UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

High Dimensional Data Analysis

2627-2-FDS02Q022

Obiettivi formativi

Questo è un corso avanzato di statistica incentrato sull'analisi di dati ad alta dimensione, in cui il numero di variabili può essere elevato rispetto al numero di osservazioni (o persino superarlo). Il corso mira a introdurre i moderni metodi statistici per la riduzione della dimensionalità, la regressione regolarizzata e la selezione delle variabili in contesti ad alta dimensione. L'accento è posto sia sui fondamenti teorici sia sugli aspetti computazionali dei metodi. In linea con gli obiettivi formativi del Corso di Studio, l'insegnamento si colloca nell'area di apprendimento Statistico-Matematica. Esso contribuisce a sviluppare le competenze analitiche e metodologiche avanzate, essenziali per definire il profilo quantitativo del laureato e per garantire la coerenza dei piani di studio individuali. Durante tutto il corso, gli studenti saranno incoraggiati a sviluppare capacità di apprendimento autonomo e l'abilità di valutare criticamente le nuove tecniche, in linea con gli obiettivi del Corso di Laurea Magistrale in Data Science.

Obiettivi formativi secondo i Descrittori di Dublino (DdD)
1. Conoscenza e capacità di comprensione

- Comprendere le sfide e i fondamenti teorici dell'analisi statistica in contesti ad alta dimensionalità.
- Apprendere tecniche chiave come la regressione penalizzata (es. Lasso, Ridge), la selezione di sottoinsiemi, la riduzione dimensionale e i criteri di selezione del modello.
- Familiarizzare con concetti quali la sparsità, il compromesso bias-varianza, l'overfitting e la "maledizione della dimensionalità".

2. Conoscenza e capacità di comprensione applicate

- Applicare tecniche di modellizzazione ad alta dimensionalità a problemi concreti in ambiti come la genomica o l'analisi di immagini.
- Utilizzare il software statistico R per implementare e valutare modelli ad alta dimensionalità.
- Interpretare i risultati dei modelli e valutarne le prestazioni attraverso studi di simulazione e tecniche di validazione incrociata.

3. Autonomia di giudizio

- Sviluppare la capacità di scegliere metodi statistici adeguati in scenari complessi e ad alta dimensionalità.
- Valutare criticamente i limiti e le assunzioni dei modelli, riconoscendo possibili fonti di bias o instabilità.
- Confrontare modelli alternativi utilizzando evidenze empiriche e criteri teorici.

4. Abilità comunicative

- Comunicare concetti e risultati statistici complessi in modo chiaro e rigoroso, sia in forma scritta sia durante le discussioni in aula.
- Presentare e discutere i risultati analitici, le assunzioni e i limiti delle tecniche di analisi ad alta dimensionalità in modo preciso e strutturato.
- Sviluppare la capacità di giustificare le scelte metodologiche e interpretare i risultati alla luce delle considerazioni teoriche ed empiriche.

5. Capacità di apprendimento

- Sviluppare la capacità di aggiornarsi autonomamente in un ambito in rapida evoluzione come l'analisi statistica e l'apprendimento in alta dimensionalità.
- Consultare e valutare criticamente la letteratura scientifica più recente, approfondendo nuove metodologie o applicazioni.

Contenuti sintetici

Il corso riguarda metodi di regressione che possono essere impiegati nel caso di dati ad alta dimensionalità

Programma esteso

1. Regressione lineare, bias/variance trade-off
2. Regressione penalizzata, ridge regression e lasso
3. Sezione del modello, metodi di validazione incrociata
4. Regressione nonparametrica. k-*nearest neighbors* (k-NN). Kernel smoothing. Regression splines, Smoothing splines, Local regression
5. High-dimensional inference
6. Selective Inference

Prerequisiti

Il corso richiede conoscenze pregresse di probabilità, inferenza statistica, algebra lineare e programmazione. Particolare attenzione è richiesta alla conoscenza del modello di regressione lineare, sia negli aspetti descrittivi che inferenziali, e alla teoria asintotica basata sull'approssimazione normale.

Metodi didattici

Il corso si svolge interamente in un'aula informatizzata, integrando l'insegnamento teorico con attività

computazionali svolte mediante il software R.
Le lezioni saranno tenute in lingua inglese e si articoleranno in:

- lezioni frontali, finalizzate all'introduzione dei concetti teorici e metodologici;
- sessioni tutoriali, dedicate all'illustrazione pratica degli strumenti open-source e alla loro applicazione a problemi reali di analisi statistica ad alta dimensionalità.

Il corso prevede 42 ore complessive, erogate in presenza, organizzate in incontri della durata di 2 o 3 ore ciascuno.

Modalità di verifica dell'apprendimento

L'apprendimento sarà verificato tramite una prova scritta individuale, composta da due domande aperte, ciascuna articolata in sottosezioni che coprono aspetti teorici e applicativi dei contenuti trattati a lezione. La valutazione tiene conto della correttezza, completezza, chiarezza espositiva e proprietà di linguaggio delle risposte.

Non sono previste prove in itinere.

Il punteggio massimo complessivo è pari a 31/30, con eventuale lode assegnata in presenza di una prova particolarmente eccellente. La valutazione non segue una rigida griglia predefinita per ciascuna domanda, ma si basa su criteri di qualità complessiva delle risposte e padronanza degli argomenti, pur garantendo trasparenza e coerenza nella correzione.

Testi di riferimento

- Materiale didattico fornito dal docente
- Azzalini, Scarpa (2012) Data analysis and data mining, an introduction . New York: Oxford University Press
- Gareth, Witten, Hastie, Tibshirani (2014) An Introduction to Statistical Learning, with Applications in R . Springer
- Hastie, Tibshirani, Friedman (2009) The Elements of Statistical Learning. Data Mining, Inference and Prediction . Springer
- Hastie, Tibshirani and Wainwright (2015) Statistical Learning with Sparsity: The Lasso and Generalizations . CRC Press

Periodo di erogazione dell'insegnamento

Primo Semestre

Lingua di insegnamento

Inglese

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ