Machine learning: a hands-off introduction

Filippo Biscarini (CNR, Milan, Italy)

filippo.biscarini@cnr.it

Filippo in one slide

- **Roma** (*born*)
- **Perugia** (*MSc degree*)
- **Cork, ICBF** (*Web-design & Database*)
- **Cremona, ANAFI** (*Quantitative Genetics*)
- **Guelph, CGIL** (*Visiting Scientist*)
- **Wageningen, WUR** (*PhD*)
- **Göttingen University** (*post-doctoral researcher*)
- **Lodi, PTP** (*'omics in animals, plants, humans*)
- **Milan** - **CNR** (*tenured researcher*)
- **Cardiff University** (*biostatistician*)
- **Milan** - **CNR** (*senior researcher*)
- **Bruxelles** - **ERC** (*seconded national expert*)
- **Milan** - **CNR** (*senior researcher*)

Overview - 6th edition of this course

Day 1

- Introduction to data mining, ‘omics data and machine learning
- Experimental design
- Advanced R libraries (data.table, tidyverse, tidymodels etc.)

Day 2

- Multivariate data generalities
- Model and variable selection: the machine learning paradigm
- Introduction to supervised learning
- Machine learning for regression problems

Overview

<u>Day 3</u>

- Overfitting and resampling techniques
- Classification problems
- Model complexity tuning
- Unsupervised learning: PCA, Umap, Self-organizing maps

<u>Day 4</u>

- p >> n problems and model regularization (Lasso)
- Workflows with tidymodels
- Bagging and Random Forest for regression and classification
- Multiclass classification with RF
- Slow learning: the boosting approach

Overview

Day 5

- SVM (snippet)
- Advanced data visualization
- Final interactive exercise
- Quiz!

timetable

repo

website

breaks: **long break at around 16:00 (30 min.)**, each day (shorter breaks in between on a case-per-case basis)

It's been a long way to machine learning

- 1925: Ronald Fisher's "*Statistical Methods for Research Workers*" (he later regretted the 0.05 p-value threshold) ⟶ **frequentist statistics**
- **Bayesian** resurgence: 1980s → **MCMC** (1986: Gibbs sampling by Geman & Geman)
- Non-parametric statistics & resampling methods (e.g. empirical tests)
- The **machine** (statistical) **learning** paradigm

A lot of math!

Increasing computer power

Big data

It’s been a long way to machine learning

Supervised learning

- **Linear regression**: late 1800-early 1900 (Francis Galton → Karl Pearson, Ronald Fisher)
- **Logistic regression**: 1940s (Berkson 1944 “Application of the Logistic Function to Bio-Assay”)
- **KNN**: 1950s (Fix & Hodges, 1951)
- **Lasso-penalisation**: late 1980s/1990s (Tibshirani 1996 “Regression Shrinkage and Selection via the lasso”)
- **SVM**: 1990s (Cortes & Vapnik 1995 “Support-Vector Networks”)
- **Boosting**: 1990s/2000s (Schapire 1990 "The Strength of Weak Learnability")
- **Random Forest**: early 2000s (Breiman 2001 “Random Forest”)

It’s been a long way to machine learning

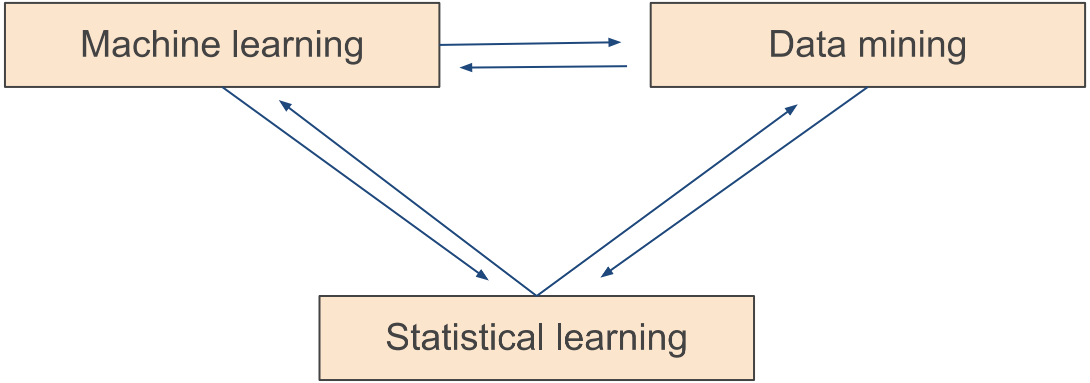
Unsupervised learning

- **PCA**: early 1900s, Karl Pearson
- **k-means clustering**: late 1950s (S. Lloyd, 1957 "Least square quantization in PCM"; published in 1982)
- **anomaly detection**: $p(x) < \varepsilon$ (Edgeworth 1987: “On discordant observations”)
  From 1990’s → ML for anomaly det. (surveyed by Hodge & Austin 2004)
- etc.

Why now?

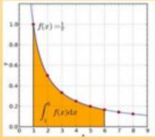
- ideas been about for several decades
- recent novelties:
   i. powerful computers / cloud computing
   ii. optimized algorithms to solve models
   iii. data deluge
   iv. programming frameworks
   v. digital applications

A bit of terminology



- **closely related terms** (very much so)
- **data mining** more for **unsupervised learning** (finding patterns in the data, novel insights) → but uses machine/statistical learning methods
- <u>statistical</u> and <u>machine</u> <u>learning</u> are quasi synonyms (approach from **different directions**: **statistics** or **computer science**)

A bit of terminology



Statistics

+

programming

~~Machine Learning~~

=

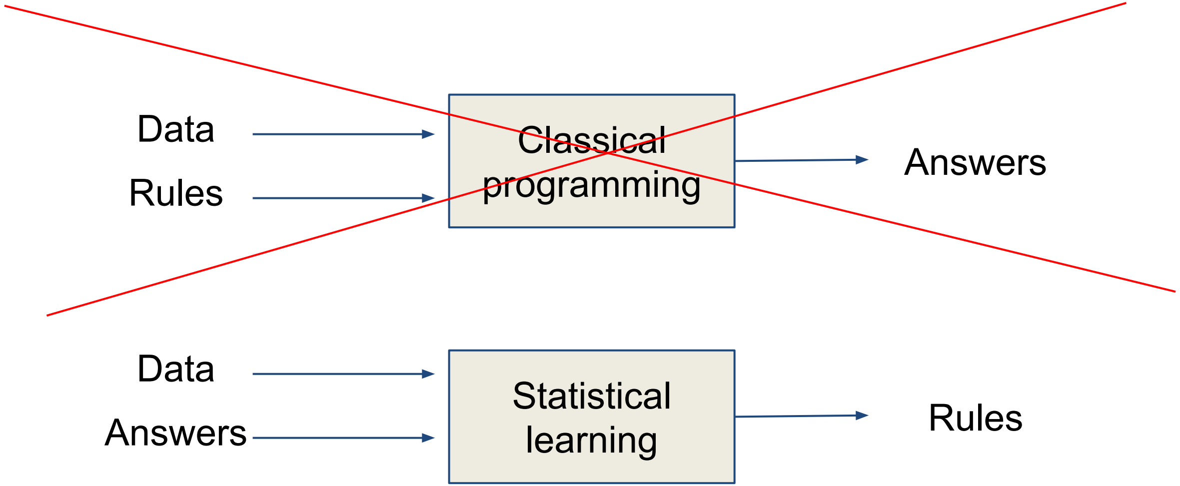
$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Statistical Machine Learning

	Statistics	programming ~~Machine Learning~~	Statistical Machine Learning
Subfield of...	Mathematics	Computer Science (AI)	Statistics & ~~Machine Learning~~ programming
Focus on...	Building models with explicitly programmed instructions	Creating systems that learn from data	Sets of tools for modeling and understanding complex data
Purpose	Inferences; Relationships between variables	Optimization; Prediction accuracy	Building statistical models for prediction; understanding data
Prior assumptions about data	Some knowledge about population usually required	None	Some knowledge about population may be required
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high dimensional data; ML learns from data	Usually applied to high dimensional data
Knowledge overlap	No ML knowledge required	Some stats knowledge usually needed: stats is basis for algorithms	Knowledge of statistics and ML required

What is learning?

What is learning?



Machine learning

- Concerned with the analysis of **complex data** to identify **patterns** that can be used to:
  - **predict** the outcomes of elections
  - **identify** and filter spam messages from e-mail
  - **foresee** criminal activity
  - automate traffic signals according to road conditions
  - produce financial estimates of storms and natural disasters
  - **identify** disease outbreaks (e.g. SoundsTalk)
  - **predict** when patients get sick
  - determine credit worthiness
  - target advertising to specific types of consumers
  - and many more ...

Machine learning

- Concerned with the analysis of **complex data** to identify patterns that can be used to:
    - **predict** the outcomes of elections
    - **identify** and filter spam messages from e-mail
    - **foresee** criminal activity
    - automate traffic signals according to road conditions
    - produce financial estimates of storms and natural disasters
    - **identify** disease outbreaks (e.g. SoundsTalk)
    - **predict** when patients get sick
    - determine credit worthiness
    - target advertising to specific types of consumers
    - and many more ...

many terms related to predictions (one of the main tasks in ML)

Machine learning - between legend and reality

1. US retailer used machine learning to analyse consumers data and identify pregnant women (customers) and predict due date
2. based on this, targeted promotional offers were sent via mail (e.g. maternity clothes, baby clothes, baby food etc.)
3. father reacted angrily to her daughter receiving such offers for maternity items
4. manger from the retailer called to apologise for the error in their ML system
5. ultimately, the father returned the apologies because his daughter was indeed pregnant

Machine learning - between legend and reality

1. US retailer used machine learning to analyse consumers data and identify pregnant women (customers) and predict due date
2. based on this, targeted promotional offers were sent via mail (e.g. maternity clothes, baby clothes, baby food etc.)
3. father reacted angrily to her daughter receiving such offers for maternity items
4. manger from the retailer called to apologise for the error in their ML system
5. ultimately, the father returned the apologies because his daughter was indeed pregnant

AI-generated news

- example 1
- example 2

May be true or not, yet:

- retailers indeed use ML to analyse purchase data
- ML can be surprisingly effective (know us better than ourselves)
- ethical implications! (“don’t be evil!” @google)

ML - beware of unexpected results!

1. ~2015, Amazon
2. Tested a ML algorithm to automatically and quickly screen CV for recruitment
3. Biased towards discriminating against applications from women
4. The algorithm was using data like team sports vs individual sports, chess playing etc. which were partially correlated with sex
5. **Emerging biases**: not by design, but emerging from high order non-linearities in the algorithm
6. Risk of using ML without understanding / controlling well what goes on

(Amazon never used this system, stopped at testing)

Machine learning - definition

- A. Samuel (1959): giving computers the **ability to learn without being explicitly programmed** (he coined the term ‘machine learning’)
- T. Mitchell (1998): a computer program **learns** from **experience E** with respect to **task T** with **performance P**, **if P on T improves with E**

Machine learning - definition: a task for you!

Which is **E**, **T**, **P**?

- diagnosing patients as sick or healthy
- watching the clinician making the diagnosis (sick/healthy)
- number of patients correctly diagnosed

Data (knowledge) representation

Source: http://collections.lacma.org/node/239578

- not a real pipe
  (picture of a pipe)
- idea of a pipe
  (concept)
- actual pipe (object)

Abstract connections, knowledge representation

Data (knowledge) representation

Source: http://collections.lacma.org/node/239578

- not a real pipe (picture of a pipe)
- idea of a pipe (concept)
- actual pipe (object)

- raw data (0s, 1s in memory)
- abstraction (what the data mean)
- reality (natural phenomenon)

Abstract connections, knowledge representation

Data (knowledge) representation → learning

- not a real pipe (picture of a pipe)
- idea of a pipe (concept)
- actual pipe (object)

- raw data (0s, 1s in memory)
- abstraction (what the data mean)
- reality (natural phenomenon)

Abstract connections, knowledge representation

model

generalization

(we want the machine to be able to learn from experience and generalise to new cases, just like we humans do)

Data representation: example from genomics

Let's work this out together!

<u>Genomic variants for diabetes</u>

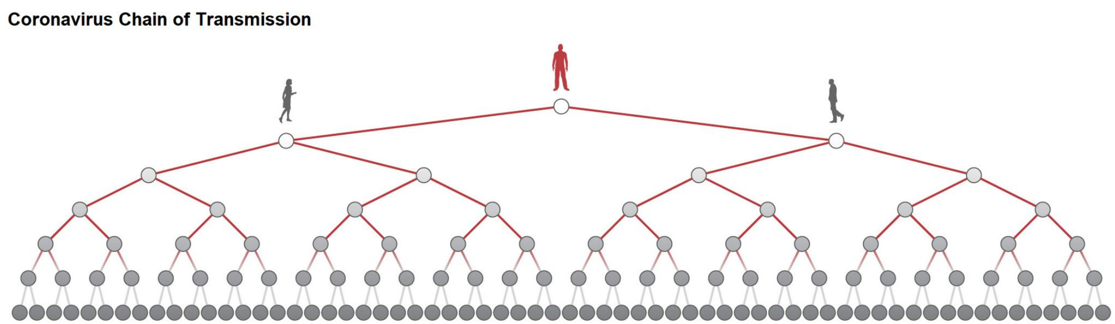
- raw data:
  - 0s and 1s stored in memory
- what the data mean (data representation):
  -
  -
- natural phenomenon (what we want to study):
  -
  -
  -
- model:
  -
  -

Data representation: example from genomics

Genomic variants for diabetes

- raw data:
    - 0s and 1s stored in memory
- what the data mean (data representation):
    - some 0/1 are genomic variants, others are disease labels
    - n. copies minor allele, presence/absence etc.
- natural phenomenon (what we want to study):
    - genetic predisposition to diabetes
    - predict diabetes based on genome
    - identify genomic variants linked to diabetes
- model:
    - knowledge that genes (co)determine phenotypes
    - $P(\text{diabetes}|x) = \text{variant}_1 + \text{variant}_2 + \ldots + \text{variant}_m + e$

Back-of-the-envelope machine learning



[from: The New York Times]

Rule: ?

day	N. cases
1	1
2	2
3	4
4	8
5	16
…	…
9	?