DMIF, University of Udine

Data Management for Big Data

A Brief Introduction to Data Mining

Andrea Brunello
andrea.brunello@uniud.it

May 2021

What is Data Mining

Basic Definitions

Data $\approx$ stored events/facts.

Information can be considered as the set of concepts, patterns, regularities that are hidden in the data.

Data Mining is the task by which useful, previously unknown information can be extracted from (possibly large) quantitites of data.

- It is a process of abstraction, that leads to the definition of a *model*.

Machine Learning represents the "technical basis" of Data Mining.

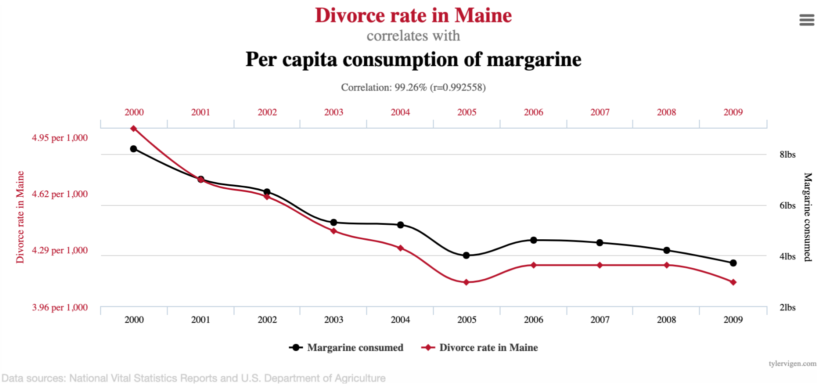
What are Patterns Good for?

The models that capture the patterns can be used to:

- **know**: that some population groups are more likely to buy a specific good
- **explain**: what are the reasons behind customer churn
- **predict**: whether an increase in advertising budget will bring to more sales

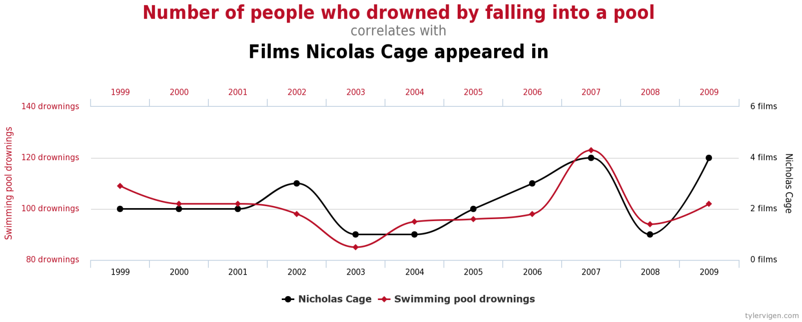
Sometimes, goals may overlap. For instance, think about a model that gives the value of a house based on a series of its characteristics.

Sometimes, the discovered patterns may be trivial, produced by random correlation, or simply wrong.



https://www.tylervigen.com/spurious-correlations

Another strange correlation



Wrap Up

To summarize:

- Data Mining is a task that relies on Machine Learning
- to (semi-)automatically extract
- information, useful patterns
- from (possibly large) quantities of data

Input of the process:

- instances, examples of the concepts that you want to learn

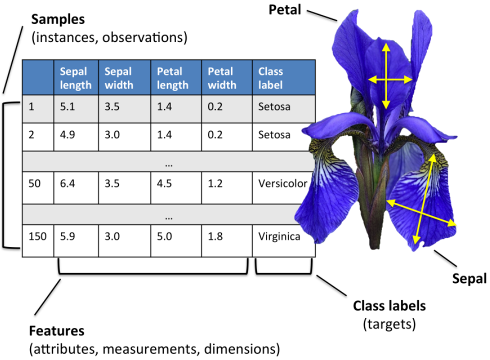
Output of the process:

- predictions
- models

Types of Learning

We will consider tabular datasets, i.e.,

- each row corresponds to an instance
- each column corresponds to a characteristic (feature)
- there may be a colum with a special role (label)



We can identify the following, main, categories of learning:

- Supervised Learning:
  - Classification tasks
  - Regression tasks
- Unsupervised Learning:
  - Association Rule Discovery
  - Clustering
  - ...

Each instance in the dataset is characterized by a set of categorical or numerical features that are used as predictors to determine the value of a specific label.

Given a training dataset of instances, each with feature values $x_1, x_2 \ldots, x_n \in X_1 \times X_2 \times \cdots \times X_n$ and a label value $l \in L$, we want to learn a function $f : X_1 \times X_2 \times \cdots \times X_n \rightarrow L$, such that:

$$f(x_1, \ldots, x_n) = \hat{l} \approx l$$

Function f is encoded into a model, that can be used to predict the value of l for new instances.

In classification tasks, the label l is categorical, thus its domain of values is discrete and finite. For instance, a set of colors, topics, . . .

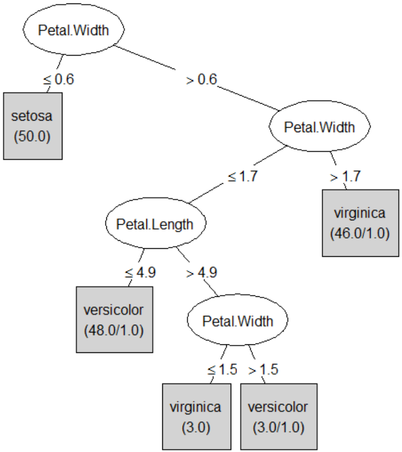
Classical models:

- decision trees and their ensembles
- logistic regression
- naive bayes classifier
- support vector machines

Exemplary tasks:

- text/image/video classification
- credit card fraud detection
- customer churn prediction

J48 decision tree with 98% accuracy on the *Iris* dataset (relying on 10-fold cross-validation).



In regression tasks, the label l is numerical, thus its domain is continuous. For instance, real estate values, probability of a failure, ...

Classical models:

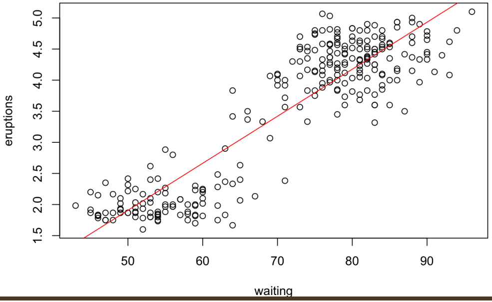
- linear regression
- decision tree ensembles
- support vector regression

Exemplary tasks:

- predictive maintenance
- sentiment analysis
- revenue forecasting

Dataset *faithful*, recordings about the Old Faithful geyser in Yellowstone National Park.

Eruption duration	Waiting time
2.883	55
1.883	54
1.600	52
1.750	47



House Prices Regression Example

$$y = w_0 + w_1 * X_1 + w_2 * x_2 + \dots w_n * x_n$$

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_baseme
0	7129300520	20141013T000000	221900.00	3	1.00	1180	5650	1.00	0	0	...	7	1180	0
1	6414100192	20141209T000000	538000.00	3	2.25	2570	7242	2.00	0	0	...	7	2170	400
2	5631500400	20150225T000000	180000.00	2	1.00	770	10000	1.00	0	0	...	6	770	0
3	2487200875	20141209T000000	604000.00	4	3.00	1960	5000	1.00	0	0	...	7	1050	910
4	1954400510	20150218T000000	510000.00	3	2.00	1680	8080	1.00	0	0	...	8	1680	0

Variable	**Parameter Estimate**	***t*-Statistic**	***p*-Value**
Intercept	$74,915.65	55.744	0.0001
House size (square feet)	36.04	66.756	0.0001
Age of house (in years)	-1,067.32	-16.964	0.0001
Year of sale*	3,349.06	18.505	0.0001
Swimming pool	18,095.18	24.354	0.0001
Subject area 1	2,128.93	3.320	0.0010
Adjusted R^2	0.830		
F-value	1389.979		
p-value	0.0001		

* 0 = 2000; 1 = 2001; 3 = 2003; 4 = 2004

Note: Analysis based on 1,426 sales of single-family residential properties from January 2000 to November 2004.

Unsupervised Learning

We are given a dataset of instances, each one with feature values $x_1, x_2 \ldots, x_n \in X_1 \times X_2 \times \cdots \times X_n$.

There is no label, the goal here is to look for any kind of interesting pattern that can be found among the features.

Still, the output of the process can be considered a model, that encodes such relationships between the features.

The goal is that of discovering "interesting" relations between features in a large dataset.

For instance, the rule $\{onions, potatoes\} \Rightarrow \{burger\}$ found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat.

Such information can be used as the basis for decisions about activities such as promotional pricing or product placements.

Many algorithms to mine association rules have been presented in the literature. Historically, the most important one is *Apriori* (Agrawal and Srikant, 1994).

Clustering is the task of grouping a set of instances in such a way that objects in the same group (cluster) are more similar to each other than to those in other groups.

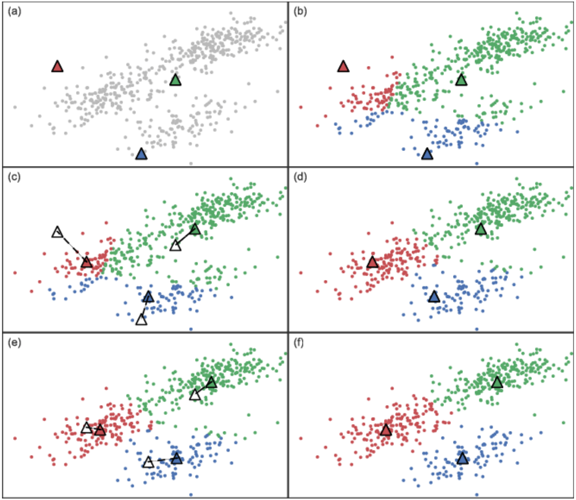
Similarity calculation relies on metrics (e.g., euclidean distance) that are applied on the instances' features.

Many kinds of clustering: soft vs hard, hierarchical vs partitional, …

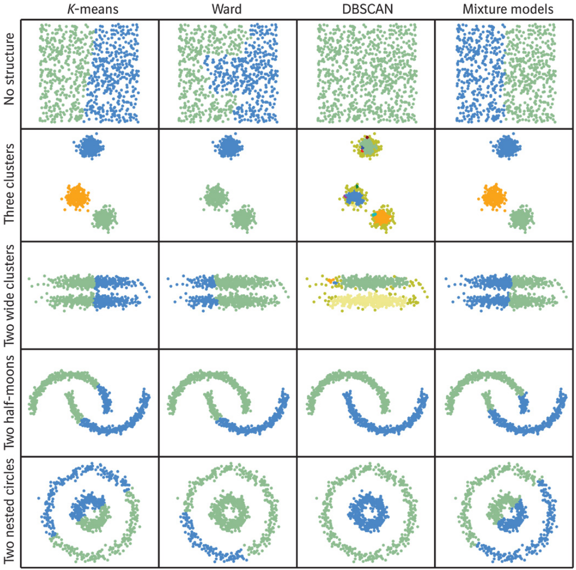
Useful, for instance, to perform customer segmentation.

A popular, partitional clustering algorithm is *K-Means*.

K-Means Example



Clustering is a Hard Task!



References

M. Hall, I. H. Witten, E. Frank, C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th Edition, 2016.

R. Tibshirani, T. Hastie, *An Introduction to Statistical Learning*, 2nd Edition, 2009.

F. Chollet, *Deep Learning with Python*, 2017.