UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Machine Learning M

2526-2-F8204B006

Obiettivi

Lo studente apprenderà le tecniche di Machine Learning più efficaci, comprendendo i fondamenti teorici di ogni tecnica e acquisendo il *know-how* per poterle applicare con successo alla risoluzione di problemi pratici. Sarà inoltre fornita una panoramica sulle più innovative soluzioni per l'identificazione del miglior algoritmo di Machine Learning e della sua configurazione ottimale (Automated Machine Learning – AutoML), dato un dataset. Lo strumento di riferimento per il corso sarà R, ma verranno anche presentate alcune soluzioni equivalenti in Python (ad esempio scikit-learn) e Java (ad esempio WEKA, KNIME).

Contenuti sintetici

Concetti basi del Machine Learning: tipologie di dati, *istanze*, *features*, *tasks* e *scenarios*, *parametri* e *iper-parametri*, misure di performance
Tecniche di apprendimento non-supervisionato
Tecniche di apprendimento supervisionato: classificazione e regressione
Modellare non-linearità nei dati: tecniche basate sul concetto di *kernel*
Automated Machine Learning: configurazione automatica di un modello di Machine Learning

Programma esteso

Introduzione

- Machine Learning scenarios & tasks, notazioni utili
- Tipi di dati e problemi: tabular, streams, text, time-series, sequences, spatial, graph, web, social, immagini, distribuzioni

Unsupervised Learning

- Concetti di similarità e distanza
- Distanze tra punti: Minkowski e casi specifici (Manhattan, Euclidea, Chebyshev/Lagrange), Mahalanobis
- Wasserstein: una distanza (non una divergenza!) tra distribuzioni di probabilità e/o nuvole di punti (datasets)
- Clustering: approcci deterministici vs probabilistici; flat vs gerarchici; basati su distanza/similarità vs densità
- Outlier and anomaly detection

Supervised Learning

- I fondamenti del "learning": classificazione binaria, processo di generazione dei dati, concetto vs ipotesi, errore empirico vs errore di generalizzazione
- Classificazione e regressione: metriche e tecniche di validazione (hold-out, *k* fold-cross, leave-one-out)
- Approcci model-free/instance-based, un semplice algoritmo: k-nearest neighbors (KNN)
- Approcci model-based: Support Vector Machine (lineari)

Supervised Learning per dati non-lineari

- Non-linearità, VC dimensions, kernel-trick
- Un richiamo a Decision Tree e Random Forest
- Kernel-based learning: kernel-SVM e Gaussian Processes per classificazione e regressione
- Dimensionlity reduction: Principal Component Analysis (PCA) e kernel-based PCA (kPCA)

L’approccio connessionista

- Artificial Neural Networks: paradigma di apprendimento
- Deep Learning: *“a fraction of the connectionist tribe”*
- Modelli neurali Generativi: Auto-Encoder (AE) e Variational-AE (VAE), Generative Adversarial Network (GAN) e Wasserstein-GAN (WGAN), Transformer

Automated Machine Learning (AutoML)

- Ottimizzazione degli iperparametri di un algoritmo di Machine Learning
- Selezione del miglior algoritmo di Machine Learning e (simultanea) ottimizzazione dei suoi iperparametri

Esercizi ed esempi pratici

Prerequisiti

Si consiglia la conoscenza di elementi di base di informatica, matematica applicata, probabilità e statistica

Modalità didattica

L'intera attività formativa viene svolta attraverso lezioni in presenza. Le lezioni riguarderanno sia aspetti teorici che applicazioni pratiche, specificatamente l'utilizzo di librerie software e dati open.

Materiale didattico

- Testo di riferimento: Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar (2018). Foundations of Machine Learning.
- Slides e materiale didattico fornito dal docente

Altri tesi suggeriti:

- Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). Mathematics for machine learning. Cambridge University Press.
- Charu C. Aggarwal (2023). Neural Networks and Deep Learning – A Textbook
- Robert B. Gramacy (2020). Surrogates – Gaussian Processes Modeling, Design, and Optimization for the Applied Statistics.
- Charu C. Aggarwal (2015). Data Mining – the Textbook

Periodo di erogazione dell'insegnamento

Primo semestre - primo periodo

Modalità di verifica del profitto e valutazione

La modalità di verifica prevede le seguenti 2 prove:

- lo svolgimento di un progetto con associata redazione di un rapporto tecnico, stile articolo scientifico,
- un esame orale (individuale e obbligatorio) finalizzato a verificare il grado di comprensione degli argomenti trattati.

Il progetto può essere svolto in *team* (max 3 studenti per gruppo) ed i dataset oggetto delle attività saranno concordati con il docente a partire da piattaforme open quali OpenML, Kaggle o UCI Repository.
La qualità del progetto è stabilita sulla base del corretto utilizzo degli algoritmi di ML e all'analisi critiva dei risultati.
L'esame orale è finalizzato alla verifica della comprensione di aspetti teorici e metodologici del ML.
Il progetto contribuisce al 60% della valutazione finale, la prova orale al restante 40%.

Non sono previste prove intermedie.

Orario di ricevimento

Su appuntamento

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ