Data Mining

Yannis Kotidis, Associate Professor AUEB, kotidis@aueb.gr

Ioanna Filippidou, Ph.D. Candidate, filippidoui@aueb.gr (Teaching Assistant)

Overview

Data-oriented techniques for extracting patterns from data. Association rules discovery and usage, ratio rule mining. Dimensionality reduction techniques, data clustering and classification. Link analysis, mining social network graphs. Similarity computations, nearest neighbors, collaborative filtering and recommendation algorithms. Mining data streams.

Key Outcomes

By completing the course the students will be able:

- To fully understand standard data mining methods and techniques such as association rules, data clustering and classification.
- Learn new, advanced techniques for emerging applications (e.g. social network analysis, stream data mining).
- Gain practical intuition about how to apply these techniques on datasets of realistic sizes using modern data analysis frameworks.

Requirements and Prerequisites

The course requires a good knowledge in computer science, algorithms and data management. A basic knowledge of statistics and probability theory is essential.

Required Course Material

There is no required textbook. All course materials will be provided in class and will be available for downloading.

Books

There are many books on the subject. The students are encouraged to download and read the following textbook

- Mining of Massive Datasets (Jure Leskovec, Anand Rajaraman, Jeff Ullman)

The textbook is available from the following web page: http://www.mmds.org/. Additional material in the form of academic papers or presentations will be available for download.

Software/Computing requirements

Students will be able to run and work with most of the course material on their own computers. The students will be able to complete the first project assignments using WEKA (https://sourceforge.net/projects/weka/). The second programming assignment will be implemented using the neo4j native graph database. The neo4j community edition is available for download at http://neo4j.com/download/.

Grading

Students will be graded on their performance in the final exam and two project assignments. More precisely, the grading is divided as follows:

- In class participation will count towards 5% of the grade.
- Final exam will count 55% towards the final grade.
- The first project assignment will be announced after the fifth unit and will count for 20% towards the final grade.
- The second assignment will be announced after the eighth unit and will count for the remaining 20% of the final grade.

Project Assignments

Late assignments will either not be accepted or will incur a grade penalty unless due to documented serious illness or family emergency. Exceptions to this policy for reasons of civic obligations will only be made available when the assignment cannot reasonably be completed prior to the due date. The students should make suitable arrangements, and give notice for late submission in advance.

Participation

In-class contribution is an important part of our shared learning experience. Your active participation helps us to evaluate your overall performance. Please arrive to class on time and stay to the end of the class period. Chronically arriving late or leaving class early is unprofessional and disruptive to the entire class. Turn off all electronic devices prior to the start of class. Cell phones tablets and other electronic devices are a distraction to everyone. In lectures you need to use laptop you will be informed to do so.

Attendance Requirements

Class attendance is essential to succeed in this course and is part of your grade. An excused absence can only be granted in cases of serious illness or grave family emergencies and must be documented. Job interviews and incompatible travel plans are considered unexcused absences. Where possible, please notify the instructor in advance of an excused absence. Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives express permission. Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one’s own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should

be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Course Syllabus

The course comprises ten units of three hours each.

Unit 1: Introduction to the basics of Data Mining

Introduction to data mining. The big data ecosystem and its implications to data mining. From big-data to big-errors. The Bonferroni's principle. New emerging techniques in mining streaming and static datasets.

Unit 2: Association rules

Motivation of association rule mining. Definition and statistical properties of association rules. The a-priori algorithm. Performance considerations and possible enhancements. The PCY algorithm. Criticism to association rule mining. Extensions to the basic scheme.

Unit 3: Singular Value Decomposition

Intuition and definition of SVD. Using SVD for mining basket data. Dimensionality reduction, compression and visualization with SVD. Using singular vectors as ratio rules. Other uses of SVD (outlier detection, similarity computation, information retrieval, correlation analysis).

Unit 4: Link Analysis

Examples of networked datasets. The WWW as a graph. Motivation and definition of pageRank. Topic-specific pageRank. HITS and its connection to SVD. Fighting spam, spamRank.

Unit 5: Similarity, Nearest Neighbor Search and Collaborative Filtering

Definition of nearest-neighbors and their applications. Set similarity measures. Converting text to sets via minhashing. Locality Sensitive Hashing for fast, all-pair similarity computations. Use of nearest neighbors in collaborative filtering.

Unit 6: Clustering

Intuition, hierarchical clustering, k-means, bisecting k-means, CURE. Density based clustering, the DBSCAN algorithm. Clustering in social graphs, introduction to graph partitioning. Contextual and structural similarity metrics for social graphs.

Unit 7: Classification

Definition and example use cases. Decision and regression trees. Split decisions, classification error, entropy and information gain. Avoiding overfitting.

Unit 8-9: Graph analytics

From relational databases to big-data systems and native graph databases. The property graph model. Social network analysis examples (centrality metrics, node clustering coefficient, friend suggestion, etc). Parallel graph computation engines. Applications examples (pageRank, PBFS, semi-clustering).

Unit 10: Data stream mining

Data stream processing: definition, assumptions and challenges. Applications of data streams in sensor networks, web traffic and social networks. Sampling data streams, challenges and restrictions. Overview of randomized algorithms for stream processing. Sketches and their applications in frequency moment estimation. Data streams counting schemes.