WHAT IS MACHINE LEARNING?

(and what is it not?)

Charles-Pierre Astolfi, cp.astolfi@crans.org

POP QUIZ!

Machine Learning…

- aims at replicating the brain.
- tries to find patterns and correlations.
- is not used in industry yet.
- is a black art more than a science.
- can make scientific discoveries.
- has no ethical ramifications yet.
- can help you find your life partner.
- saves 5 millions lives per year.

A DAY IN MY LIFE

WHAT IS MACHINE LEARNING?

~~Science~~ Black art whose goal is to:

- Classify data.
- Capture characteristics from empirical data.
- Generate data "in the style of" what has been seen.
- Learn to take decisions based on the past course of actions.

Classification (and ranking)

Clustering

Regression

Reinforcement learning

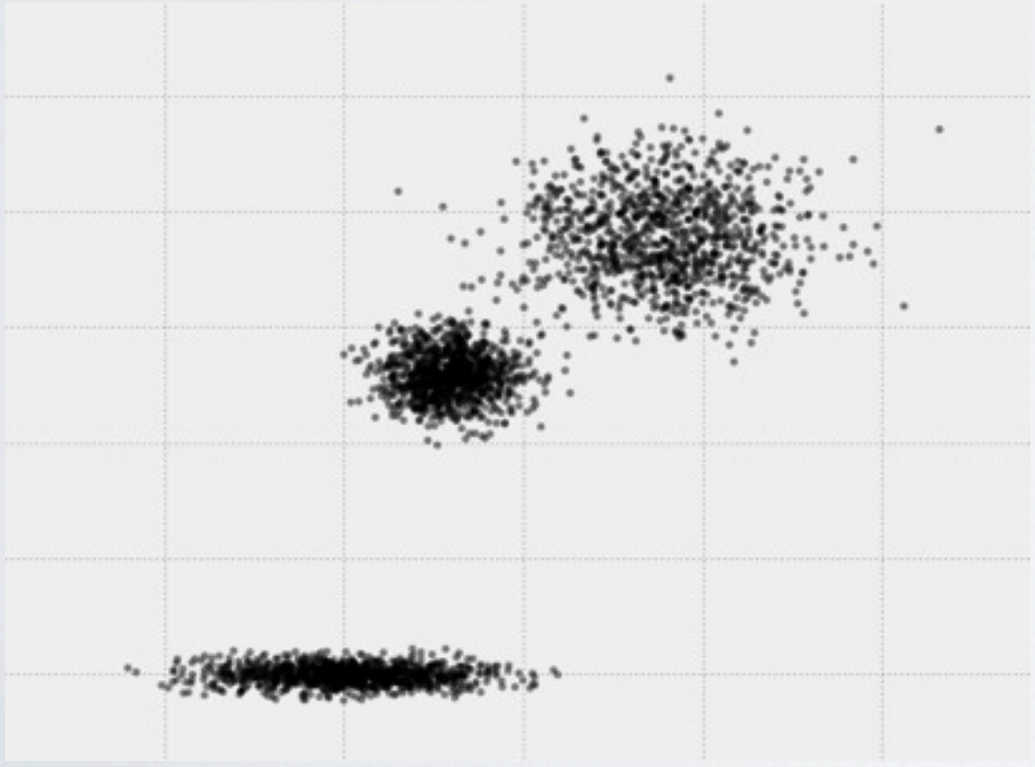
CLASSIFICATION
(SUPERVISED LEARNING)

Input	Output
Age + Year of operation + Number of axillary nodes detected	0 if the patient died within 5 years 1 if the patient survived 5 years or longer

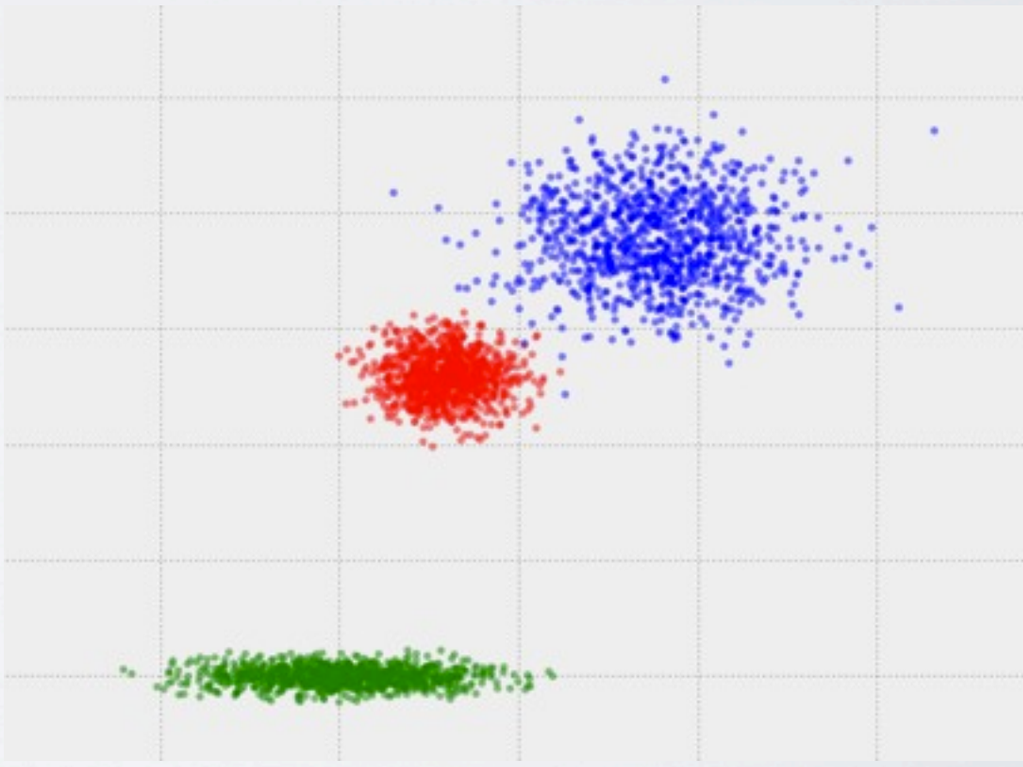
CLUSTERING
(UNSUPERVISED LEARNING)

Like classification, but the labels are unknown.

Input



Output



UNSUPERVISED LEARNING

State of the art:

- Andrew Ng & al. trained an unsupervised large-scale (16,000 cores) neural network

- This is a neuron that detects faces

- Precision: 19% on 22000 classes.

REGRESSION

- Like classification, but one has to predict a value rather than a label.

- E.g.: given some statistics about crime in a neighborhood, predict the number of crimes next year.

REINFORCEMENT LEARNING

- Predictions are decisions !
- Demo: Pendulum swing up learning

SUMMARY

If given...	My predictions are...	Then I'm doing...
Vector	(Known) finite set of labels	Classification
	(Unknown) finite set of labels	Clustering
	Real value	Regression
Past events	Actions	Reinforcement learning

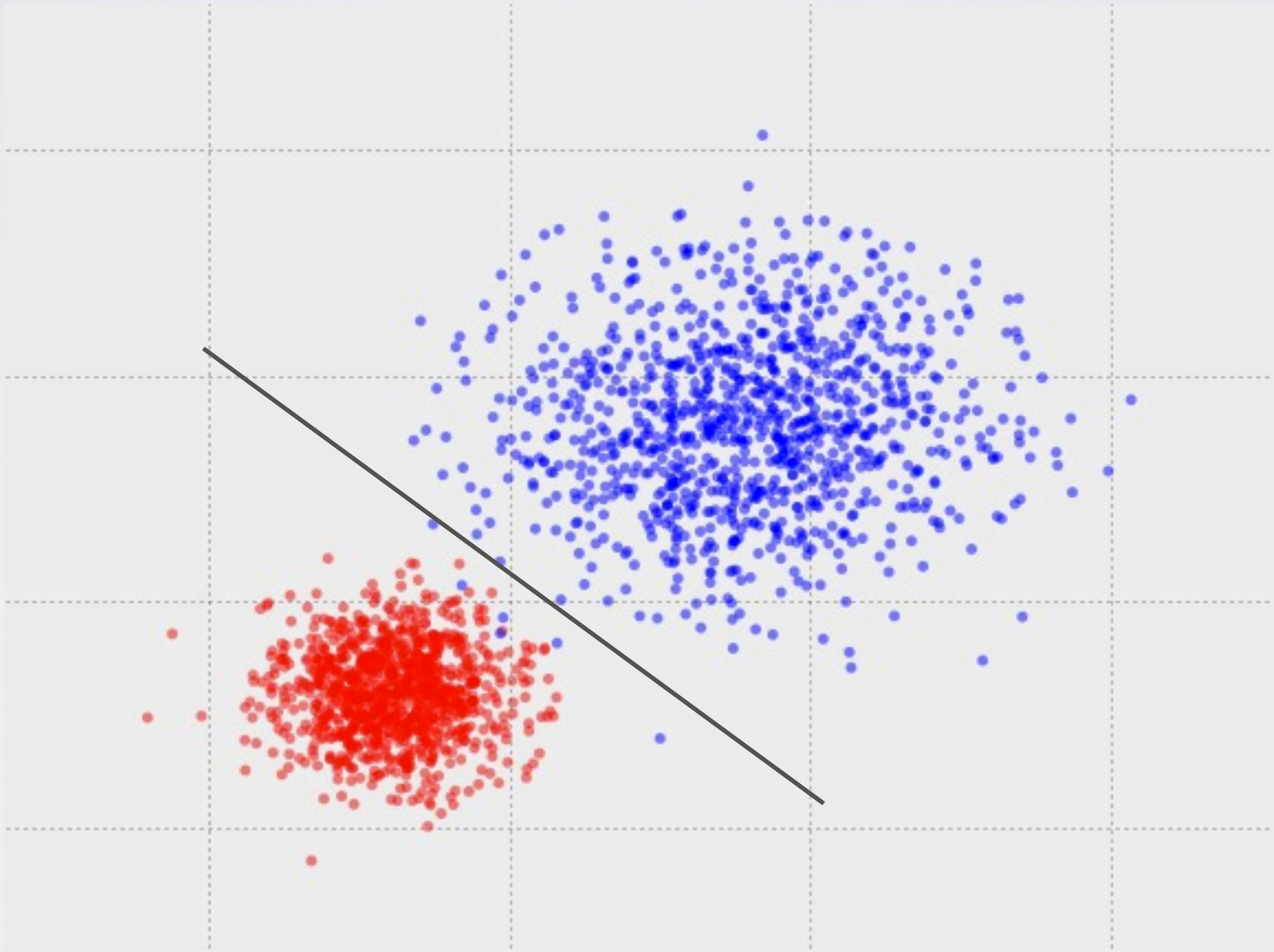
RIDGE

Given $\mathbf{X} \in \mathbb{R}^n \times \mathbb{R}^m$ (training data) and $Y \in \mathbb{R}^n$ (outcomes),

Find w that satisfies:

$$\min_w \sum_{i=1}^{n} (Xw - Y)_i^2 + \alpha \sum_{i=1}^{m} w_i^2$$

RIDGE RESULTS



ML APPLICATIONS

- Finding conservation equations for the double pendulum (a chaotic dynamic system!)

- Web search

- Match people (meetic, eharmony and okcupid hire a lot of ML people!)

- Discriminate gender on Twitter
  Most common words for females:
  "!, love, :), haha, so"
  For males: "Goog, googl, google, http"

ML IN INDUSTRY

- Automated mining: Rio Tinto and Nicta
- Web search: Google
- Ad selection: Google, Facebook
- Medical research
- Machine Vision: Driverless cars, animal census via drones
- Speech Recognition: Help desks, banking.
- Intelligence agencies !
- Snail mail: address recognition
- Sentiment mining: who's thinking what?
- Recommender systems: Netflix (1M$ prize), Air France
- Automated translation
- Rare event detection (people fighting on CCTV)
- Stock prediction (finance or logistics)
- Energy consumption prediction
- Weather forecasting (that one's hard)
- Signal analysis
- Behavior analysis
- Job finding

ML DRAWBACKS

- No silver bullet. (SVM? Ridge? Lasso? Random Forests?)
- NP-Hardness is often an issue.
- Even for heuristics, complexity is usually more than linear.
- It's hard to get clean data.
- It's hard to select the good features.
- It's often hard to understand your predictive model.
- It's next to impossible to ensure statistical significance.
- There's this thing we call the "Curse of dimensionality"...

I THINK WE'RE DONE HERE.

Questions?