Data Mining and Machine Learning
Bioinspired computational methods
Biological data mining

Introduction

Francesco Marcelloni

Department of Information Engineering
University of Pisa
ITALY

e-mail: francesco.marcelloni@unipi.it

Programme 6 ECTs

- Data Preprocessing
- Classification and Prediction
- Cluster Analysis
- Mining Frequent Patterns, Associations and Correlations

Programme 12 ECTs

- Data Preprocessing
- Classification and Prediction
- Cluster Analysis
- Mining Frequent Patterns, Associations and Correlations
- Outlier detection
- Mining Stream, Time-series and Sequence Analysis
- Graph Mining
- Steaming Data Mining
- Hadoop and Mahout

Course

- Lectures: prof. Francesco Marcelloni
  - Reception hours: Wednesday 15-18 (please, send me an e-mail for confirmation)
- Practical and laboratory work: prof. Fabrizio Ruffini
  - Reception hours: Monday 16-18

Material

- Teaching material:
  - J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, Third edition
  - Jiawei Han, Jian Pei, Hanghang Tong, Data Mining: Concepts and Techniques. Morgan Kaufmann, Fourth edition
  - Papers on the different algorithms described during the course

- Slides available in the Teams channel of the course

Acknowledgement

Some slides belong to the collection

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University

This collection accompanies the book

J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011

Project (12 CFU)

- Project
  - The project consists of the **development of one application** which exploits one or more techniques introduced in the lectures. Applications can be stand-alone, Web applications, mobile applications and so on. For instance, event detection, recommender systems, user profiling, sentiment analysis. **The applications can be developed in groups of two persons at most**.
- Schedule
  - Application specification (**has to be approved**)
  - Analysis and Design of the application
  - Implementation (**recommended Python**) and Validation
  - Presentation of the application 2-3 days before the examination (you have to deliver source code, executable version and documentation)

Examination (12 CFU)

- Examination
  - 2-3 days before the official date of the examination:
    - **Presentation of the application and discussion**. The date for the presentation will be agreed with us. The application will be assessed in terms of originality, coherence with the specifications, appropriate choice of the data mining techniques, results obtained in the validation, usability.
    - A mark (between 18 and 30 in case of positive evaluation) will be assigned to each project. Only students who have obtained positive evaluations will be entitled to take the examination.
  - Date of the examination: written test on theoretical aspects of the course and final discussion on the answers or directly an oral.
    - The final mark will be computed as the average of the marks obtained in the two evaluations.

Examination (6 CFU)

Examination

- Date of the examination:
  - **Practical test using Python**. We will give you some dataset to be analyzed by using specific methods
  - After the analysis you will prepare a short presentation with the results of the analysis
  - A mark (between 18 and 30 in case of positive evaluation) will be assigned to the presentation. Only students who have obtained positive evaluations will be entitled to participate to the second test.
- Written\Oral test: written test on theoretical aspects of the course and final discussion on the answers or oral.
  - The final mark will be computed as the average of the marks obtained in the two evaluations.

Glossary

- Machine Learning
  - (The English Oxford Dictionary)
    "The capacity of a computer to learn from experience, i.e. to modify its processing on the basis of newly acquired information.

- Data Mining
  - (Merrian-Webster)
  - the practice of searching through large amounts of computerized data to find useful patterns or trends

Machine Learning

- Machine Learning

Supervised Learning
Learning with a labeled training set

Unsupervised Learning
Discovering patterns in unlabeled data

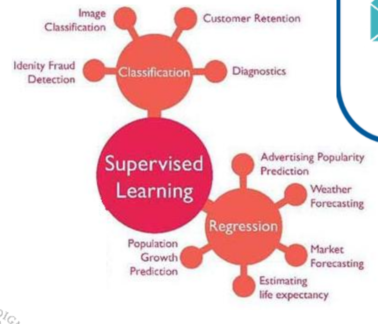
Reinforcement Learning
Learning based on feedback or reward

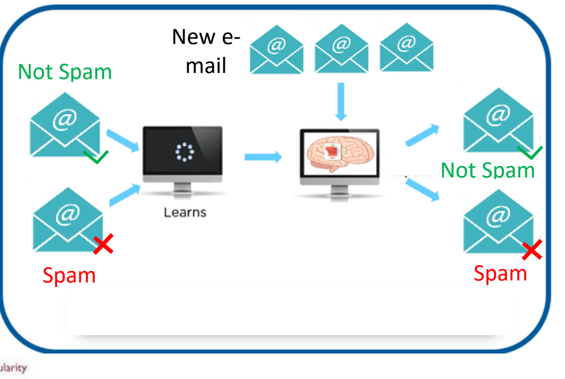
Semi-supervised Learning

Supervised Learning

- Supervised Learning

Learning with a labeled training set

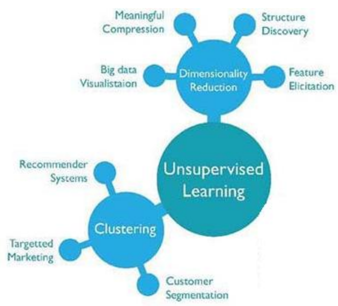


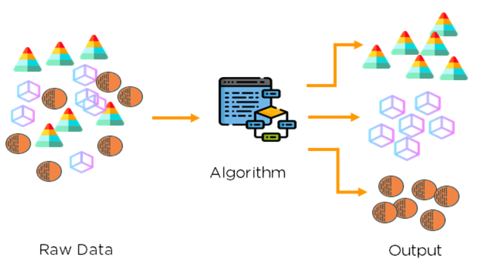


Unsupervised Learning

- Unsupervised Learning

Discovering patterns in unlabeled data

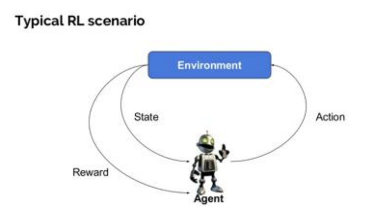


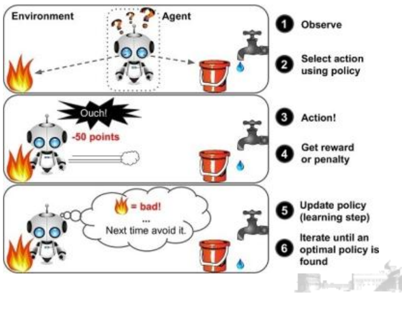


Reinforcement Learning

- Reinforcement Learning

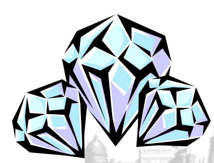
Learning based on feedback or reward.
No dataset needed at beginning



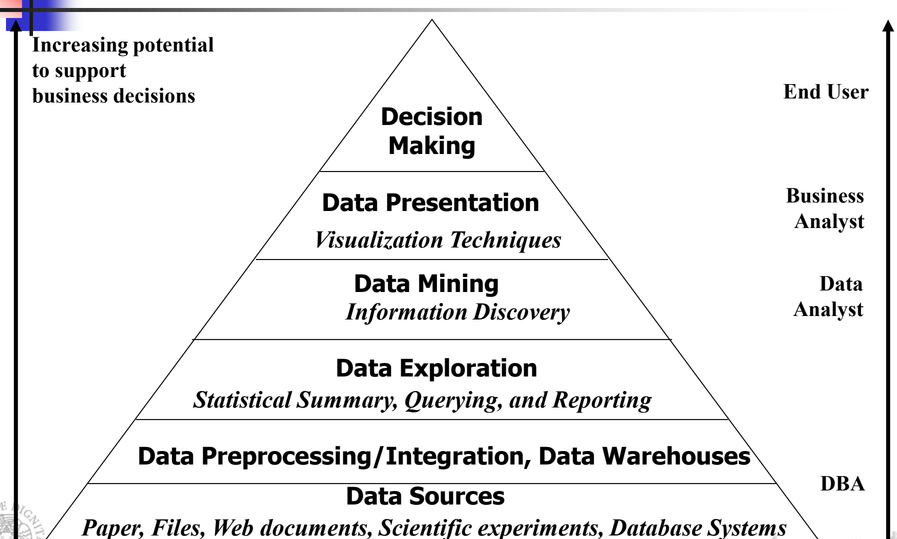


What is Data Mining

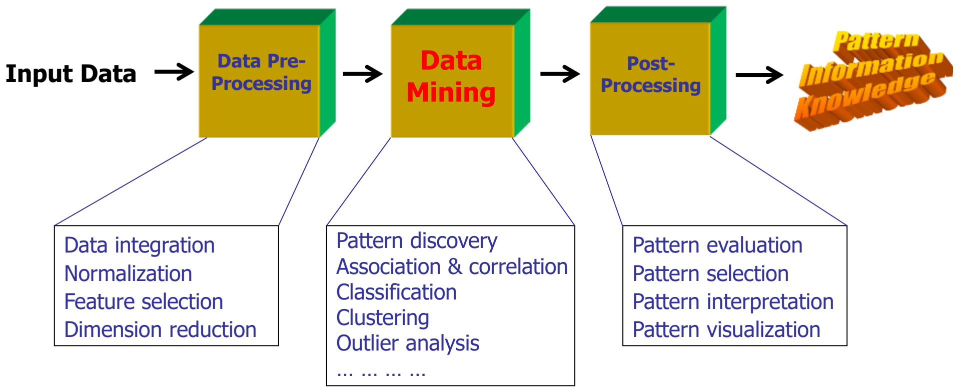
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge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



Data Mining in Business Intelligence



A Typical View from ML and Statistics



This is a view from typical machine learning and statistics communities

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Data Mining Function: Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper -> Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Sequence Motifs (nucleotide or amino-acid sequence pattern) and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

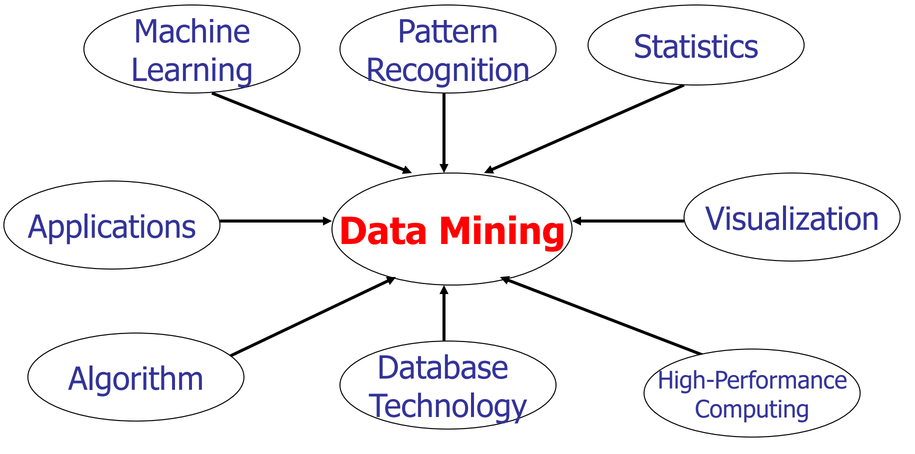
Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Major Issues in Data Mining (1)

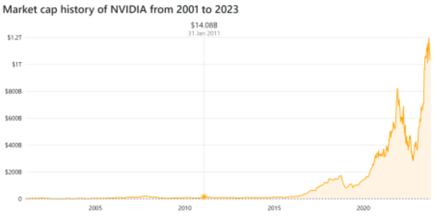
- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining
  - Trustworthiness

Why only today?

- **Technological development**: vast computing power and nearly infinite storage capacity enable the execution of increasingly complex machine learning algorithms using enormous amounts of data.



The sixth most valuable company in the world by market capitalization.

Why only today?

- **New learning algorithms:** increasingly advanced and sophisticated algorithms capable of training ever more complex models

GPT-1: 117 million parameters
GPT-2: 1.5 billion parameters
GPT-3: 175 billion parameters
Google Bard: 137 billion parameters
GPT-4: 1.8 trillion parameters (estimated)
Llama 4: 2 trillion parameters (estimated)

GPT-3: **26 days and 1248 MWh** for training (the equivalent of the annual electricity consumption of approximately 400 households with 4 people)
Google Bard: 13 days and 312 MWh for training

University of Pisa
Francesco Marcelloni

Why only today?

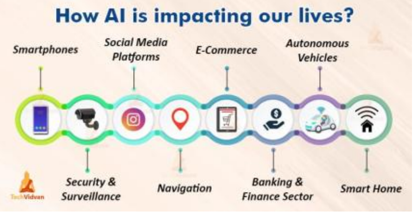
- **Availability of large amounts of data:** thanks to social networks, digital platforms, and the Internet of Things, which allows us to collect data from sensors deployed everywhere, etc.

GPT-3: **45TB** of text data from different datasets

Dataset	Quantity (tokens)
Common Crawl (filtered)	410 billion
WebText2	19 billion
Books1	12 billion
Books2	55 billion
Wikipedia	3 billion

Why only today?

- Impact on everyday life: AI has become pervasive



https://techvidvan.com/tutorials/ai-in-human-life/

- Significant investments: AI is a disruptive technology, and both companies and nations are investing heavily in it.

GPT-3: $10 billion (with $3 billion already invested)

Bard: $300 million