INTRODUZIONE AL DATA MINING

INF/01 - 9 CFU - 1° semestre

Docente titolare dell'insegnamento

GIOVANNI MICALE
Email: gmicale@dmi.unict.it
Edificio / Indirizzo: Dipartimento di Matematica e Informatica, Blocco III, Stanza 40
Telefono: 0957383071
Orario ricevimento: Mercoledì - 11-13

OBIETTIVI FORMATIVI

Obiettivi formativi generali dell'insegnamento in termini di risultati di apprendimento attesi.

1. **Conoscenza e capacità di comprensione (knowledge and understanding):** Il corso mira a formare le conoscenze e le competenze di base per l'analisi, la rappresentazione, e l'organizzazione di dati.
2. **Capacità di applicare conoscenza e comprensione (applying knowledge and understanding):** lo studente acquisirà conoscenze riguardo ai modelli e gli algoritmi per l'analisi dei dati quali: mining ad alto supporto, sistemi di raccomandazione, ricerca di similarità, classificazione, clustering, reti neurali, analisi di reti.
3. **Autonomia di giudizio (making judgements):** Attraverso esempi concreti e casi di studio, lo studente sarà in grado di elaborare autonomamente soluzioni a determinati problemi legati all'analisi dei dati.
4. **Abilità comunicative (communication skills):** lo studente acquisirà le necessarie abilità comunicative e di appropriatezza espressiva nell'impiego del linguaggio tecnico nell'ambito generale dell'analisi dei dati.
5. **Capacità di apprendimento (learning skills):** il corso si propone, come obiettivo, di fornire allo studente le necessarie metodologie teoriche e pratiche per poter affrontare e risolvere autonomamente nuove problematiche che dovessero sorgere durante una attività lavorativa. A tale scopo diversi argomenti saranno trattati a lezione coinvolgendo lo studente nella ricerca di possibili soluzioni a problemi reali, utilizzando benchmark disponibili in letteratura.

MODALITÀ DI SVOLGIMENTO DELL'INSEGNAMENTO

Lezioni frontali

Qualora l'insegnamento venisse impartito in modalità mista o a distanza potranno essere introdotte le necessarie variazioni rispetto a quanto dichiarato in precedenza, al fine di rispettare il programma previsto e riportato nel syllabus.

PREREQUISITI RICHIESTI

Programmazione, strutture dati, algoritmi su grafi.

FREQUENZA LEZIONI

Le risorse principali messe a disposizione dello studente sono le **lezioni frontali**, la cui frequenza è **fortemente consigliata**.

Per seguire meglio le lezioni, vengono messe a disposizione le **slide** utilizzate per il corso. Le slide non costituiscono un mezzo di studio: forniscono un dettaglio puntuale sugli argomenti trattati a lezione.

Per ulteriori approfondimenti sugli argomenti del corso, verranno indicati riferimenti a libri di testo e risorse online.

CONTENUTI DEL CORSO

Il corso è diviso in due parti, una teorica in cui verranno illustrati i principali problemi di Data Mining e una pratica in cui verrà introdotto il linguaggio R e mostrato come tali problemi possono essere risolti in R. Le due parti saranno portate avanti in parallelo.

Gli argomenti affrontati nel corso sono:

- Introduzione al Data Mining
- Richiami su calcolo delle probabilità
- Linguaggio R
- Data Mining ad alto supporto (apriori, insiemi frequenti, regole di associazione)
- Classificazione (alberi decisionali, SVM, classificatori bayesiani, classificatori lazy, estrattori di regole)
- Clustering (gerarchico, k-means, density-based)
- Sistemi di raccomandazione
- Catene di Markov e HMM
- Introduzione alle reti (Misure di centralità, Coefficiente di Clustering)
- Modelli random di reti
- Graph matching
- Graph mining
- Reti neurali (Feed-Forward, Convolutional, Recurrent, Long-Short Term Memory)

Nell'ambito del linguaggio R, oltre alle funzioni di base, verranno introdotte diverse librerie per l'analisi dei dati, quali "caret" (per la classificazione), "igraph" (per l'analisi e la visualizzazione di reti) e "keras" (per la costruzione di reti neurali).

TESTI DI RIFERIMENTO

Per la parte teorica sull'analisi dei dati, si farà principalmente riferimento a diversi capitoli del libro:

- Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman (http://www.mmds.org).

Altri testi di Data Mining suggerito è

- Data Mining: Concepts and Techniques, Jiawei Han and Micheline Kamber, The Morgan Kaufmann Series in Data Management Systems
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie,Robert Tibshirani, Jerome Friedman, Springer

Per la parte relativa al calcolo delle probabilità il testo suggerito è:

- Statistical methods in Bioinformatics: an Introduction (Second Edition), Warren J. Ewens e Gregory R. Grant, Springer.

Per il linguaggio R si può far riferimento al seguente libro di testo disponibile online:

- The R book (Second Edition), Michael J. Crawley, Wiley (https://www.cs.upc.edu/~robert/teaching/estadistica/TheRBook.pdf).

ALTRO MATERIALE DIDATTICO

Il materiale didattico sarà pubblicato su Studium (https://studium.unict.it/) e, in caso di svolgimento delle lezioni in modalità mista o a distanza, sul canale Teams del corso.

PROGRAMMAZIONE DEL CORSO

	Argomenti	Riferimenti testi
1	Introduzione al data mining	materiale didattico fornito dal docente
2	Richiami su calcolo delle probabilità	materiale didattico fornito dal docente
3	Linguaggio R	materiale didattico fornito dal docente
4	Data Mining ad alto supporto (apriori, insiemi frequenti, regole di associazione)	materiale didattico fornito dal docente
5	Classificazione (alberi decisionali, SVM, classificatori bayesiani, classificatori lazy, estrattori di regole)	materiale didattico fornito dal docente
6	Clustering (gerarchico, k-means, density-based)	materiale didattico fornito dal docente
7	Sistemi di raccomandazione	materiale didattico fornito dal docente
8	Catene di Markov e HMM	materiale didattico fornito dal docente
9	Introduzione alle reti (Misure di centralità, Coefficiente di Clustering)	materiale didattico fornito dal docente
10	Modelli random di reti	materiale didattico fornito dal docente
11	Graph matching	materiale didattico fornito dal docente

12	Graph mining	materiale didattico fornito dal docente
13	Reti neurali (Feed-Forward, Convolutional, Recurrent, Long-Short Term Memory)	materiale didattico fornito dal docente

VERIFICA DELL'APPRENDIMENTO

MODALITÀ DI VERIFICA DELL'APPRENDIMENTO

L'esame finale consiste in **una prova scritta** ed una **prova orale**.

La **prova scritta** è costituita da tre domande di teoria a risposta aperta su argomenti presentati a lezione.

Il voto ottenuto con la prova scritta è il voto di base dell'esame, che può essere incrementato di 2 o 4 punti con la prova orale, a seconda del tipo di esame orale scelto.

La **prova orale** può essere, a scelta:

- Un **progetto** (proposto dallo studente o dal docente, e comunque sempre concordato tra le due parti) che consiste genericamente in un'implementazione pratica (preferibilmente nel linguaggio R visto a lezione) di una soluzione ad un problema di data mining. Col progetto si può ottenere un **incremento massimo di 4 punti** sul voto base dello scritto.
- Un **seminario** (scelto dallo studente da una lista di articoli scientifici proposta dal docente alla fine del corso o proposto dallo studente stesso in accordo col docente), che consiste in una breve presentazione (massimo 10 minuti) in power-point del contenuto di un articolo scientifico che approfondisce problemi visti durante il corso. Col seminario si può ottenere un **incremento massimo di 2 punti** sul voto base dello scritto.

Le due parti dell'esame (prova scritta e prova orale) possono essere sostenute in qualsiasi ordine e anche in sessioni e appelli d'esame diversi.

Il progetto, una volta assegnato, dovrà essere completato entro **90 giorni**.

Salvo diversa comunicazione, la prova scritta si svolgerà alle **ore 9:00** e avrà durata di **1 ora**.

Note:

- È **vietato** l'uso di qualsiasi strumento hardware (calcolatrici, tablet, smartphone, cellulari, auricolari BT etc.), di libri o documenti personali durante la prova scritta.
- Per sostenere gli esami è **obbligatorio prenotarsi** utilizzando l'apposito modulo del portale CEA.
- Non sono ammesse prenotazioni tardive tramite email. In mancanza di prenotazione, l'esame non può essere verbalizzato.
- La verifica dell’apprendimento potrà essere effettuata anche per via telematica, qualora le condizioni lo dovessero richiedere, con la prova scritta rimpiazzata da una prova orale con 3 domande a risposta aperta sui contenuti del corso (Learning assessment may also be carried out on line, should the conditions require it.)