Machine Learning and Data Mining

Claudio Sartori

DISI
Department of Computer Science and Engineering – University of Bologna, Italy
claudio.sartori@unibo.it

Reference materials

[Alpaydin, 2014] an excellent theoretical book on Machine Learning (available in the G.P. Dore library)

[Hastie et al., 2009] an excellent book on Statistical Learning foundations

[Witten et al., 2016] a practical textbook which is companion of the popular software suite *Weka*

Pre–requisites

Refresh your knowledge on:

- elements of statistics and probability / applied mathematics
- python programming
- (databases and SQL)

An extremely short story of Information Technology

'60s the early data collections and databases

'70s the early database management systems

'80s DBMS maturity, new types of data, new access paradigms, early attempts to derive insights from data

'90s web, data warehousing, knowledge discovery in databases

2000– big data explosion

From day by day operations to analysis

- in the beginning data were mainly used to perform day-by-day operations: inventory, billing, census, ...
- most of the data were used once, then merely stored for archival purposes
- can we use stored data to improve our *decision processes*?
- can we learn from data?

Drowning in data

- Data explosion
  - automatic data collection, mature DBMS technology, cheap storage

  $\Rightarrow$ huge amounts of stored data are available
- it is much easier to store data than to analyze them

$\Rightarrow$ increasing distance between data generation and data comprehension

- we are drowning in data and starved for information [1]

1 The original quotation was *We are drowning in information but starved for knowledge*, John Naisbitt, Megatrends, 1982

How to learn from data 1/3

In the beginning was the *statistics*

- Numerical statements of facts in any department of inquiry placed in relation to each other
- Since the 18th century
  - descriptive
  - inferential
  - statistical models

How to learn from data 2/3

Machine Learning

- Field of study that gives computers the ability to learn without being explicitly programmed
- Since late '50s of XX century
  - learning by being told
  - *learning from examples*

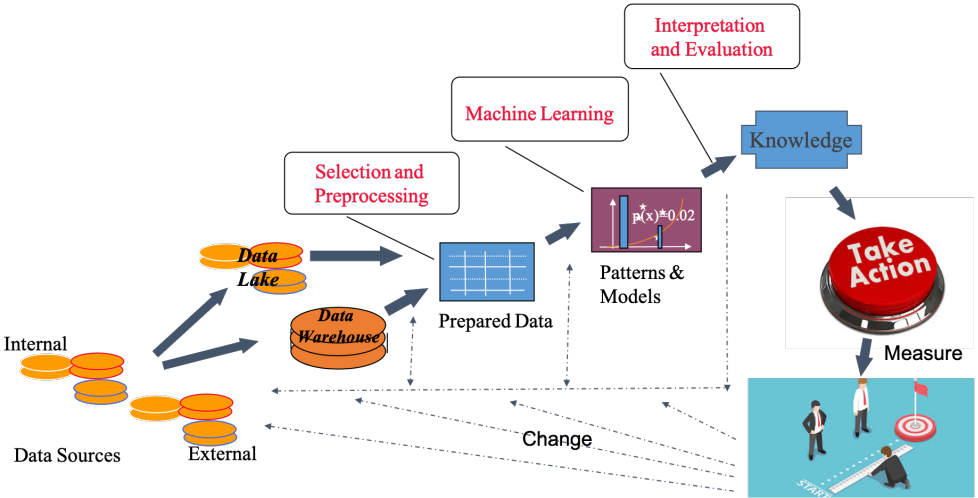
How to learn from data 3/3

Data Mining [2]

- a computational process to discover patterns in data
- discovery of patterns in large data sets digitally stored
- the term was used since the early '90s of the XX century
- Data Mining uses concepts and tools from:
  - artificial intelligence
  - machine learning
  - statistics
  - DBMS technology
- *data–driven approach*

2 Curiosity: the name is wrong, because we are not mining for *data*, which are already there and available, but for *patterns*

Discovery *process*



A very short terminology

Business Intelligence – Analyse massive amounts of data with various tools for business purposes

Analytics – Learn to draw specific conclusions from the observation of raw data

Example – decide if a single credit card transaction is fraudulent

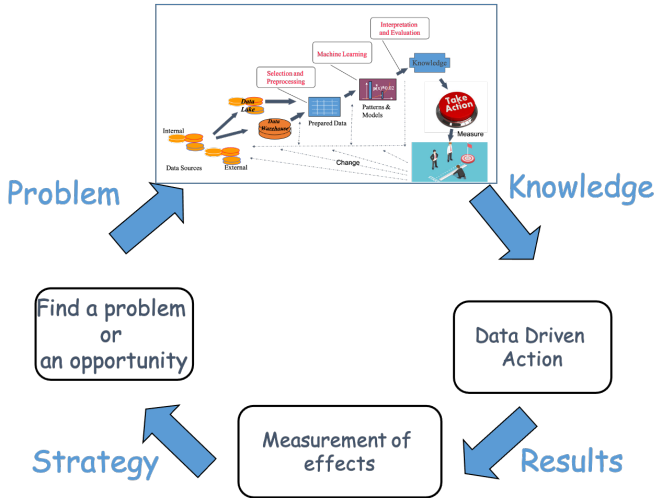
Data Mining – The entire *discovery process* sets up a *pipeline* starting from the data and ending with the general patterns and their translation into useful actions

Machine Learning – the application of methods and algorithms to extract the patterns from the data

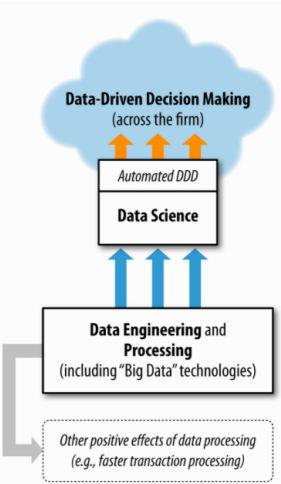
What is the Data Science?

- A broader term including data mining and all of the terms mentioned in the previous slide
- A *data scientist* is intended to have skills in all of the areas mentioned above

The virtuous loop



Data–driven decision making and IT



Soybean diseases [Michalski and Chilausky, 1980] 1/4

An early success of *machine learning*

- diagnosis of soybean diseases
- experts use *rules*
- available a database of diagnosed examples
  - 307 individuals
  - 35 attributes
  - 19 different diseases
  - diagnosed disease for each individual
- see the dataset at https://archive.ics.uci.edu/ml/datasets/Soybean+(Large)

Soybean diseases 2/4

Example of formalized rules from experts

```
If leaf condition = normal and
   stem condition = abnormal and
   stem cankers = below soil line and
   canker lesion color = brown
then diagnosis is rhizoctonia root rot

If leaf malformation = absent and
   stem condition = abnormal and
   stem cankers = below soil line and
   canker lesion color = brown
then diagnosis is rhizoctonia root rot
```

Soybean diseases 3/4

Problems:

- the elicitation of rules from expert is difficult and time consuming
- the rules are not independent
  - e.g. if the leaf condition is normal it is obvious that the leaf malformation should be absent
  - the rules should be carefully checked
- the diagnosis accuracy obtained by the rules alone, without expert assistance, is 72%
- the rules are not able to capture all the expert knowledge *by being told*

Soybean diseases 4/4

Alternative approach:

- the example with diagnosis have been processed by a machine learning algorithm, to generate the classification rules
- the new set of rules got an accuracy of 97.5%, comparable with that of a junior expert

$\Rightarrow$ learning from examples

Wal–Mart and hurricane Frances 1/4

[Provost and Fawcett, 2013]

from a New York Times story from 2004:

> *Hurricane Frances was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons predictive technology.*
>
> *A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes worth of shopper history that is stored in Wal-Mart's data warehouse, she felt that the company could start predicting what's going to happen, instead of waiting for it to happen, as she put it. (Hays, 2004)*

Wal–Mart and hurricane Frances 2/4

Usefullness of data–driven approach in this scenario

- predict that people in the path of the hurricane would buy more bottled water: *too easy*
- project the amount of increase in sales due to the hurricane, to ensure that local Wal-Marts are properly stocked
- discover patterns due to the hurricane that were not obvious
- mining the data could reveal that a particular DVD sold out in the hurricane's path but maybe it sold out that week at Wal-Marts across the country, not just where the hurricane landing was imminent

Wal–Mart and hurricane Frances 3/4

What to do

- examine the huge volume of Wal–Mart data from prior, similar situations (such as Hurricane Charley) to identify unusual local demand for products
- anticipate unusual demand for products and rush stock to the stores ahead of the hurricane's landfall

Wal–Mart and hurricane Frances 4/4

Indeed, that is what happened; the New York Times (Hays, 2004) reported

- the experts mined the data and found that the stores would indeed need certain products and not just the usual flashlights
- *strawberry Pop-Tarts* increase in sales, like seven times their normal sales rate, ahead of a hurricane!

- the pre-hurricane top-selling item was beer

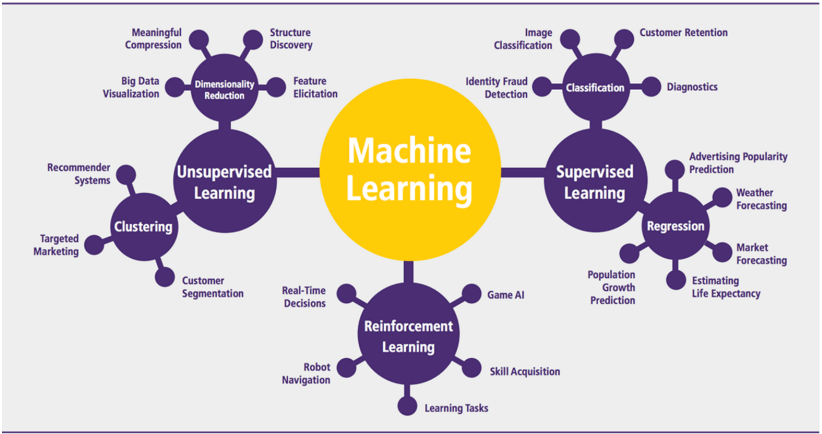
General application areas 1/2

- decision support
  - market analysis
    - how can I improve my business looking at the market basket?
  - risk management
    - should I make a loan to this customer?
  - fraud detection
    - is this credit card transaction fraudulent?
- data analysis
  - text mining
    - e.g. news aggregation
  - social mining
    - analysis of social networks, e.g. twitter, facebook, ...
  - image analysis
    - look at satellite images to find leaks in a pipeline in the desert

General application areas 2/2

- prediction
  - which will be the power requirement in a power supply network in the next day?
- advanced diagnosis and predictive maintenance
  - according to the status of a complex plant, can I forecast some failure, and possibly perform maintenance before the failure?
- ...

Machine Learning at a glance



Source: https://medium.com/analytics-vidhya/which-machine-learning-algorithm-should-you-use-by-problem-type-a53967326566

From Business Problems to Tasks

- data mining is a process
- process has stages
- the process influences the project and the software deployment
- several alternative and well–defined *tasks*
- necessary general knowledge on the base tasks
- a process can easily involve a sequence of different tasks

Tasks 1/4

- *classification* and class *probability estimation*
  - among the customers of a phone company, which are likely to respond to a given offer?
  - can I sort them on the basis of the probability of responding?
- *regression* (value estimation)
  - given a set of numeric attribute values for an individual, estimate the value of another numeric attribute
    - how much will a given customer, whose characteristics are known, use a given service?
  - it is related to *classification*, but the methods are completely different

Tasks 2/4

- *similarity matching* identify similar individuals based on data known about them
  - e.g. which are the companies similar to our best customers? they could be target of our next customer acquisition campaign
  - necessary *similarity measures*
- *clustering* groups individual in a population on the basis of their similarities
  - dna sequences could be clustered in functional groups
- *co–occurrence grouping* attempts to find *associations* between entities according to the transactions in which the appear together
  - (also known as *frequent itemset mining*, *association rule discovery*, *market basket analysis*)
  - what items are commolny purchased together?

Tasks 3/4

- *profiling*
  - also known as *behavior description*
  - What is the typical cell phone usage of this customer segment?
  - the behavior can be described in a complex way over an entire population
  - usually the population is divided in groups of similar behavior
  - useful also to *detect anomalies*
- *link analysis and prediction*, in a world where there exist items and connections (i.e. a *graph*), try to infer missing connections from the existing ones
  - since you and Karen share ten friends, may be you would like to be Karen's friend?

Tasks 4/4

- *data reduction* attempts to take a large set of data and replace it with a reduced one, preserving most of the *important information*
  - the smaller set can be easier to manipulate, or even show more general insights
  - involves *loss of information*
  - looks for the best trade–off between information loss and improved insight
- *causal modeling* understand what events or actions actually *influence* others
  - consider that we use predictive modeling to target advertisements to consumers, and we observe that indeed the targeted consumers purchase at a higher rate subsequent to having been targeted. Was this because the advertisements influenced the consumers to purchase? Or did the predictive models simply do a good job of identifying those consumers who would have purchased anyway?

Supervised vs unsupervised methods 1/3

Example questions?

- Do our population *naturally* fall into different groups?
  - there is no specific *purpose* or *target* for grouping: it should *emerge* by observing the characteristics of the individuals
  - this is an example of *unsupervised* mining
- Can we find groups of customers who have particularly high likelihoods of canceling their service soon after their contracts expire?
  - a specific target is defined: cancelling or not
  - this is an example of *supervised* mining
  - this problem is called *churn analysis*

Supervised vs unsupervised methods 2/3

- The techniques for supervised situations are *substantially different* from those of the unsupervised ones
- Being supervised or unsupervised is a characteristic of the problem and/or the data, it is not a design choice
- Supervised information is usually *added* to the attributes of the individuals

Supervised vs unsupervised methods 3/3

Two main ways to obtain *supervised information*:

- information provided by *experts*
  - e.g. the *soybean disease labels* of the example of page 18
- history
  - e.g. we have an history of the customers who cancelled their service subscription
  - the supervised information is not available run–time, when we must decide what to do
  - later on the history will tell us the value of the unknown attribute which influences our actions
  - we want to learn how to guess the unknown attribute from the known ones

Reinforcement Learning

- target: a sequence of actions which obtains the best result
- learn: a policy
- how: try a policy – get a reward – change the policy
- focus: the overall policy, rather than the single actions

Software – *Open source programming languages with libraries*

An extreme selection

R – a complete, interpreted language, open source, with an infinite suite of specialized libraries; top in user choices, since several years; originally designed for statistical analysis

Python – a complete, interpreted language, open source, with a growing suite of specialized libraries (**scikit-learn**)

Software – *Open source tools with GUI*

An extreme selection

Weka – an open source java collection for the entire mining process

RapidMiner – open source platform which is receiving increasing interest [3]

Knime – open source platform which is receiving increasing interest [3]

3 It has also a commercial version

Software – *Commercial tools*

An extreme selection

SAS – commercial high–end software

IBM SPSS Statistics – commercial software

MATLAB – commercial software

SQL Server, **Oracle**, ... – major DBMS vendors provide their integrated solutions for data warehousing and data mining

What are we going to study?

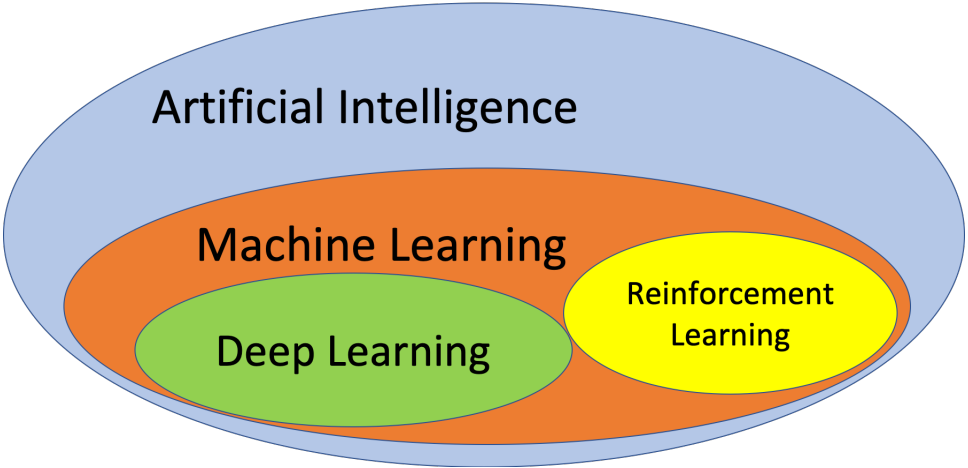
Very general view

- Learning functions
  - Principles
  - Methods
  - Software and examples with Python and scikit-learn
- Classification
  - several specific flavours
  - several algorithms
- Regression
- Clustering
- Associations

What do we include in Machine Learning?

- Methods studied in the Artificial Intelligence area since the '50 of the XX century
- Methods for *data driven decisions*
- Methods which have been known for several decades in *Statistics*

Some relationships



Caveat: the figure represents only logical inclusions, the relative sizes do not reflect the relative importance of the topics

What is a *data set*?

Narrow view

- a set of N individuals
- each individual is described by D values
- in essence it could be seen as a relational table with N rows and D columns

Broader view

- in most of the cases data are not so nicely arranged
- many machine learning techniques require that the dataset is provided as a relational table
  - transformation
- we will see examples of data sets with different formats and of transformations

Bibliography I

- Alpaydin, E. (2014).
  Introduction to machine learning.
  Adaptive computation and machine learning. MIT Press, third edition edition.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009).
  The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition.
  Springer Series in Statistics. Springer.

Bibliography II

- Michalski, R. and Chilausky, R. (1980).
  Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis.
  International Journal of Policy Analysis and Information Systems.
- Provost, F. and Fawcett, T. (2013).
  Data Science for Business.
  O'Reilly.

Bibliography III

- Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.