



UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Data Science and Statistical Models for Unstructured Data

2223-3-E4102B076

Obiettivi formativi

Introdurre gli studenti alle moderne tecniche statistiche di classificazione non-supervisionata e riduzione della dimensionalità dei dati, con particolare enfasi su:

1. L'unità della struttura concettuale del problema e il legame tra questa e le tecniche statistiche.
2. La struttura logica e matematica sottostante gli algoritmi.
3. Le differenze e le diverse caratteristiche delle tecniche e degli algoritmi, con il relativo riflesso sui criteri di utilizzo nell'analisi dei dati.
4. I limiti intrinseci degli algoritmi statistici

In sintesi, l'obiettivo è fornire agli studenti una conoscenza e una competenza allo stato dell'arte degli strumenti di classificazione non-supervisionata, insieme ad una comprensione profonda della struttura delle tecniche e ad una capacità critica per quanto riguarda il loro utilizzo, in termini di concettualizzazione del problema da affrontare, scelta degli algoritmi, loro implementazione e validazione, analisi e interpretazione dei risultati.

In tal modo, l'insegnamento permette l'acquisizione di solide basi per le applicazioni in campo demografico, socio-economico, biostatistico e, in generale, in tutti gli ambiti in cui devono essere analizzati sistemi complessi e multidimensionali di dati.

Contenuti sintetici

Il corso introduce e motiva il problema della classificazione non-supervisionata e della riduzione della dimensionalità, mostrandone le sue applicazioni ad ambiti concreti, richiama gli strumenti matematici di base, principalmente algebrici, e illustra le principali tecniche e i principali algoritmi di classificazione/riduzione, sia di tipo

lineare che non-lineare e sia per dati numerici che non-numerici e parzialmente ordinati. La spiegazione delle tecniche è supportata da numerosi esempi su dati reali o simulati e dall'illustrazione dei codici sw che ne permettono l'implementazione.

Programma esteso

1. Il problema della classificazione non-supervisionata e della riduzione della dimensionalità dei dati: esempi di analisi di dati sociali, di dati economici e di dati provenienti da discipline umanistiche.
2. Richiami di algebra lineare: spazi vettoriali, prodotti scalari, proiezioni ortogonali, norme matriciali.
3. Tecniche di analisi lineari:
 - Decomposizione a Valori Singolari (SVD) e suo legame con l'analisi delle Componenti Principali.
 - Non-negative Matrix Factorization e suo confronto con SVD.
 - Il lemma di Johnson-Lindestrauss e la riduzione della dimensionalità a distorsione limitata: Proiezioni Casuali
 - Multidimensional Scaling.
4. Tecniche non-lineari:
 - Dati su varietà differenziabili: Isomap
 - Self-organizing map (SOM)
 - Entropia, divergenza di Kullback-Liebler e riduzione della dimensionalità: SNE e t-SNE
5. Dati categoriali e parzialmente ordinati:
 - Analisi delle Corrispondenze
 - Estrazione di ranking da dati multidimensionali
6. I limiti degli algoritmi statistici: il teorema "no-free lunch".

Prerequisiti

Non sono formalmente previsti prerequisiti, ma è necessaria una competenza di base di algebra lineare, statistica descrittiva e analisi dei dati.

Metodi didattici

Lezioni frontali ed esercitazioni e simulazioni, mediante linguaggio R, condotte dal docente. Il metodo didattico prevede anche l'utilizzo di canali social, per agevolare discussioni e confronti fra studenti e con il docente.

Modalità di verifica dell'apprendimento

Esame orale, per verificare, la comprensione delle logiche di fondo delle metodologie studiate e la loro

formalizzazione analitica.

La scelta di questa modalità di verifica è dettata da:

1. La tipologia di contenuti del corso, di natura metodologica.
2. L'importanza che gli studenti acquisiscano una capacità argomentativa e di organizzazione del pensiero e siano in grado di effettuare collegamenti analogici fra le parti del programma, sollecitati dalle domande del docente.

Non sono previste modalità alternative di esame per i non frequentanti, né prove in itinere.

Criteri di valutazione:

1. Conoscenza dei concetti fondamentali del corso.
2. Padronanza delle tecniche fondamentali di riduzione della dimensionalità
3. Copertura degli argomenti.
4. Capacità di collegare le differenti parti del corso e di discuterle e confrontarle criticamente.

Testi di riferimento

Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Wang J. - Springer 2012.

Methods of Multivariate Analysis, Rencher A. C., Wiley 2002

Introduction to Lattices and Order (second edition), Davey B.A., Priestley H. A., CUP 2002 (capitolo 1).

Dispense e articoli forniti dal docente in modalità online.

Periodo di erogazione dell'insegnamento

I semestre II ciclo

Lingua di insegnamento

Italiano

Sustainable Development Goals
