UNIVERSITÀ DI PISA

DATA MINING

DINO PEDRESCHI

Anno accademico	2020/21
CdS	DATA SCIENCE AND BUSINESS INFORMATICS
Codice	420AA
CFU	12

Moduli	Settore/i	Tipo	Ore	Docente/i
DATA MINING: ASPETTI AVANZATI E APPLICAZIONI	INF/01	LEZIONI	48	RICCARDO GUIDOTTI
DATA MINING: FONDAMENTI	INF/01	LEZIONI	48	MIRCO NANNI DINO PEDRESCHI

Obiettivi di apprendimento

Conoscenze

Il corso è suddiviso in due moduli.

DATA MINING: FONDAMENTI

I formidabili progressi della potenza di calcolo, della capacità di acquisizione e memorizzazione dei dati e di connettività hanno creato quantità di dati senza precedenti. Il data mining, ovvero la scienza dell'estrazione di conoscenza da tali masse di dati, si è quindi affermato come ramo interdisciplinare dell'informatica.

Le tecniche di data mining sono state applicate a molti problemi in ambito industriale, scientifico e sociale, e si ritiene che avranno un impatto sempre più profondo sulla società. L'obiettivo del corso è quello di fornire un'introduzione ai concetti di base del data mining e del processo di estrazione della conoscenza, con approfondimenti sui modelli analitici e gli algoritmi più diffusi.

DATA MINING: ASPETTI AVANZATI E APPLICAZIONI

La seconda parte del corso completa le conoscenze del primo modulo con una rassegna delle tecniche avanzate per il mining per dati tabulari e tecniche avanzate di mining per nuove forme di dati. Le tecniche avanzate di classificazione sono relative a reti neurali, SVM, metodi ensemble. Nuovi problemi affrontati sono outlier detection, transactional clustering, time series forecasting e sequential pattern mining. Inoltre vengono analizzati problemi relativi alla privacy, explainability e fairness.

Modalità di verifica delle conoscenze

Per la verifica delle conoscenze acquisite nel corso gli studenti dovranno sostenere una prova scritta che coprirà tutti gli argomenti trattati a lezione. Inoltre sarà chiesto agli studenti di organizzarsi in gruppi di tre unità per collaborare alla realizzazionee un progetto che ha l'obiettivo di analizzare un dataset con i diversi metodi di mining presentati a lezione.

Infine, lo studente dovrà anche sostenere una prova orale sugli argomenti trattati nei due moduli.

Capacità

Al termine dei due moduli lo studente sarà in grado di:

- progettare un KDD process
- applicare le diverse tecniche di mining sulla base delle domande analitiche a cui rispondere
- usare strumenti di mining e librerie python
- simulare il funzionamento di ogni algoritmo di mining presentato a lezione

Modalità di verifica delle capacità

- Lo studente durante lo scritto dovrà svolgere degli esercizi che richiedono la simulazione degli algoritmi di mining
- Lo studente dovrà realizzare e presentare un progetto che richide di analizzare un dataset con i diversi metodi di mining presentati a lezione
- A corredo del progetto, lo studente dovrà preparare anche una relazione scritta che riporti i risultati dell'attività di progetto e l'interpretazione dei risultati trovati.
- Lo studente alla fine svolgerà un esame orale per la verifica delle conoscenze teoriche

Comportamenti

Lo studente potrà maturare abilità nel lavoro di gruppo. Inoltre potrà acquisire e/o sviluppare opportune sensibilità nelle scelte progettuali e di impostazione del processo analitico. Infine, lo studente potrà imparare come intepretare i risultati analitici e come visualizzarli in modo opportuno.

Modalità di verifica dei comportamenti

In fase di esame saranno valutate le scelte progettuali effettuate dal gruppo di studenti e la capacità di elaborazione dei dati con strumenti di analytica e di mining. Inoltre, saranno valutate l'accuratezza e la precisione applicata dal gruppo nello svolgere le attività progettuale.

Indicazioni metodologiche

- Il corso si basa su lezioni frontali, con ausilio di slide, ed esercitazioni sia sulla simulazione degli algoritmi che sull'uso di Python per l'analisi e il mining dei dati (uso del PC personale)
- Tutto il materiale didattico verrà caricato sulla pagina del corso presente sul portale Didawiki.
- Lo studente potrà comunicare con il docente nelle ore di ricevimento e durante le esercitazioni

Programma (contenuti dell'insegnamento)

DATA MINING: ASPETTI AVANZATI E APPLICAZIONI

- CRISP
- Instance-based Classifiers
- Naive Bayes Classifiers
- Linear and Logistic Regression
- Imbalanced Learning
- Dimensionality Reduction
- Outlier Analysis
- Advanced Clustering Methods
- Transactional Clustering
- Support Vector Machines
- (Deep) Neural Networks
- Ensemble Classifiers
- Time Series Distances and Clustering
- Time Series Forecasting
- Time Series Classification
- Sequential Patterns
- Privacy
- Explainability
- Fairness

Bibliografia e materiale didattico

BOOKS

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar. **Introduction to Data Mining**. Addison Wesley, ISBN 0-321-32136-7, 2006
  - http://www-users.cs.umn.edu/~kumar/dmbook/index.php
  - I capitoli 4, 6, 8 sono disponibili sul sito del publisher. – Chapters 4,6 and 8 are also available at the publisher's Web site.
- Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F. **GUIDE TO INTELLIGENT DATA ANALYSIS.** Springer Verlag, 1st Edition., 2010. ISBN 978-1-84882-259-7

SOFTWARE

- KNIME The Konstanz Information Miner. Download page
- Python - Anaconda: Anaconda is the leading open data science platform powered by Python. Download page (the following libraries are already included)
- Scikit-learn: python library with tools for data mining and data analysis Documentation page
- Pandas: pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Documentation page

Indicazioni per non frequentanti

Le esercitazioni svolte in aula e le slides possono essere scaricati dal sito web del corso: http://didawiki.di.unipi.it/doku.php/dm/start

Modalità d'esame

L'esame consiste in una prova scritta sugli argomenti trattati a lezione(sostituita nel periodo Covid-19 dalla prova orale) , un progetto svolto in

gruppi di tre unità e una prova orale che prevede la discussione del progetto e la verifica dell'acquisizione delle nozioni teoriche

Ultimo aggiornamento 11/09/2020 09:56