Corso di Laurea Magistrale in Informatica
a.a. 2017-18

anno di corso	codice	denominazione	cfu	settore scientifico disciplinare	tipo	semestre	tipo attività didattica	cfu	ore	docente
2	F1801Q105	DATA AND TEXT MINING	6	INF/01	a scelta	Primo Semestre	lezione e-learning	5	40	Stella Fabio
							laboratorio	1	12	Stella Fabio

CV docenti: http://www.unimib.it/go/176181440)

Contenuti Analisi dei dati ed estrazione della conoscenza secondo il modello concettuale del data mining. Metodologie per l'analisi di dati strutturati, semi-strutturati e non strutturati. Problemi, modelli ed algoritmi di classificazione, supervisionata e non supervisionata, nel caso di variabili continue, discrete ordinali e nominali e nel caso di variabili miste. Algoritmi di estrazione automatica delle associazioni presenti nei dati. Modelli grafico probabilistici, generativi e discriminativi, per l'analisi di dati testuali semi-strutturati e non-strutturati. Metodi di valutazione della performance previsiva dei modelli. Progettazione ed implementazione del ciclo di data e text mining. Ambienti software, dati e risorse computazionali.

Testi di riferimento Slide e materiale proprietario del docente. In aggiunta verranno resi disponibili capitoli di libri selezionati in base allo specifico argomento trattato.

Obiettivi formativi Pianificazione e conduzione di studi di data mining o di text mining. Progettazione ed implementazione di pipeline e di componenti software per condurre studi di data mining o di text mining. Organizzazione e gestione di progetti di analisi dei dati seguendo le metodologie di data mining o di text mining. Utilizzo di risorse computazionali distribuite per la risoluzione di pipeline o modelli di data mining o di text mining ed la conseguente pubblicazione dei risultati ottenuti. Redazione di report di analisi e commento dei risultati ottenuti tramite uno studio di data o text mining. Redazione di documentazione per l'illustrazione della rilevanza e dei possibili vantaggi competitivi emergenti dai risultati ottenuti tramite il progetto di data mining o di text mining.

Metodi didattici Tutto il corso è fruibile in modalità e-learning. Ogni lezione frontale ed ogni esercitazione è resa disponibile per mezzo di unità video, slide, file di supporto dati e modelli. Il corso è pensato in modo tale che lo studente abbia a disposizione 4 giorni prima della lezione in aula il materiale videoregistrato, le slide associate, i dati ed i modelli necessari. Lo studente assiste alle lezioni video registrate, segnando punti non chiari, dubbi, domande che desidera porre ed argomenti da approfondire.
Lo studente nel corso della lezione frontale o dell'esercitazione è meglio focalizzato sugli argomenti trattati, è maggiormente interattivo con il docente e con i compagni. Inoltre, al termine della lezione può rivedere piu' volte la video registrazione eventualmente arricchita con le sue note che può rendere disponibili ai colleghi di corso. Tutto il corso è organizzato e realizzato per rendere autonomo lo studio da parte dello studente.
Lezioni frontali (circa il 75% del corso) ed esercitazioni (25%) a calcolatore con ambienti software open source su problemi realistici in ambito medico, biologico, finanziario, pubblicitario, social networking, ...

Modalità di verifica dell'apprendimento

Tre prove che richiedono di sviluppare modelli o pipeline di data mining o text mining con gli strumenti software presentati a lezione. Sette prove che prevedono quesiti con risposta multipla, le prove sono da svolgere durante il periodo di attività del corso.
Progetto concordato con il docente da svolgere entro scadenza e da illustrare con presentazione basata su slide.

Programma esteso

1) Data e text mining
1.1) motivazioni
1.2) il ciclo di estrazione della conoscenza a partire dai dati
1.3) tipologie di problemi

2) Esplorazione dati
2.1) tipi di variabili
2.2) misure univariate, bi-variate e multi-variate
2.3) rappresentazioni grafiche

3) Preprocessamento
3.1) binning, normalizzazione, standardizzazione, ...
3.2) selezione delle variabili e costruzione delle feature
3.3) tecniche di riduzione della dimensionalità e di compressione dei dati

4) Classificazione supervisionata e regressione
4.1) problema di classificazione supervisionata; binari, multiclasse e multilabel
4.2) modelli di regressione, euristici, di separazione e probabilistici
4.3) misure di prestazione; accuratezza, precisione, recall
4.4) schemi di stima; hold-out, k-folds cross validation, LOOCV
4.5) regression lineare; modelli, stima, critica e inferenza

5) Classificazione non supervisionata
5.1) problema di clustering
5.2) modelli di partizione, gerarchici, density-based, graph-based, prototype-based
5.3) misure di prestazione e valutazione, misure interne ed esterne

6) Regole di associazione
6.1) problema di associazione
6.2) algoritmo a-priori
6.3) algoritmo tertius

7) Text mining
7.1) preprocessing, vocabolari e tassonomie
7.2) bag-of-word, term frequency e term frequency – inverse document frequency
7.3) classificazione di documenti, pagine web, notizie, .
7.4) latent dirichlet allocation per la scoperta di argomenti
7.5) information extraction e co-reference problem

Master of Science in Computer Science
a.a. 2017-18

year	code	course name	ECTS	type	semester	educational activity type	ECTS	hours	faculty
2	F1801Q105	Data and text mining	6	elective	First semester	e-learning lecture	5	40	Stella Fabio
						laboratory	1	12	Stella Fabio

Professors' CV: http://www.unimib.it/go/176181440

Contents Knowledge discovery in databases. Methodologies for analyzing structured, semistructured and un-structured data. Problems, models and algorithms for supervised and un-supervised classification of continuous data, discrete, ordinal and nominal, data and for mixed data. Association discovery algorithms to automatically extract rules from data. Probabilistic graphical models, generative and discriminative, for analyzing semi-structured and un-structured textual data. Performance evaluation methods for supervised and un-supervised classification. Design and development of the data and text mining cycles. Software environments, data and computational resources.

Textbooks Videolecture, slides, datasets, data mining and text mining workflow, pipelines and models. Occasionally book chapters will be made available to students for specific arguments.

Course objectives

Design and development of data mining or text mining analysis. Design and implementation of pipeline and software components for data mining and text mining. Organize and lead data analysis projects by implementing the data mining and text mining methodologies. Exploitation of distributed computational resources to execute data mining or text mining pipelines and to publish extracted knowledge and results. Write analysis report to describe and comment on obtained results from data mining and text mining projects.

Metodi didattici The course is self contained and it is entirely organized as a set of e-learning lectures. Lectures and recitations are made available on video media, slides, data files and model files. The course has beed designed to allow students to access video material, slides, data and model files, for each lecture at least 4 days before the lecture will take place. The student attends the lecture after having watched the videlectures, and having used data sets and models. Therefore, the student can focus on questions to asks to the teacher when the lecture will take place.
Furthermore, after each lecture the student can watch many times the videolecture eventually enriched with his/her notes or with notes that he/she can share with other students attending the course. Lectures are about 75% while the remaining 25% consists of recitations, i.e. hands on open source software and studying data

mining and text mining problems in several domanins as; medicine, biology, finance, advertisment, social networking, ...

Learning assessments

Three assignments consisting of developing data mining and text mining models, workflow and pipelines by using the open source environment presented during recitations. Seven assignments consisting of a set of multiple choice questions. Assignments are to be done during the active period of the course. 1 project to agree on with the teacher, the project will be documented with a technical report and with a final oral discussion.

Extended syllabus

1) Data and text mining
1.1) motivations
1.2) the knowledge discovery cycle
1.3) problems types

2) Data Exploration
2.1) types of variable
2.2) univariate, bivariate an dmulti-variate measures
2.3) grapical representation"

3) Preprocessing
3.1) binning, normalization, standardization, ...
3.2) variables selection and features construction
3.3) dimensionality reduction and data compression

4) Supervised classification and regression
4.1) supervised classification problem; binary, multiclass and multilabel
4.2) regression models, heuristic, separation and probabilistic
4.3) performance measures; accuracy, precision, recall
4.4) estimation; hold-out, k-folds cross validation, LOOCV
4.5) linear regression; models, estimatiuon, criticism and inference

5) Unsupervised Classification
5.1) the clustering problem
5.2) partitioning, hierarchical, density-based, graph-based, prototype-based
5.3) performance evaluation and validation, internal and external measures

6) Association rules
6.1) the assocaition problem
6.2) a-priori algorithm
6.3) tertius algorithm

7) Text mining
7.1) preprocessing, vucabulary and taxonomies
7.2) bag-of-word, term frequency e term frequency – inverse document frequency
7.3) document classification, web pages, newswire, .
7.4) latent dirichlet allocation for topic extraction
7.5) information extraction and co-reference problem