UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Basi di Dati

1920-2-F6302N030

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.

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.

- **Regole di Associazione;** viene mostrato come possibile apprendere in modo automatico regole di associazione nel caso di insiemi di dati dove le osservazioni sono caratterizzate dalla natura transazionale. In questo caso il compito dell'esperto di machine learning consiste nel progettare e implementare un workflow che consenta di estrarre regole di associazione tra attributi in modo da fornire capacità predittiva e decisionale. 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

- **Regole di Associazione**
    - Introduzione alle regole associative
    - Tipi di itemsets e loro rilevanza
    - Principio ed algoritmo Apriori
    - Misure di prestazione, procedure di valutazione e selezione di regole associative

Prerequisiti

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

Modalità didattica

L'intera attività formativa viene svolta attraverso lezioni in presenza. 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*.

Materiale didattico

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

- http://www.uokufa.edu.iq/staff/ehsanali/Tan.pdf
- http://dsd.future-lab.cn/members/2015nlp/Machine_Learning.pdf
- http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0470276800.html

Periodo di erogazione dell'insegnamento

Primo semestre

Modalità di verifica del profitto e valutazione

La verifica si basa su due componenti complementari, lo svolgimento di un *progetto di machine learning con conseguente redazione di un rapporto tecnico*, stile articolo scientifico, e lo svolgimento di una *prova d'esame, in laboratorio ed a calcolatore* volta a verificare il grado di comprensione metodologica e teorica del candidato. Gli studenti sono incoraggiati al *team working per quanto riguarda la componete progetto dell'esame*, favorendo pertanto il *confronto*, la *discussione* e lo *spirito critico*, componenti irrinunciabili per un contesto complesso come quello oggetto del corso. Il progetto viene scelto dal candidato tra quelli segnalati dal docente come eleggibili a partire da quanto offre la piattaforma Kaggle (**https://www.kaggle.com/**), *piattaforma digitale* ed *internazionale* che offre uno spazio virtuale altamente *professionalizzante* dove si incontrano *domanda ed offerta* nell'ambito della *Data Science*.
Il *progetto di machine learning* attrbuisce un massimo di 21 punti assegnati i base ai seguenti criteri

- Merito tecnico; rigore notazionale, accuratezza e correttezza (massimo 5 punti).
- Chiarezza espositiva e di comunicazione delle idee del candidato, includendo leggibilità e visione critica (massimo 5 punti)
- Inquadramento appropriato del problema trattato (massimo 2 punti)
- Bilanciamento complessivo tra le diverse componenti del report (massimo 3 punti)
- Assenza di ripetizioni, plagio ed auto plagio (massimo 3 punti)
- Qualità dei diagrammi, tabelle, grafici, figure, ... (massimo 3 punti)

La *prova d'esame in laboratorio ed a calcolatore* attribuisce un massimo di 11 punti, ripartiti come segue; 6 punti per 6 quiz a risposta chiusa aventi per oggetto i *concetti presentati nel corso* e massimo 5 punti per una domanda aperta volta a *valutare la capacità critica del candidato*.
Infine, il candidato che lo desiderasse può richiedere di sostenere una *prova orale* (*prevede domande di ragionamento e deduzione su argomenti presentati nel corso*) che attribuisce un massimo di 3 punti.

Orario di ricevimento

Su appuntamento