UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Machine Learning

2627-1-F5109P027

Obiettivi

Formare la figura professionale dell'analista dati tramite la metodologia informatica del machine learning.
L'obiettivo viene perseguito;

- fornendo competenze di progettazione, sviluppo e documentazione di studi di machine learning,
- fornendo competenze su software open source professionale per l'estrazione della conoscenza a partire dai dati,
- stimolando e promuovendo il team working come metodo professionale di lavoro e collaborazione.

Al termine del corso lo studente avrà maturato competenze e conoscenze tali da progettare, sviluppare, documentare e presentare uno studio di machine learning. **(DdD 1, DdD 2)**.
Inoltre, l'organizzazione della prova d'esame, tramite un progetto di gruppo da sviluppare e documentare, consentirà al candidato di sviluppare capacità critica **(DdD 3)**, di comunicazione **(DdD 4)**, mentre la disponibilià di materiale da leggere, suggerito dal docente consentirà al candidato di proseguire in modo autonomo lo studio e l'approfondimento della materia **(DdD 5)**.

Contenuti sintetici

Il corso tratta i seguenti argomenti;
Esplorazione dei dati; mostra come progettare e sviluppare workflow di esplorazione dati e di pre-processamento dei dati stessi. Nello specifico mostra come effettuare caricamento di un insieme di dati, come riassumerne quantitativamente le principali caratteristiche, per variabili categoriche, nominali, ordinali e per variabili numeriche. Inoltre, viene mostrato come trattare il problema dei valori mancanti e come ridurre la dimensione dell'insieme di dati sia in termini di attributi che in termini di osservazioni.

- Classificazione Supervisionata; introduce alla formulazione, valutazione e risoluzione di problemi di

classificazione supervisionata, vale a dire problemi per i quali a fronte di un insieme di attributi di input si dispone di uno o più attributi di output che rappresentano le quantità da prevedere utilizzando gli attributi di input a disposizione. Viene dedicata attenzione al tema della progettazione dell'esperimento di apprendimento ed alla valutazione dei relativi risultati. Nello specifico si mostra come stimare le prestazioni di un modello di classificazione, come si comparano le prestazioni di due modelli di classificazione. Infine, viene riservata particolare attenzione al problema della selezione e/o generazione degli attributi tramite opportune procedure algoritmiche.

- Classificazione non Supervisionata; questo argomento è dedicato alla formulazione, valutazione e risoluzione di problemi di classificazione non supervisionata, vale a dire problemi per i quali si dispone solamente di un insieme di attributi di input. In questo caso il compito dell'esperto di machine learning consiste nel progettare ed implementare un workflow che consenta di raggruppare le osservazioni dell'insieme di dati disponibili in modo tale da rendere ottimale l'omogeneità delle osservazioni associate allo stesso gruppo e rendere massima la differenza tra osservazioni assegnate a gruppi differenti. Vengono presentate e discusse diverse misure di similarità utilizzate per valutare l'omogeneità dei gruppi formati dalle procedure e dagli algoritmi di classificazione non supervisionata. Infine, particolare attenzione viene data agli indici di valutazione e comparazione di soluzioni alternative.

Programma esteso

Esplorazione dei dati e Pre-Processing

- Tipi di dati ed attributi
- Esplorazione dei dati, grafica e tabellare
- Trattamento delle osservazioni mancanti
- Pre-processamento dei dati

Classificazione Supervisionata

- Introduzione alla classificazione supervisionata
- Tecniche di classificazione supervisionata; reti neurali, classificatori Bayesiani, alberi di decisione, ...
- Misure di prestazione, procedure di valutazione e comparazione di classificatori
- Classi sbilanciate e problemi di classificazione non binaria

Classificazione Non Supervisionata

- Introduzione alla classificazione non supervisionata
- Misure di prossimità per attributi continui, nominali, ordinali
- Algoritmi di clustering; k-means,algoritmi gerarchici, dbscan, opossum, ...
- Misure di prestazione, procedure di valutazione e comparazione delle soluzioni di clustering

Prerequisiti

Sono utili nozioni base di informatica, calcolo delle probabilità e statistica.

Modalità didattica

L’intero corso è reso disponibile in formato digitale in lingua Inglese. Il corso si compone di lezioni audio-narrate sia per la componente metodologica che per la componente pratica, vale a dire l’impiego di software open source per il machine learning. Lo studente è stimolato a verificare il proprio livello di apprendimento tramite esercizi guidati da realizzarsi con l’impiego del software open source presentato a lezione. Il corso rende inoltre disponibili circa 170 quiz a risposta multipla, tramite la piattaforma Moodle, per consentire allo studente di verificare il proprio livello di preparazione.

- 14 lezioni da 2 ore svolte in modalità erogativa in presenza.
- 9 esercitazioni da 2 ore svolte in modalità erogativa in presenza.

Materiale didattico

Materiale audiovisivo, slides, dataset e workflow progettati e realizzati dai docenti del corso.

https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
https://github.com/kerasking/book-1/blob/master/ML%20Machine%20Learning-A%20Probabilistic%20Perspective.pdf

Periodo di erogazione dell'insegnamento

Primo semestre

Modalità di verifica del profitto e valutazione

La verifica prevede una prova di 4 ore durante la quale il/la candidato/a è chiamato/a a rispondere a 6 quiz (1 punto per ogni risposta corretta) e 1 domanda a riposta aperta (fino a 9 punti) **(DdD 5)**. Inoltre, verrà assegnato un porblema di machine learning da risolvere. Nello specifcio verrà presentato un problema, fornito un dataset e si dovrà progettare e sviluppare un workflow KNIME **(DdD 3, DdD 4)** (ogni nodo del workflow andrà commentato utilizzando le funzionalità offerte dalla piattaforma KNIME, vengono assegnati massimo 10 punti), e si dovranno redigere delle slide secondo un template fornito dal docente (fino ad un massimo di 5 punti. Infine, dietro richiesta del/la candidato/a è possibile richiedere una prova orale che assegna un punteggio compreso tra -3 e +3 punti.
Non sono previste prove in itinere.

Il workflow e le slides verranno valutati in base ai seguenti criteri;

1. Merito tecnico; accuratezza e correttezza del workflow (fino a 10 punti)
2. Chiarezza espositiva e di comunicazione delle idee, includendo leggibilità e visione critica of the slides (fino a 5 punti)

Orario di ricevimento

Su appuntamento, inviare mail a fabio.stella@unimib.it

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | RIDURRE LE DISUGUAGLIANZE