

Data Mining – An introduction

Marc Plantevit

Université Claude Bernard Lyon 1 – LIRIS CNRS UMR5205



* Slides from different research school lectures and conference tutorials.

2021

About me.

marc.plantevit@univ-lyon1.fr or marc.plantevit@liris.cnrs.fr

Associate Professor, HDR

Computer Science Dept.

University Claude Bernard Lyon 1.

Lab: LIRIS UMR 5205

Team: Data Mining & Machine Learning (head since 2019)

Research Interest: Foundations of constraint-based pattern mining, sequences, augmented graphs, subgroup discovery, XAI.



Evolution of Sciences

Before 1600: Empirical Science

- ▶ Babylonian mathematics: 4 basis operations done with tablets and the resolution of practical problems based on words describing all the steps. \Rightarrow that worked and they manage to solve 3 degree equations.
- ▶ Ancient Egypt: No theorization of algorithms. We give only examples made empirically, certainly repeated by students and scribes. Empirical knowledge, transmitted as such, and not a rational mathematical science.
- ▶ Aristotle also produced many biological writings that were empirical in nature, focusing on biological causation and the diversity of life. He made countless observations of nature, especially the habits and attributes of plants and animals in the world around him, classified more than 540 animal species, and dissected at least 50.
- ▶ ...

1600-1950s: Theoretical Science

Each discipline has grown a theoretical component. Theoretical models often motivate experiments and generalize our understanding.

- ▶ Physics: Newton, Max Planck, Albert Einstein, Niels Bohr, Schrödinger
- ▶ Mathematics: Blaise Pascal, Newton, Leibniz, Laplace, Cauchy, Galois, Gauss, Riemann
- ▶ Chemistry: R. Boyle, Lavoisier, Dalton, Mendeleev,
- ▶ Biology, Medicine, Genetics: Darwin, Mendel, Pasteur



1950s–1990s, Computational Science

- ▶ Over the last 50 years, most disciplines have grown a third, computational branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
- ▶ Computational Science traditionally meant simulation. It grew out of our inability to find closed form solutions for complex mathematical models.




The Data Science Era

1990's-now, Data Science

- ▶ The flood of data from new scientific instruments and simulations
- ▶ The ability to economically store and manage petabytes of data online
- ▶ The Internet and computing Grid that makes all these archives universally accessible
- ▶ Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes.

The Fourth Paradigm: Data-Intensive Scientific Discovery

Data mining is a major new challenge!

 The Fourth Paradigm. Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Research, 2009.

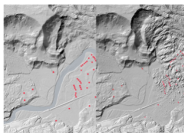
Evolution of Database Technology

- ▶ 1960s: Data collection, database creation, IMS and network DBMS
- ▶ 1970s : Relational data model, relational DBMS implementation
- ▶ 1980s: RDBMS, advanced data models (extended-relational, OO, deductive, etc.), application-oriented DBMS (spatial, scientific, engineering, etc.)
- ▶ 1990s: Data mining, data warehousing, multimedia databases, and Web databases
- ▶ 2000s: Stream data management and mining, Data mining and its applications, Web technology (XML, data integration) and global information systems, NoSQL, NewSQL.

Why Data Mining?

- ▶ The Explosive Growth of Data: from terabytes to petabytes
 - ▶ Data collection and data availability
 - ▶ Automated data collection tools, database systems, Web, computerized society
- ▶ Major sources of abundant data
 - ▶ Business: Web, e-commerce, transactions, stocks, . . .
 - ▶ Science: Remote sensing, bioinformatics, scientific simulation, . . .
 - ▶ Society and everyone: news, digital cameras, social network, . . .
 - ▶ "We are drowning in data, but starving for knowledge!" – John Naisbitt, 1982 –

Applications



- ▶ Human mobility (ANR VEL'INNOV 2012–2016)
- ▶ Social media (GRAISearch - FP7-PEOPLE-2013-IAPP, Labex IMU project RESALI 2015–2018)
- ▶ Soil erosion (ANR Foster 2011–2015)
- ▶ Neuroscience (olfaction)
- ▶ Chemoinformatics
- ▶ Fact checking (ANR ContentCheck 2016 – 2019)
- ▶ Industry (new generation of product, failure detection)
- ▶ ...

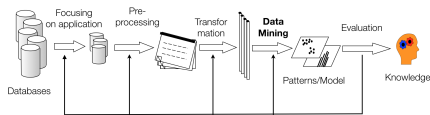
What is Data Mining

- ▶ Data mining (knowledge discovery from data)
 - ▶ Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- ▶ Alternative names:
 - ▶ KDD, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- ▶ **Watch out: Is everything “data mining”?**
 - ▶ simple search or query processing
 - ▶ (Deductive) expert systems


KDD Process

Data Mining

- ▶ Core of KDD
- ▶ Search for knowledge in data



Iterative and Interactive Process

 Fayad et al., 1996

Functionalities

- ▶ **Descriptive data mining** vs Predictive data mining
- ▶ **Pattern mining**, classification, clustering, regression
- ▶ Characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.







Major Issues In Data Mining

- ▶ Mining methodology
 - ▶ Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web.
 - ▶ Performance: efficiency, effectiveness, and scalability
 - ▶ Pattern evaluation: the interestingness problem
 - ▶ Incorporation of background knowledge.
 - ▶ Handling noise and incomplete data
 - ▶ Parallel, distributed and incremental mining methods.
 - ▶ Integration of the discovered knowledge with existing one: knowledge fusion.
 - ▶ Completeness or not.
- ▶ User interaction
 - ▶ Data mining query languages and ad-hoc mining.
 - ▶ Expression and visualization of data mining results.
 - ▶ Interactive mining of knowledge at multiple levels of abstraction
- ▶ Applications and social impacts
 - ▶ Domain-specific data mining & invisible data mining
 - ▶ Protection of data security, integrity, and privacy.

Where to Find References? DBLP, Google Scholar

- ▶ Data Mining and KDD
 - ▶ Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - ▶ Journals: Data Mining and Knowledge Discovery, ACM TKDD
- ▶ Database Systems
 - ▶ Conferences: : ACM-SIGMOD, ACM-PODS, (P)VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - ▶ Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- ▶ AI & Machine Learning
 - ▶ Conferences: Int. Conf. on Machine learning (ICML), AAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc
 - ▶ Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- ▶ Web and IR
 - ▶ Conferences: SIGIR, WWW, CIKM, etc
 - ▶ Journals: WWW: Internet and Web Information Systems,

Recommended Books

-  U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996
-  J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd ed., 2006
-  D. J. Hand, H. Mannila, and P. Smyth, *Principles of Data Mining*, MIT Press, 2001
-  P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Wiley, 2005
-  Charu C. Aggarwal, *Data Mining*, Springer, 2015.
-  Mohammed J. Zaki, Wagner Meira, Jr. *Data Mining and Analysis Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.

ML versus DM

Predictive (global) modeling

- ▶ Turn the data into an as accurate as possible prediction machine.
- ▶ Ultimate purpose is **automatization**.
- ▶ E.g., autonomously driving a car based on sensor inputs



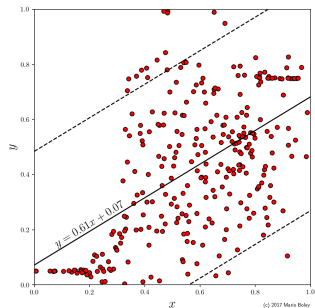
M. Boley www.realkd.org

Exploratory data analysis.

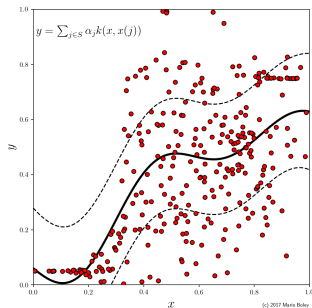
- ▶ Automatically discover novel insights about the domain in which the data was measured.
- ▶ Use machine discoveries to synergistically **boost** human expertise.
- ▶ E.g., understanding commonalities and differences among PET scans of Alzheimer's patients.

ML versus DM

“A good prediction machine does not necessarily provide explicit insights into the data domains”



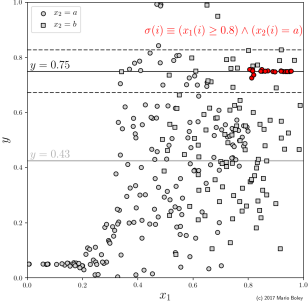
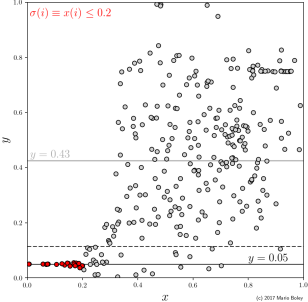
Global linear regression model



Gaussian process model.

ML versus DM

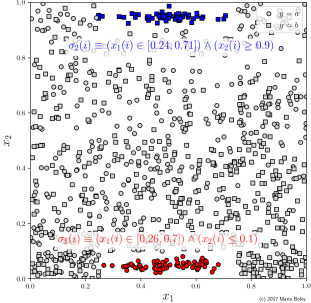
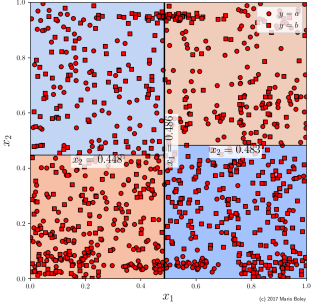
“A complex theory of everything might be of less value than a simple observation about a specific part of the data space”



Identifying interesting subspace and the power of saying “I don't know for other points”

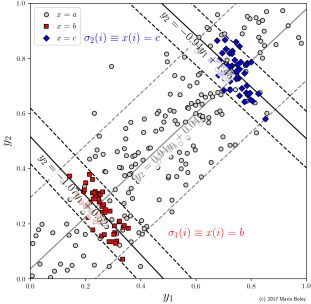
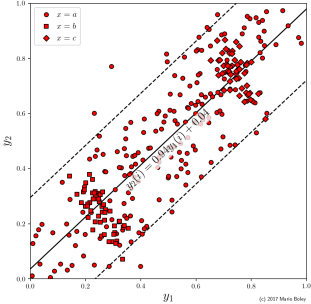
ML versus DM

“Subgroups look similar to decision trees but good tree learners are forced to brush over some local structure in favor of the global picture”



ML versus DM

“Going one step further, we can find local trends that are opposed to the global trend”



Roadmap

We will focus on **descriptive data mining** especially on Constraint-based Pattern Mining with an **inductive database vision**.

$$Th(\mathcal{L}, \mathcal{D}, \mathcal{C}) = \{\psi \in \mathcal{L} \mid \mathcal{C}(\psi, \mathcal{D}) \text{ is true}\}$$

- ▶ Pattern domain: (itemset, sequences, graphs, dynamic graphs, etc.)
- ▶ Constraints: How to efficiently push them?

 Imielinski and Mannila: Communications of the ACM (1996).

Roadmap

- ▶ Clustering techniques
- ▶ (constraint-based) Pattern mining
- ▶ Subgroup discovery / Exceptional model mining
- ▶ Classification