UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Machine Learning

2627-1-FDS02Q002-FDS02Q00201

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

I contenuti del corso sono i seguenti:

- Esplorazione dei dati, per ispezionare e sintetizzare i dati disponibili e per progettare e sviluppare un workflow di pre?processing.
- Classificazione supervisionata, per apprendere una mappatura dagli attributi in input agli attributi in output o target da classificare o predire.
- Classificazione non supervisionata, per formare gruppi omogenei di osservazioni e/o attributi utilizzando una data misura di prossimità.

Il corso include inoltre due moduli verticali:

- **Dai modelli linguistici statistici ai modelli linguistici neurali;** questo modulo introduttivo illustra l'evoluzione dei modelli linguistici, dagli approcci statistici basati su N?grammi alle architetture neurali e ai moderni Large Language Models (LLM). Gli studenti acquisiranno una panoramica dei concetti fondamentali relativi alla rappresentazione del linguaggio, al neural language modeling e al paradigma pretraining–fine?tuning che alimenta i sistemi di IA contemporanei.
- **Un'introduzione graduale alle Reti Bayesiane Causali;** questo modulo verticale introduce le reti bayesiane causali, un modello grafico probabilistico ampiamente utilizzato in diversi domini applicativi come economia, sanità e medicina.

Imparerai a sviluppare workflow di machine learning utilizzando la piattaforma open source KNIME. Non è richiesto programmare, anche se, qualora lo desiderassi, KNIME consente di utilizzare potenti e professionali linguaggi di programmazione open source e ambienti software commerciali: R, Weka, Matlab, Python, Java, …

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 2/3 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 da 3 a 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 problema 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
2. Chiarezza espositiva e di comunicazione delle idee, includendo leggibilità e visione critica delle slides

Orario di ricevimento

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

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | RIDURRE LE DISUGUAGLIANZE