani sia in ambito teorico che applicativo, oltre che per i contesti lavorativi di sbocco (biostatistico/statistico/demografico e affini) degli studenti del corso di laurea in Biostatistica.

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-Dirichlet .

Inferenza Bayesiana non parametrica.

Metodi di sintesi della distribuzione a posteriori, regioni 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 ed estensioni del modello con covariate sia nel modello osservato che nel modello latente.

Metodi Markov Chain Monte Carlo: algoritmi Metropolis-Hastings e Gibbs sampling.

Test diagnostici per la convergenza.

Esercitazioni svolte in relazione a specifici problemi applicativi utilizzando l'ambiente R, RStudio ed il software SAS.

Programma esteso

All'inizio del corso vengono riprese 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 altre famiglie coniugate: Gaussiana, Poisson-gamma, multinomiale-Dirichlet. Enfasi viene posta anche sulla distribuzione predittiva. L'inferenza Bayesiana viene confrontata con l'inferenza intesa in senso classico. Vengono illustrate le caratteristiche di scelta e di determinazione della distribuzione a priori sia informativa che non informativa. La nozione di scambiabilità viene illustrata attraverso il teorema di rappresentazione di De Finetti. La distribuzione a posteriori viene sintetizzata attraverso le regioni di credibilità, e gli 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 farmacoepidemiologia, 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. La passeggiata casuale viene illustrata 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 riferite sia alle analisi grafiche che ai test statistici che permettono la valutazione diagnostica della 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. Vengono utilizzati sia l'ambiente R e Rstudio, ed i seguenti principali pacchetti: probBayes, learnBayes, LMest, LaplaceDemon, RMarkdown attraverso la libreria knitr per integrare il codice, i risultati delle analisi ed i commenti, oltre al software SAS attraverso la libreria proc MCMC.

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 affiancate da esercitazioni pratiche che permettono allo studente di sviluppare l'aspetto della scienza dei dati. Le lezioni sono impartite in presenza. Durante il corso con l'ausilio di R nell'ambiente RStudio, con il 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 dove gli studenti vengono incoraggiati ad affrontare il problema applicativo riferito all'ambito teorico illustrato a lezione con lo scopo ulteriore di sviluppare l'apprendimento cooperativo. Le ore previste di didattica erogativa sono 30 e quelle di didattica interattiva sono 24.

Modalità di verifica dell'apprendimento

Le seguenti modalità di verifica dell'apprendimento si applicano sia agli studenti frequentanti che a quelli non frequentanti le lezioni frontali. L'esame è in forma scritta con orale facoltativo, non sono previste prove intermedie. L'esame scritto ha una durata di circa un ora e mezza e si svolge in laboratorio informatico. Durante l'esame, gli studenti devono rispondere a domande aperte di teoria e risolvere gli esercizi applicativi basandosi sugli argomenti teorici trattati e sulle esercitazioni pratiche assegnate settimanalmente durante il corso. Le domande di teoria valutano l'apprendimento dei concetti essenziali dell'inferenza statistica con metodi avanzati. Le analisi empiriche devono essere condotte utilizzando l'ambiente R, RStudio e RMarkdown ed il software SAS e permettono di verificare la capacità degli studenti di applicare modelli statistici Bayesiani a dati reali o simulati e di elaborare report riproducibili che descrivano i dati, le procedure e i risultati ottenuti. Durante l'esame è consentito l'utilizzo del materiale di studio e del materiale fornito dal docente comprendente anche i codici R e SAS implementati durante il corso. Ogni domanda avrà un punteggio di circa 2 o 3 punti. Lo studente supera l'esame con una votazione di almeno 18/30.

Testi di riferimento

Il materiale didattico principale consiste nelle dispense preparate dal docente, che coprono, gli argomenti teorici, le applicazioni sviluppate con il software R, gli esercizi e le soluzioni. Queste dispense saranno rese disponibili sulla pagina della piattaforma e-learning dell'università dedicata al corso. Inoltre, il docente pubblica alla fine di ogni lezione le slides, i programmi di calcolo e i dataset utilizzati. Settimanalmente vengono assegnati esercizi, e le relative soluzioni. Sulla stessa pagina web sono disponibili degli esempi del testo d'esame.
I riferimenti bibliografici principali sono elencati nella bibliografia delle dispense alcuni dei quali sono i seguenti che risultano disponibili presso la biblioteca di Ateneo anche in formato ebook:

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. (2024). 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 (2023). 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 2024 - Gennaio 2025

Lingua di insegnamento

Il corso viene erogato in lingua italiana. Gli studenti Erasmus possono utilizzare il materiale didattico predisposto in lingua inglese e fornito dal docente su richiesta. Possono inoltre richiedere di svolgere la prova d’esame in lingua inglese.

Sustainable Development Goals

SALUTE E BENESSERE