UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Mining M

2425-1-F8204B014

Obiettivi formativi

Il corso si pone come obiettivo l'approfondimento di tecniche per l'analisi dei dati e di *data mining* e il perfezionamento delle abilità di modellizzazione con finalità previsiva, con relative implementazioni nell'ambiente di programmazione R.

Il corso contribuisce al raggiungimento degli obiettivi formativi nell'area di apprendimento del CdS: "**Statistica**".

Contenuti sintetici

Il programma dettagliato è disponibile nella pagina web del corso. Gli argomenti principali sono:

- A-B-C: modelli lineari ed aspetti computazionali
- Compromesso distorsione e varianza, ottimismo
- Selezione del modello e metodi penalizzati per modelli lineari (regressione ridge, lasso, *elastic-net*)
- Regressione nonparametrica (regressione lineare locale, splines di regressione e di lisciamento)
- Modelli additivi (GAM and MARS)

Programma esteso

- **A-B-C**
  - Il modello lineare e: ripasso e notazione
  - Equazioni normali, scomposizione di Cholesky ed algoritmi efficienti per i minimi quadrati
  - Scomposizione QR, metodo delle ortogonalizzazioni successive
  - Minimi quadrati iterati

    - Modelli lineari generalizzati: ripasso e notazione
- **Compromesso distorsione e varianza, ottimismo**
    - Regressione polinomiale
    - Insieme di stima ed insieme di verifica
    - Ottimismo, compromesso distorsione varianza, indice di Mallows
    - Convalida incrociata e convalida incrociata generalizzata
    - Criteri di informazione (AIC, BIC, etc.)
- **Selezione del modello e metodi penalizzati per modelli lineari**
    - *Best subset selection*
    - Regressione tramite componenti principali
    - Regressione *ridge*
    - Regressione *LARS* e Lasso
    - *Elastic-net*
- **Regressione nonparametrica**
    - Regressione lineare locale
    - Splines di regressione e di lisciamento
    - Regressione nonparametrica, caso bivariato
    - Maledizione della dimensionalità
- **Modelli additivi**
    - *Generalized Additive Models* (GAM)
    - *Multivariate Adaptive Regression Splines* (MARS)

Prerequisiti

È richiesta la conoscenza di (i) nozioni di algebra lineare, (ii) modelli di regressione lineare, (iii) modelli di regressione lineare generalizzati (GLM), (iv) inferenza statistica, (v) calcolo delle probabilità. È inoltre richiesta una solida conoscenza del software R.

Si raccomanda inoltre la conoscenza degli argomenti avanzati di probabilità e statistica inferenziale trattati nei corsi *Probabilità e Statistica Computazionale M* e *Statistica Avanzata M*.

Metodi didattici

Le lezioni si svolgono sia in aula che in laboratorio, integrando aspetti di carattere teorico con quelli pratico-applicativi di analisi dei dati e di programmazione in R.

Le 47 ore di didattica saranno così suddivise:

- 35 ore di lezione svolte in modalità erogativa in presenza;
- 12 ore di attività di laboratorio svolte in modalità interattiva da remoto.

Modalità di verifica dell'apprendimento

L'esame è composto da due parti, entrambe obbligatorie:

- (20 punti su 30) Prova scritta a **domande aperte**, in cui vengono valutati gli aspetti teorici del corso.

- (10 punti su 30) **Progetto individuale** (*data challenge*).

Il voto finale è dato dalla somma dei punteggi delle due parti.

Nella seconda metà del corso viene annunciata il tema del progetto individuale (*data challenge*). Gli studenti dovranno produrre ed inviare al docente delle **previsioni** relative al caso studio assegnato, congiuntamente ad una **relazione** di 4-5 pagine. Il materiale del progetto deve essere inviato al docente prima dell'esame scritto e ha validità di un anno, a partire dal momento in cui la competizione è stata annunciata.

Testi di riferimento

Riferimenti principali

- Azzalini, A. and Scarpa, B. (2011), *Data Analysis and Data Mining*, Oxford University Press.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, Second Edition, Springer.

Approfondimenti

- Efron, B. and Hastie, T. (2016), *Computer Age Statistical Inference*, Cambridge University Press.
- Lewis, Kane, Arnold (2019) *A Computational Approach to Statistical Learning*. Chapman And Hall/Crc.

Ulteriore materiale didattico verrà messo a disposizione nella pagina web del corso.

Periodo di erogazione dell'insegnamento

Secondo semestre

Lingua di insegnamento

Inglese

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ