UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data and Text Mining (blended)

2021-2-F1801Q105

Obiettivi

Formare la figura professionale dell'*esperto di estrazione della conoscenza da dati strutturati, semi-strutturati e non-strutturati*.
La metodologia adottata è rappresentata da **Data e Text Mining**.
L'obiettivo viene perseguito;

- fornendo competenze di *progettazione*, *sviluppo* e *documentazione* di studi di data e text mining,
- presentando i principali algoritmi di apprendimento per dati strutturati, semi-strutturati e non strutturati.
- fornendo competenze su *software open source professionale* per l'*estrazione della conoscenza* a partire dai dati strutturati, semi-strutturati e non strutturati,
- stimolando e promuovendo il *team working* come metodo professionale di lavoro e collaborazione.

Al termine del corso lo studente sarà in grado di *progettare*, *sviluppare*, *documentare* e *presentare* uno studio di **Data e Text Mining**.

Contenuti sintetici

Il corso è strutturato nei seguenti argomenti;

- **Preprocessing di dati strutturati, semi strutturati e non strutturati;** viene mostrato come progettare e

sviluppare workflow di esplorazione e pre-processamento dati strutturati, semi-strutturati e non strutturati.

- **Classificazione Supervisionata;** introduce alla formulazione, valutazione e risoluzione di problemi di classificazione supervisionata. In problemi in questione sono carettirizzati dal fatto che 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. Vengono presentati diversi algoritmi di apprendiment che costituiscono lo stato dell'arte. INoltre, viene dedicata attenzione al tema della progettazione dell'esperimento di apprendimento ed alla valutazione dei relativi risultati.

- **Classificazione non Supervisionata;** 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. Viene mostrato come progettare ed implementare un workflow che consenta di raggruppare osservazioni omogenee e separae osservazioni non omogenee. Vengono presentati i principali algorimi di partizionamento, gerarchici, basati sulla densità, e basati sul concetto di grafo. Viene posta attenzione agli indici di valutazione e comparazione di soluzioni alternative.

- **Regole di Associazione;** mostra come apprendere regole di associazione nel caso di dati transazionali. Viene presentato come progettare e implementare un workflow che consenta di estrarre regole di associazione tra attributi in modo da fornire capacità predittiva e decisionale. Anche in questo caso particolare attenzione viene dedicata agli indici di valutazione e comparazione di regole associative.
- **Text Pre-processing;** vengono illustrati i principali metodi ed algoritmi utilizzati per trasformare testo in linguaggio naturale al fine di renderlo utilizzabile da algoritmi di apprendimento. Vengno nello specifico illustrati i principali passi di preprocessing previsti dal Natural Language Processing.
- **Auto-organizzazione dei Documenti**: vengono presentati i Topic Models, modelli grafico-probabilistici di tipo generativo che consentono di estrarre in mod automatico temi nascosti nel testo in linguaggio naturale. Si tratta di modelli indipendenti dalla lingua e basati solo sul dato di conteggio di occorrenze e co-occorrenze di termini. Questa classe di modelli ha trovato grandissima applicazione nel web, per i sistemi di raccomandazione e per l'indicizzazione del testo in liguaggio naturale.
- **Estrazione dell'Informazione:** tecniche ed algoritmi che consentono di estrarre in modo automatico diversi tipi di entità quali persone, istituzioni, città, luoghi, valute, ... Inoltre, vengono presentati algoritmi per estrarre relazioni tra entità e per il riempimento automatico di template a partire da testo in linguaggio naturale.

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
- **Preprocessing del Testo**
    - Tokenizzazione
    - Filtering e Stemming
    - Modello 0/1, basato su frequenza e modello bag-of-words
    - Misura TF-IDF
- **Categorizzazione del Testo**
    - Schema binario
    - Schema multi-classe
    - Schema multi-etichetta
- **Organizzazione dei Documenti**
    - Clustering di documenti
    - Topic Models
    - Latent Dirichlet Allocation
    - Misure di validazione dei topic
- **Estrazione dell'Informazione**
    - Estrazione di entità
    - Estrazione di relazioni tra entità
    - Previsione di sequenze
    - Applicazioni industriali e commerciali
- **Deep Learning**
    - Introduzione
    - Feedforward neural network
    - cenni alle reti convoluzionali
    - cenni alle reti sequenziali

Prerequisiti

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

Modalità didattica

Le lezioni si svolgeranno in modalità blended, con eventi in presenza in forma di seminari tematici. L'intero corso è reso *disponibile in formato digitale ed 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 data e text mining*. 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 *230 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.
Inoltre, i seguenti testi sono consigliati

- http://www.uokufa.edu.iq/staff/ehsanali/Tan.pdf
- http://dsd.future-lab.cn/members/2015nlp/Machine_Learning.pdf
- https://www.researchgate.net/file.PostFileLoader.html?id=526bc9cfd3df3efa3ec519ee&assetKey=AS%3A272156041121792%401441898465154
- http://www.springer.com/us/book/9781447125655
- https://link.springer.com/book/10.1007%2F978-1-4020-4993-4
- 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 Data e/o Text Mining 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