MATHEMATICS (LM39)

(Università degli Studi)

Teaching DATA MINING

GenCod A004898

Owner professor Massimo CAFARO

Teaching in italian DATA MINING

Teaching DATA MINING

SSD code ING-INF/05

Reference course MATHEMATICS

Course type Laurea Magistrale

Credits 6.0

Teaching hours Ore-Attivita-frontale: 42.0

For enrolled in 2020/2021

Taught in 2020/2021

Course year 1

Language INGLESE

Curriculum PERCORSO COMUNE

Location

Semester Secondo-Semestre

Exam type Orale

Assessment Voto-Finale

Course timetable https://easyroom.unisalento.it/Orario

BRIEF COURSE DESCRIPTION

The course provides a modern introduction to data mining, which spans techniques, algorithms and methodologies for discovering structure, patterns and relationships in data sets (typically, large ones) and making predictions. Applications of data mining are already happening all around us, and, when they are done well, sometimes they even go unnoticed. For instance, how does the Google web search work? How does Shazam recognize a song? How does Netflix recommend movies to its users? The principles of data mining provide answers to these and others questions. Data mining overlaps the fields of computer science, statistical machine learning and data bases. The course aims at providing the students with the knowldedge required to explore, analyze and leverage available data in order to turn the data into valuable and actionable information for a company, for instance, in order to facilitate a decision-making process.

REQUIREMENTS

Calculus. Probability theory. Linear Algebra. Programming skills.

COURSE AIMS

Knowledge and understanding. The course describes methods and models for the analysis of large amounts of data. Students must have a solid background with a broad spectrum of basic knowledge related to data mining:

- the students must have the basic cognitive tools to think analytically, creatively, critically and in an inquiring way, and have the abstraction and problem-solving skills needed to cope with complex systems;
- they must have solid knowledge of data mining models and methodologies;
- they must be able to work on large data collections, including heterogeneous and produced at high speed data, in order to integrate them - in particular by knowing how to manage their origin and quality - and to carry out in-depth thematic analyses, drawing on this knowledge to improve the decision-making process.

Applying knowledge and understanding. After the course the student should be able to:

- describe and use the main data mining techniques;
- understand the differences among several algorithms solving the same problem and recognize which one is better under different conditions;
- tackle new data mining problems by selecting the appropriate methods and justifying his/her choices;
- tackle new data mining problems by designing suitable algorithms and evaluating the results;
- explain experimental results to people outside of statistical machine learning or computer science.

Making judgements. Students must have the ability to process complex and/or fragmentary data and must arrive at original and autonomous ideas and judgments, and consistent choices in the context of their work, which are particularly delicate in the profession of data scientist. The course promotes the development of independent judgment in the appropriate choice of technique/model for data processing and the critical ability to interpret the goodness of the results of the models/methods applied to the datasets under examination.

Communication. It is essential that students are able to communicate with a varied and composite audience, not culturally homogeneous, in a clear, logical and effective way, using the methodological tools acquired and their scientific knowledge and, in particular, the specialty vocabulary. Students should be able to organize effective dissemination and study material through the most common presentation tools, including computer-based ones, to communicate the results of data analysis processes, for example by using visualization and reporting tools aimed at different types of audiences.

Learning skills. Students must acquire the critical ability to relate, with originality and autonomy, to the typical problems of data mining and, in general, cultural issues related to other similar areas. They should be able to develop and apply independently the knowledge and methods learnt with a view to possible continuation of studies at higher (doctoral) level or in the broader perspective of cultural and professional self-improvement of lifelong learning. Therefore, students should be able to switch to exhibition forms other than the source texts in order to memorize, summarize for themselves and for others, and disseminate scientific knowledge.

TEACHING METHODOLOGY	The course aims to provide students with advanced tools for data analysis, through which to extrapolate relevant information from large datasets and guide the related decision-making processes. The course consists of frontal lessons using slides made available to students via the Moodle platform, and classroom exercises. The frontal lessons are aimed at improving students' knowledge and understanding through the presentation of theories, models and methods; students are invited to participate in the lesson with autonomy of judgement, by asking questions and presenting examples. The exercises are aimed at understanding the algorithms and models presented.
ASSESSMENT TYPE	Oral exam. During the exam the student is asked to illustrate theoretical topics in order to verify his/her knowledge and understanding of the selected topics. The student must demonstrate adequate knowledge and understanding of the issues presented or indicated, applying in a relevant manner the theories and conceptual models covered by the study programme.
OTHER USEFUL INFORMATION	**Office Hours** By appointment; contact the instructor by email or at the end of class meetings.

FULL SYLLABUS

Introduzione al corso. Streams. Funzioni hash uniformi, 2-universal e pairwise independent. Streaming: modello turnstile, strict turnstile e cash register. Frequency estimation. Sketches. Count-Sketch. Count-Min. Confronto comparativo tra Count-Sketch e Count-Min. Frequent items. Phi-frequent items. The majority problem. Algoritmo di Boyer-Moore. Algoritmo di Misra-Gries. Algoritmo Frequent.Algoritmo Space Saving. Proprieta' di Space Saving. Confronto comparativo con Frequent. Introduzione al paradigma di programmazione parallela Map-Reduce. Implementazione open-source Hadoop. Pro e contro di Hadoop e Map-Reduce. Distributed File System. Chunk servers, Master node. Map Function. Sort and Shuffle. Reduce Function. Map Tasks. Reduce Tasks. Word counting. Gestione dei guasti. Numero di Map e Reduce jobs. Granularita' dei tasks e pipelining. Mitigare il problema degli strugglers task: spawning di backup tasks. Combiners. Partition (hash) function. Altri esempi di algoritmi Map-Reduce: natural join, two-pass matrix multiply, single pass matrix multiply. Misure di costo per un algoritmo Map-Reduce. Discovery di association rules. Modello market-basket. Esempi di possibili applicazioni. Frequent itemsets. Supporto di un itemset. Association rules. Confidence e Interest.Association rules con elevato interesse positivo o negativo. Mining di association rules. Maximal e closed frequent itemsets. Lattice degli itemsets. Naive approach to counting frequent pairs. Algoritmo A-priori. Monotonicity. Algoritmo PCY. Raffinamenti di PCY: multistage e multihash. Frequent itemsets in 2 passate: random sampling. Frequent itemsets in 2 passate: Random sampling e scelta della soglia opportuna, algoritmo SON, monotonicità, SON parallelo mediante Map-Reduce in 2 passate, algoritmo di Toivonen, bordo negativo. Scene completion problem. Near neighbors in spazi di dimensionalità elevata. Document similarity. Coppie di documenti candidati. Near neighbor search. Jaccard similarity e distance. Shingling: convertire documenti email etc in insiemi. k-shingles. Compressione mediante hashing di k-shingles. Min-Hashing: conversione di insiemi di cardinalità elevata in brevi signatures preservando la similarità. Similarità e distanza di Jaccard per vettori booleani. Boolean matrices. Min-hash signatures. Implementazione. Locality-Sensitive Hashing: determinare coppie di documenti candidate. Matrix partitioning in b bande di r righe: analisi del grado di accuratezza associato rispetto ai falsi positivi ed ai falsi negativi. Link analysis. PageRank. Dead ends. Spider traps. Flow formulation. Matrix formulation. Random walk interpretation. Stationary distribution of a Discrete-Time Markov Chain. Perron-Frobenius Theorem. Google matrix and teleportation. Sparse matrix encoding. Block update algorithm. Topic-specific PageRank. Matrix formulation. Topic vector. Web Spam. Term spam. Spam farms. Analisi del valore di PageRank ottenuto tramite Spam Farm. TrustRank. Trust propagation. Spam Mass estimation. Introduzione al problema del clustering. Curse of dimensionality. Clustering in spazi euclidei e non euclidei. Distanze. Hierarchical clustering: agglomerative and divisive algorithms. Clustering by point assignment. Centroid and clustroid. K-means e K-means++. Scelta di k: elbow criterion. Algoritmo BFR. Discard, Compression e Retained sets. Summarizing points. Distanza di Mahalanobis. Algoritmo CURE. Punti rappresentativi e loro scelta. Input space e feature space. Kernel methods. Kernel matrix. Linear kernel. Kernel trick. Kernel operations in feature space. Represenattive clustering: K-means e Kernel K-means. Expectation-Maximization clustering. Hierarchical clustering. Density-based clustering. Algoritmo DBSCAN. Recommender systems. Recommendations. The long tail phenomenon. Content-based systems. Utility function and matrix. Ratings. Extrapolation of ratings (utilities). Item profiles. User profiles. Collaborative filtering. k-NN. Similarity metrics. User-user and item-item collaborative filtering. Evaluation of systems. Error metrics. RMSE, precision, rank correlation. Complexity of collaborative filtering. The Netflix challenge. Bellkor recommender system. Modeling local and global effects. Learning the optimal weights: optimization problem and gradient descent. Latent factor models. SVD decomposition. Learning the P and Q matrices. Preventing overfitting: regularization. Stochastic Gradient Descent. Biases and interactions. Temporal biases and factors. Machine learning: supervised and unsupervised approaches. Attributi numerici e categorici. Attributi categorici nominali ed ordinali. Probabilistic classifiers. Parametric approach: Bayes and naive Bayes classifiers. Data centering. Non parametric approach (density based): K-nearest neighbors classifier. Decision Trees. Hyperplans. Split points. Data partion and purity. Split Point Evaluation Measures:

entropy, split entropy, information gain, Gini index, CART. Valutazione di split points numerici e categorici. Support Vector machines. Hyperplanes. Support Vectors and Margins. Linear and Separable Case. Soft Margin SVM: Linear and Nonseparable Case. Kernel SVM: Nonlinear Case. SVM Training Algorithms. Multiclass SVM. Analisi delle prestazioni di un classifier. Metriche di valutazione. ROC curve e AUC. K-fold cross-validation. Bootstrapping. Intervalli di confidenza. Paired t-Test. Bias and variance decomposition. Ensemble classifiers. Bagging. Random Forest. Boosting. Stacking. Introduction to neural networks. History. The Biological Inspiration. Learning in Biological vs Artificial Networks. An Alternative View: The Computational Graph Extension of Traditional Machine Learning. Machine Learning versus Deep Learning. Single Layer Networks: the Perceptron. Bias neurons. Training a Perceptron. Perceptron vs Linear SVMs. Activation and Loss Functions. Multilayer Neural Networks. Reasons for nonlinearity of hidden layers. Role of Hidden Layers. The Feature Engineering View of Hidden Layers. Multilayer Networks as Computational Graphs. Connecting Machine Learning with Shallow Neural Networks: Perceptron versus Linear SVM, RBF Network versus kernel SVM.

REFERENCE TEXT BOOKS

Mining of Massive Datasets
J. Leskovec, A. Rajaraman and J. Ullman
Freely availableonline: http://www.mmds.org

Data Mining and Analysis
M. J. Zaki and W. Meira
Freely available online: http://dataminingbook.info

Neural Networks and Deep Learning
Charu C. Aggarwal
Springer