Machine Learning: general concepts and applications

Paolo Frasconi
Università degli Studi di Firenze, Italy
`http://ai.dinfo.unifi.it/paolo/`

Workshop — Firenze, May 29, 2019

Part 1

Introduction to Machine Learning

What is machine learning

- The discipline is roughly as old as computer science itself but with older roots in statistics, mathematics, and physics

- Many facets exist today, in general a collection of algorithms, methods, and analysis techniques where data plays a central role

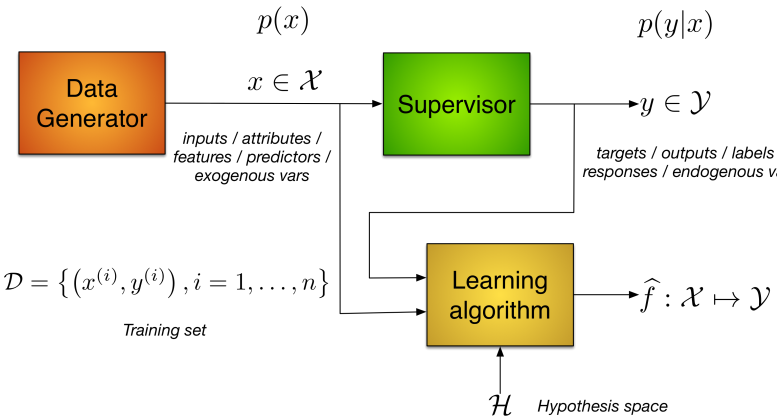
- The ability of a computer program to *aquire knowledge from data and improve performance through experience* has been one of the core issues of Artificial Intelligence since ever

- Today machine learning is the fundamental reason behind the success of AI and many people talk about AI when they really mean machine learning

- Some tasks involve computing a function on data points to obtain useful predictions
    - Object recognition in images
    - Speech recognition
    - Text categorization
    - Translation (e.g. from Czech to Dutch)
    - Molecule activity prediction
    - Protein structure prediction
- Not enough domain knowledge to formalize the task and enable traditional algorithmic solutions

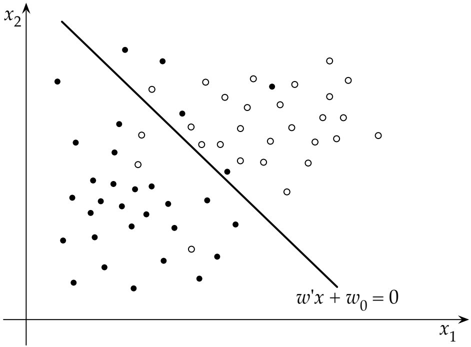
- In supervised learning you collect examples of instances paired with the corresponding solution, e.g.
  - This is a picture of a lemon tree
  - In this utterance the speakers says "Good morning"
  - This molecule is confirmed active against HIV
  - Dobré ráno $\mapsto$ Goede morgen
- A machine learning algorithm will take as input many such pairs and will return a function

Supervised Learning (standard)



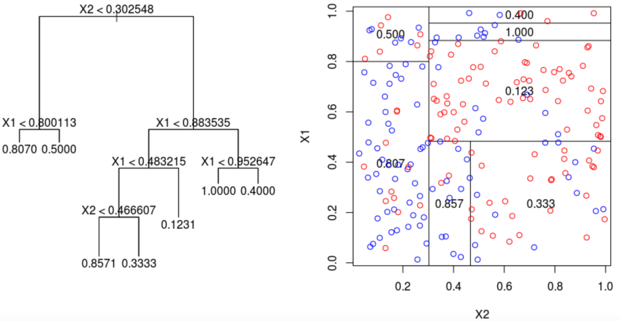
- (Q1) how do we define $\mathcal{H}$?
- (Q2) how fo we compute $\widehat{f} = A(\mathcal{D}, \mathcal{H})$?

- Long history, e.g., Fisher 1936, McCulloch & Pitts 1943, Berkson 1944, Rosemblatt 1957
- Heavy criticisms in 1969 by Seymour Papert and Marvin Minsky, end of early connectionism, prelude to the AI winter of the 1970s



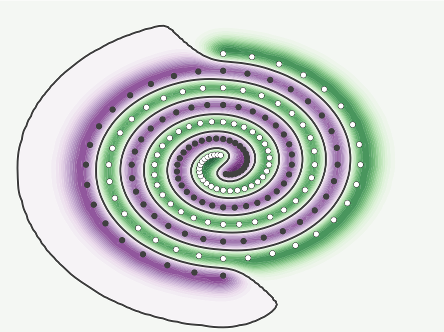
- Also have a long history, e.g., Morgan & Sonquist 1963, Hunt 1966, Breiman et al. 1984, Quinlan 1986

- Some very effective techniques (e.g. random forests, gradient boosting, etc) still based on this idea



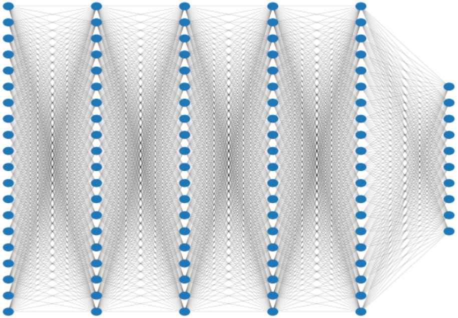
Picture by James Le

- See e.g. Mercer 1909, Kimeldorf & Wahba 1970, Cortes & Vapnik 1995
- Praised for over a decade because searching boils down to solving a convex optimization problem



- See e.g. Turing 1948, Fukushima 1980, Rumelhart, Hinton, & Williams 1985, Hinton et al. 2006

- Deep learning (where you actually use any differentiable computational graph) responsible for the AI explosion of this decade



(Q2): Solving the standard supervised learning problem

- Need to define a quality measure of the generated function
- Typically this is done through a loss function $L : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$ where $L(f(x), y)$ is the cost incurred when you predict $f(x)$ and you should have predicted y
- Once we have L (not always obvious) the problem looks like optimization:

$$\widehat{f}_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathbb{E}\, L(f(x), y)$$

- But unfortunately this cannot be done: the objective is not observed!

- Be happy with the "data distribution" (and regularize)

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \sum_{(x,y) \in \mathcal{D}} L(f(x), y)$$

$$\text{Subject to} \qquad \Omega[f] < R$$

- L needs to be appropriate for the task at hand (but tractability also needs to be taken into account)
- Choice (and size) of the hypothesis space also very important
- Hyperparameter R trades complexity for overfitting
- No universally better learning algorithm (no free lunch theorem)

What's new today

Many better algorithms and methods but perhaps most importantly these few things:

1. Powerful hardware (2 petaFLOPS for $400k, compare to .28 of top supercomputer Blue Gene when deep learning started, $290 millions)

2. Vast software codebase

3. Large datasets

4. Really strong industrial interest