Statistics and learning

An introduction to Machine Learning

Emmanuel Rachelson and Matthieu Vignes

ISAE SupAero

Friday 11th January 2013

Machine Learning

Let's talk about Machine Learning!

Keywords?

A few examples

- Given 20 years of clinical data, will this patient have a second heart attack in the next 5 years?

A few examples

- Given 20 years of clinical data, will this patient have a second heart attack in the next 5 years?
- What price for this stock, 6 months from now?

A few examples

- Given 20 years of clinical data, will this patient have a second heart attack in the next 5 years?
- What price for this stock, 6 months from now?
- Is this handwritten number a 7?

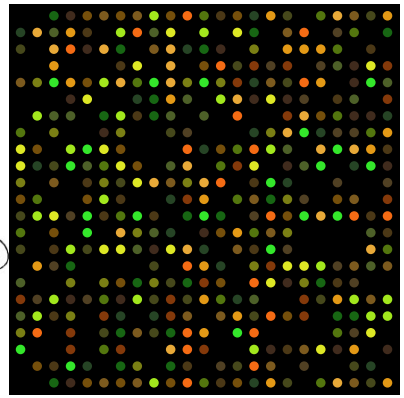
A few examples

- Given 20 years of clinical data, will this patient have a second heart attack in the next 5 years?
- What price for this stock, 6 months from now?
- Is this handwritten number a 7?
- Is this e-mail a spam?

Enlarge your thesis!

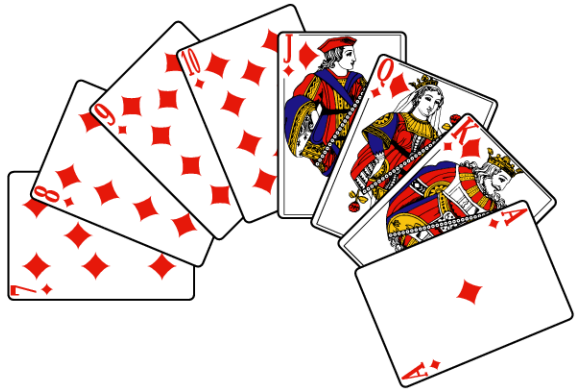
A few examples

- Given 20 years of clinical data, will this patient have a second heart attack in the next 5 years?
- What price for this stock, 6 months from now?
- Is this handwritten number a 7?
- Is this e-mail a spam?
- Can I cluster together different customers? words? genes?



A few examples

- Given 20 years of clinical data, will this patient have a second heart attack in the next 5 years?
- What price for this stock, 6 months from now?
- Is this handwritten number a 7?
- Is this e-mail a spam?
- Can I cluster together different customers? words? genes?
- What is the best strategy when playing Counter Strike? or "coinche"?



A (tentative) taxonomy

Different kinds of learning tasks:

Task	Data: based on...	Target: learn...
▶ Supervized	$\mathcal{T} = \{(x_i, y_i)\}_{i=1..n}$	$f(x) = y$
▶ Unsupervized	$\mathcal{T} = \{x_i\}_{i=1..n}$	$x \in X_k$
▶ Reinforcement	$\mathcal{T} = \{(x_i, u_i, r_i, x'_i)\}_{i=1..n}$	$\pi(x) = u \ / \ \max \sum r_t$

A (tentative) taxonomy

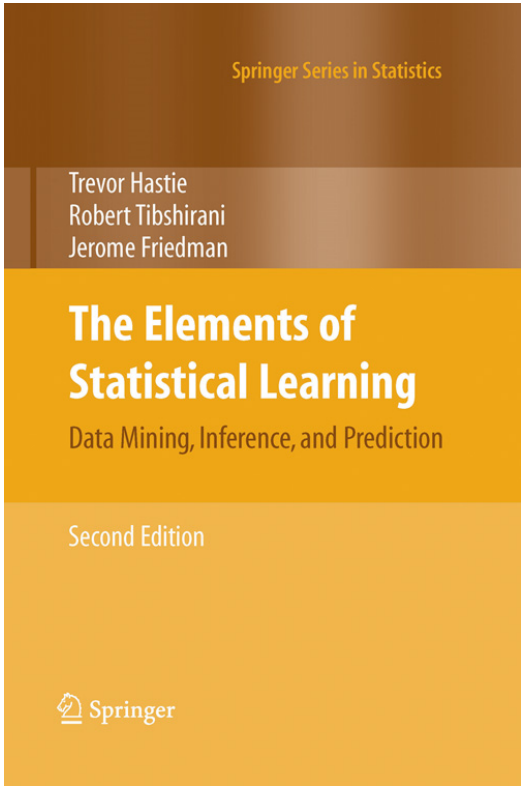
Different kinds of learning tasks:

Task	Data: based on...	Target: learn...
▸ Supervized	$\mathcal{T} = \{(x_i, y_i)\}_{i=1..n}$	$f(x) = y$
▸ Unsupervized	$\mathcal{T} = \{x_i\}_{i=1..n}$	$x \in X_k$
▸ Reinforcement	$\mathcal{T} = \{(x_i, u_i, r_i, x'_i)\}_{i=1..n}$	$\pi(x) = u \ / \ \max \sum r_t$

Different kinds of learning contexts:

- Offline, batch, non-interactive: all samples are given at once.
- Online, incremental: samples arrive one after the other.
- Active: the algorithm asks for the next sample.

Reference textbook



The Elements of Statistical Learning, second edition.
Trevor Hastie, Robert Tibshirani, Jerome Friedman.
Springer series in Statistics, 2009.

Supervized Learning – vocabulary

inputs	outputs
independent variables	dependent variables
predictors	responses
features	targets
X (random variables)	Y (random variables)
x_i (observation of X)	y_i (observation of X)

Outputs

Nature of outputs:

- Quantitative or ordered: $y_i \in \mathbb{R}$
  $\rightarrow$ Regression task.
- Qualitative or unordered: $y_i \in \{0; 1\}$
  $\rightarrow$ Classification task.

In both cases: fitting a function $f(x) = y$ to the data.

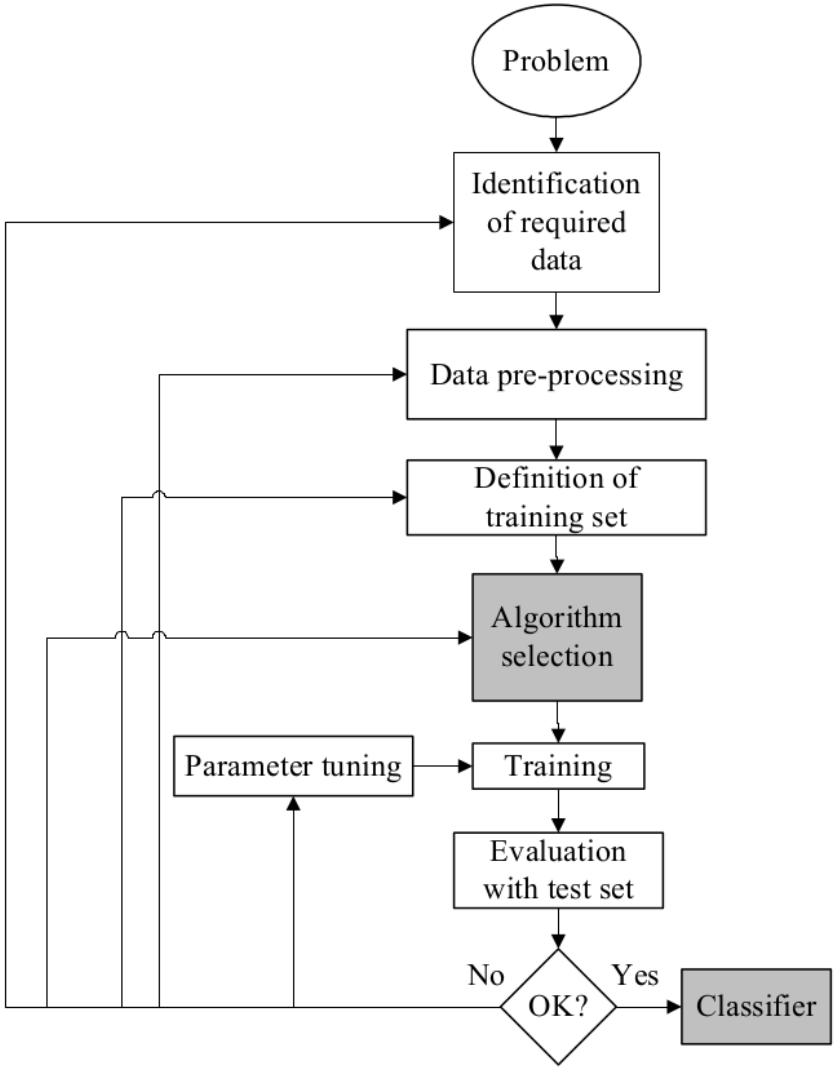
Questions:

- $y_i \in \mathbb{N}$? $y_i \in \{\text{red}, \text{blue}, \text{green}, \text{yellow}\}$? $y_i \in \mathbb{R}^N$?
- What about noise; still fitting $f(x) = y$?
- What about generalization? Overfitting? Overspecialization?

Supervized learning problem

Given the value of X,
make a good prediction $\hat{Y}$ of the dependent variable Y,
given a *training set* of samples $\mathcal{T} = \{(x_i, y_i)\}_{i=1..n}$.

The process of Supervized Learning



From **Supervized Machine Learning: A Review of Classification Techniques**, S. B. Kotsiantis, *Informatica*, 31:249–268, 2007.

Focus of the next classes

An introduction to:

- Naive Bayes classification
- Support vector machines and kernel methods,
- Neural networks,
- Decision trees and Boosting,
- Markov Chain Monte Carlo (MCMC) model selection.

Examples of other, uncovered topics in supervised learning and keywords:

- Wavelets,
- Bias-variance tradeoff,
- Cross-validation,
- L1 regularization and the LASSO,
- Vapnik-Chernovenkis dimension,
- Bagging,
- Nearest-neighbour methods,
- Random forests,
- and much more!

Welcome to the wonderful world of Machine Learning!