UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Bayesian Inference

2526-2-F8203B042-F8203B042M

Obiettivi formativi

L'insegnamento rientra nelle aree di apprendimento delle scienze statistiche, dell'informatica e delle scienze sociali e 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. L'attività formativa permette agli studenti di acquisire solidi elementi di teoria Bayesiana, di sviluppare le applicazioni attraverso un approccio di "problem solving" con dati reali e simulati in riferimento a problemi applicativi inerenti alla biostatistica. Lo studente acquisisce anche abilità comunicative testuali scritte in quanto è richiesta la stesura di testi a completamento dei risultati ottenuti dalle analisi svolte.

Conoscenza e comprensione

Lo studente viene introdotto ai principali modelli statistici Bayesiani per l'analisi di dati con diverse tipologie di variabili risposta, alle ipotesi sottostanti tali modelli e ai modelli per l'analisi dei dati longitudinali. Acquisisce inoltre la comprensione delle procedure Markov Chain Monte Carlo e dei relativi algoritmi di stima, nonché la capacità di valutarne l'efficacia.
Lo studente viene anche introdotto al linguaggio di programmazione R, nell'ambiente RMarkdown, che consente di creare documenti riproducibili contenenti codice, risultati, commenti e specifiche procedure per l'analisi Bayesiana tramite il software SAS.
Gli esempi applicativi riguardano dati reali e simulati provenienti da diversi ambiti di riferimento del corso di studi. Lo studente apprende inoltre a fornire una descrizione scritta dei risultati ottenuti alla luce delle domande di ricerca, insieme a una valutazione critica dei limiti delle analisi svolte. In questo modo sviluppa autonomia di giudizio e affina le proprie abilità comunicative, sia nella presentazione sia nella giustificazione dei procedimenti adottati per le analisi, sia nell'illustrazione dei risultati ottenuti in relazione alle domande di ricerca.

Capacità di applicare conoscenza e comprensione

L'insegnamento fornisce competenze nell'utilizzo dei modelli bayesiani coniugati, nella scelta delle distribuzioni a priori e nell'impiego degli algoritmi di stima per modelli complessi.
Attraverso l'ambiente R e RStudio, gli studenti imparano a impostare in modo organico il ragionamento statistico,

mediante l’analisi dei dati e la redazione di relazioni che illustrino il codice, le analisi e i risultati.
Con l’uso del software SAS, gli studenti apprendono a stimare modelli bayesiani complessi tramite simulazioni e a impostare correttamente gli input richiesti dagli algoritmi di stima.
L’insegnamento consente agli studenti di acquisire solide basi teoriche e la capacità di applicare i modelli statistici presentati a dati reali. Gli studenti sono in grado di valutare il modello più appropriato in base ai dati disponibili e alle domande di ricerca. Imparano inoltre a scrivere e commentare il codice utilizzato per generare i risultati, adottando un approccio open source che garantisca la riproducibilità e la replicabilità delle analisi.
Al termine dell’insegnamento, grazie al materiale fornito (le dispense del docente corredate da un’ampia bibliografia, i codici per i software R e SAS e l’interfaccia RMarkdown), lo studente è in grado di proseguire in modo autonomo nell’approfondimento di questa disciplina.

L’insegnamento è indispensabile per il successivo percorso universitario in quanto fornisce i concetti essenziali per lo sviluppo dei metodi Bayesiani sia attraverso solide basi teoriche che applicative 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-Dirichelet .

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: Algoritmo Metropolis-Hastings e campionamento Gibbs.

Test diagnostici per la convergenza.

Esercitazioni svolte in relazione a specifici problemi applicativi utilizzando l’ambiente R, RStudio con il maratore di testo RMarkdown ed il software SAS.

Programma esteso

Durante l’insegnamento 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 altre famiglie coniugate: Gaussiana, modello Poisson-gamma, multinomiale-Dirichlet e enfasi viene posta anche sulla distribuzione predittiva. L’inferenza puntuale Bayesiana viene confrontata 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.

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) per approssimare la distribuzione a posteriori: 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 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. Si intende inoltre facilitare 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, e seguenti principali pacchetti: probBayes, learnBayes, LMest, LaplaceDemon, RMarkdown attraverso la libreria knitr per integrare il codice, i risultati delle analisi ed i commenti. L'analisi con il software SAS viene svolta attraverso la libreria proc MCMC.

Prerequisiti

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

Metodi didattici

Sono previste lezioni frontali svolte in modalità erogativa in presenza riguardano la parte di teoria e sono affiancate da esercitazioni pratiche. Le lezioni sono impartite presso il laboratorio informatico. Durante l'attività formativa con l'ausilio di R nell'ambiente RStudio, con il marcatore di testo RMarkdown oppure del software SAS, gli studenti imparano ad analizzare i datil, a stimare i modelli Bayesiani e a corredare le analisi con i commenti, elaborando documenti che permettono di replicare le analisi e riprodurre i risultati. 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 17. Le lezioni da 3 ore in particolare sono svolte nella seconda parte in modo da coinvolgere gli studenti in modo interattivo. Le esercitazioni sono svolte in modalità interattiva in presenza presso il laboratorio informatico. Vengono rese disponibili nella pagina di e-elearning le video-registrazioni in asincrono sia delle lezioni che delle esercitazioni.

Modalità di verifica dell'apprendimento

Le seguenti modalità di verifica dell'apprendimento si applicano sia agli studenti frequentanti sia a quelli non frequentanti le lezioni frontali. L'esame è in forma scritta, con domande aperte, e include un colloquio orale facoltativo. Non sono previste prove intermedie.

La prova scritta ha una durata massima di due ore e si svolge in laboratorio informatico. Durante l'esame, gli studenti devono rispondere a domande aperte di teoria e risolvere esercizi applicativi, basandosi sugli argomenti teorici trattati e sulle esercitazioni pratiche assegnate settimanalmente.
Le domande di teoria mirano a valutare la comprensione dei concetti fondamentali dell'inferenza statistica bayesiana con metodi avanzati. Le analisi empiriche, condotte utilizzando l'ambiente R, RStudio, RMarkdown e il software SAS, consentono di verificare la capacità degli studenti di applicare modelli statistici bayesiani a dati reali o simulati, nonché di redigere report riproducibili che descrivano i dati, le procedure e i risultati ottenuti.
La prova è finalizzata a valutare la capacità dello studente di affrontare le problematiche oggetto di studio, di utilizzare il codice fornito e di comunicare in modo chiaro i risultati raggiunti, oltre a saper gestire efficacemente il tempo a disposizione.
Durante l'esame è consentito l'uso del materiale di studio e di quello fornito dal docente, compresi i codici R e SAS implementati durante lo svolgimento dell'attività didattica. Ogni punto di ogni esercizio ha una valutazione di circa 3 punti. L'esame si considera superato con una votazione minima di 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 all'insegnamento. 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 di alcuni testi d'esame.

I riferimenti bibliografici principali sono elencati nella bibliografia delle dispense. I principali sono elencati nel seguito e 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.

Periodo di erogazione dell'insegnamento

1° semestre, Ciclo II, Novembre 2025 - Gennaio 2026

Lingua di insegnamento

L'insegnamento 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 | ISTRUZIONE DI QUALITÁ