UNIVERSITÀ DI PISA

DATA MINING

ANNA MONREALE

Anno accademico	2020/21
CdS	INFORMATICA
Codice	309AA
CFU	9

Moduli	Settore/i	Tipo	Ore	Docente/i
DATA MINING	INF/01	LEZIONI	72	ANNA MONREALE

Obiettivi di apprendimento

Conoscenze

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 e metodi del data mining e del processo di estrazione della conoscenza, con approfondimenti sui modelli analitici e gli algoritmi più diffusi considerando anche diverse forme di dato. Inoltre saranno studiati per ogni task di mining aspetti algoritmici e eventuali approfondimenti di varianti algoritmiche delle tecniche di base.

Modalità di verifica delle conoscenze

La verifica delle conoscenze acquisite nel corso si svolgerà attraverso un orale e un progetto.
Per la realizzazione del progetto agli studenti sarà chiesto di organizzarsi in gruppi di due/tre unità.
Il progetto sarà realizzato per step con delle consegne intermedie.
Infine, lo studente dovrà anche sostenere una prova orale sugli argomenti trattati e il progetto, per cui dovrà preparare una presentazione. Lo studente dovrà dimostrare la padronanza dei concetti ma anche dei contenuti del progetto realizzato.
Gli studenti non frequentanti dovranno contattare il docente per concordare il progetto da realizzate e la data dell'orale. Il progetto coprirà tutti gli argomenti trattati nelle varie scadenze intermedie a cui sono stati sottoposti gli altri studenti.

Capacità

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 e implementare strumenti di mining e librerie python
- simulare il funzionamento di ogni algoritmo di mining presentato a lezione

Modalità di verifica delle capacità

- Lo studente dovrà realizzare e presentare un progetto che richiede di analizzare dei dati 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 analitica e di mining. Inoltre, saranno valutate l'accuratezza e la precisione applicata dal gruppo nello svolgere le attività progettuale.

UNIVERSITÀ DI PISA

Indicazioni metodologiche

- Il corso si basa su lezioni 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)

- Concetti di base del data mining e del processo di estrazione della conoscenza
- Preprocessing ed analisi esplorativa dei dati
- Classificazione: alberi di decisione, metodi Bayesiani, Tecniche avanzate di classificazione
- Clustering: metodi basati su partizione, gerarchici, basati su densità, clustering transazionale
- Pattern frequenti, pattern sequenziali e regole associative
- Time series analysis

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

- 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.cli.di.unipi.it/doku.php/magistraleinformatica/dmi/start

Modalità d'esame

L'esame consiste in un progetto svolto in gruppi di due/tre unità e una prova orale che prevede la discussione del progetto e la verifica dell'acquisizione delle nozioni teoriche

Ultimo aggiornamento 09/09/2020 14:53