Machine Learning 1
Introduction

Master 1 Computer Science

Yohann de Castro, Aurélien Garivier

2019-2020

UNIVERSITÉ DE LYON | ENS DE LYON

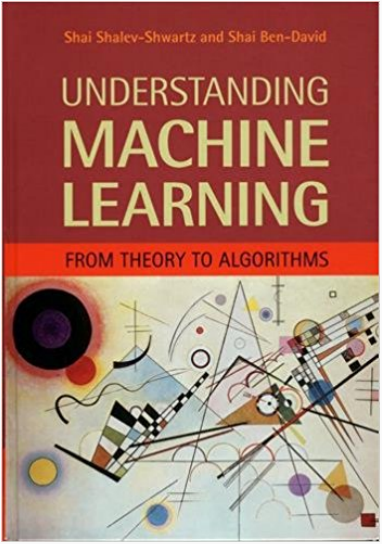
Table of contents

Before we start

Outline

- 1. 01.17 Introduction to ML, Statistics 101
- 2. 01.24 Clustering
- 3. 01.31 Dimensionality Reduction: PCA, random projections
- 4. 02.07 Supervised Learning, Nearest Neighbors
- 5. 02.14 Bias-Variance Tradeoff, CART
- 6. 02.21 Ensemble methods: Boosting, Bagging, Random Forests
- 7. 02.28 Empirical risk minimization, Linear Separators
-     03.06 holidays
- 8. 03.13 Structural Risk Minimization, Kernels, Regularization
- 9. 03.20 Neural networks and stochastic gradient descent
- 10. 03.27 Online Learning
- 11. 04.03 Revisions
- 12. 04.10 free
- 13. 04.17 Final Exam

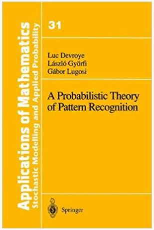
Reference textbook

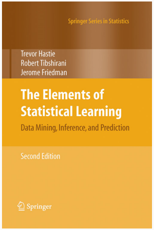


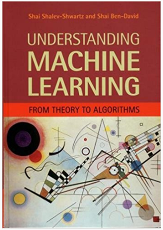
General introduction to Machine Learning theory, by two leading researchers of the field.

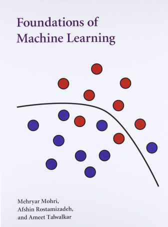
Covers a good part of the content of this course (other references will be provided for specific topics).

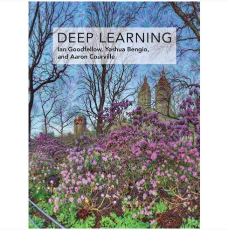
Additional References











WikiStat

scikit learn

Evaluation

- 50% final exam
- 50% exercises and project; bonus for scribes

Project: a "challenge-like" data science problem (to be presented later).
By groups of 4-5.

What is Machine Learning?

Why Machine Learning?

Actualité

Yann LeCun, Geoffrey Hinton et Yoshua Bengio reçoivent le prix Turing

Par **Stephane Nachez** - 27 mars 2019

SHARE

SPECIAL VIEWPOINTS

Machine Learning for Science: State of the Art and Future Prospects

Eric Mjolsness*, Dennis DeCoste
+ See all authors and affiliations

Science 14 Sep 2001:
Vol. 293, Issue 5537, pp. 2051-2055
DOI: 10.1126/science.293.5537.2051

Article Figures & Data Info & Metrics eLetters PDF

Abstract

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

LE MACHINE LEARNING PROVOQUE UNE CRISE DANS LE DOMAINE DE LA SCIENCE

Bastien L 19 février 2019 Analytics, Data Analytics, Intelligence artificielle 1 commentaire

Le Machine Learning est en train de provoquer une grave crise de reproductibilité dans le domaine de la science. C'est ce qu'affirme la statisticienne Genevera Allen de la Rice University dans le cadre de la conférence AAAS Annual Meeting.

De plus en plus de chercheurs utilisent le Machine Learning pour analyser des données et y détecter des tendances. Cependant, dans le cadre de la conférence scientifique AAAS Annual Meeting, la statisticienne Genevera Allen de la Rice University a tenu à tirer la sonnette d'alarme. Selon elle, le Machine Learning est en passe de provoquer **une crise de reproductibilité dans le domaine de la science**.

PUBLIC RELEASE: 15-FEB-2019

Can we trust scientific discoveries made using machine learning?

Rice U. expert: Key is creating ML systems that question their own predictions

RICE UNIVERSITY

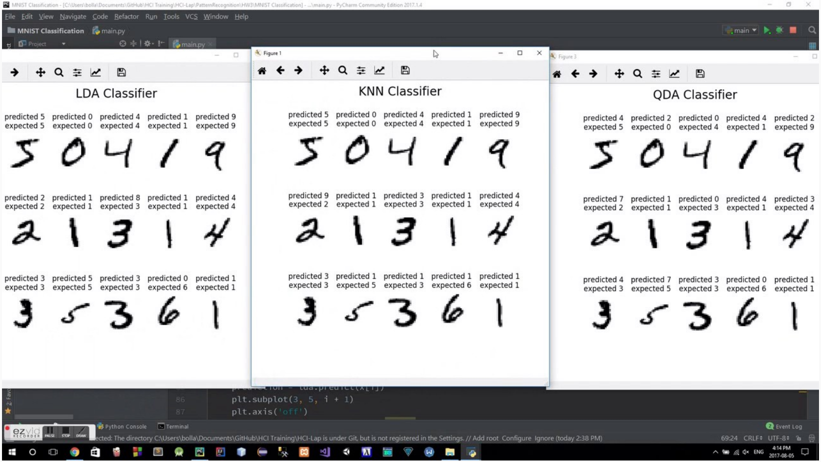
What is Machine Learning?

- Algorithms operate by building a model from **example** inputs in order to make data-driven **predictions or decisions**...
- ...rather than following strictly static program instructions: useful when designing and programming explicit algorithms is unfeasible or poorly efficient.

Within Artificial Intelligence

- evolved from the study of pattern recognition and computational learning theory in artificial intelligence.
- AI: emulate cognitive capabilities of humans
(big data: humans learn from abundant and diverse sources of data).
- a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

Example: MNIST dataset



Machine Learning (ML): Definition

Arthur Samuel (1959)

Field of study that gives computers the ability to learn without being explicitly programmed

Tom M. Mitchell (1997)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

Machine Learning: Typical Problems

- spam filtering, text classification
- optical character recognition (OCR)
- search engines
- recommendation platforms
- speach recognition software
- computer vision
- bio-informatics, DNA analysis, medicine
- etc.

For each of this task, it is possible but very inefficient to write an explicit program reaching the prescribed goal.

It proves much more succesful to have a machine infer what the good decision rules are.

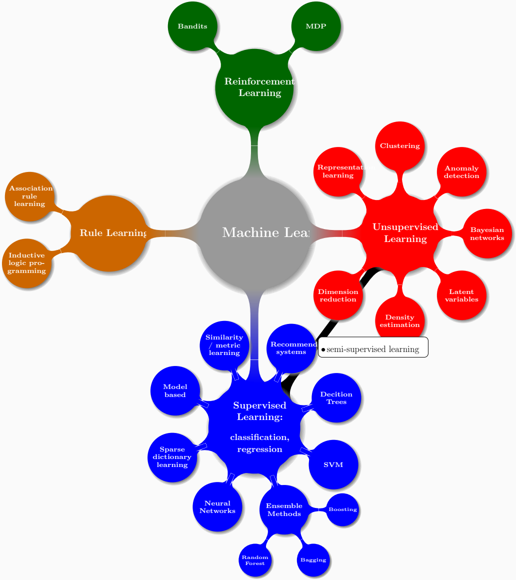
What is Statistical Learning?

$=$ Machine Learning using statistics-inspired tools and guarantees

- Importance of **probability**- and **statistics**-based methods
  $\rightarrow$ **Data Science** (Michael Jordan)
- **Computational Statistics**: focuses in prediction-making through the use of computers together with statistical models (ex: Bayesian methods).
- **Data Mining** (unsupervised learning) focuses more on exploratory data analysis: discovery of (previously) unknown properties in the data. This is the analysis step of Knowledge Discovery in Databases.
- Machine Learning has more **operational** goals
  Ex: ~~consistency~~ $\rightarrow$ oracle inequalities
  Models (if any) are *instrumental*.
  ML more focused on *correlation*, less on *causality* (now changing).
- Strong ties to **Mathematical Optimization**, which furnishes methods, theory and application domains to the field

The Learning Models

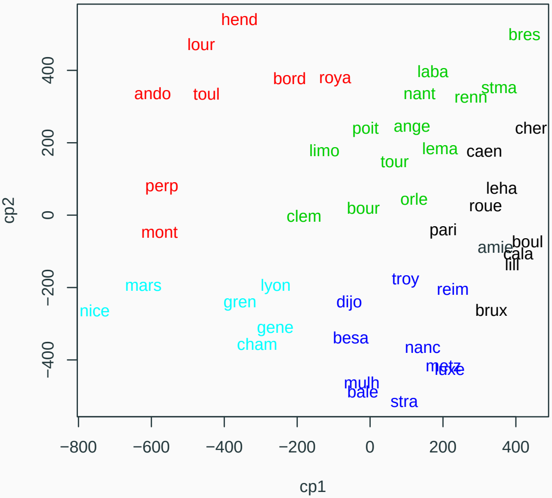
What ML is composed of



Unsupervised Learning

- (many) observations on (many) individuals
- need to have a simplified, structured overview of the data
- *taxonomy*: untargeted search for *homogeneous clusters* emerging from the data
- Examples:
  - customer segmentation
  - image analysis (recognizing different zones)
  - exploration of data

Example: representing the climate of cities



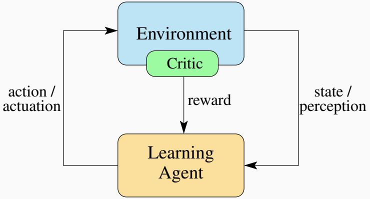
Supervised Learning

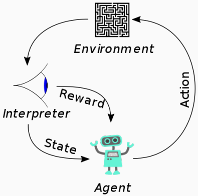
- Observations = pairs (X_i, Y_i)
- Goal = learn to *predict* Y_i *given* X_i
- Regression (when Y is continuous)
- Classification (when Y is discrete)

Examples:

- Spam filtering / text categorization
- Image recoginition
- Credit risk ranking

Reinforcement Learning

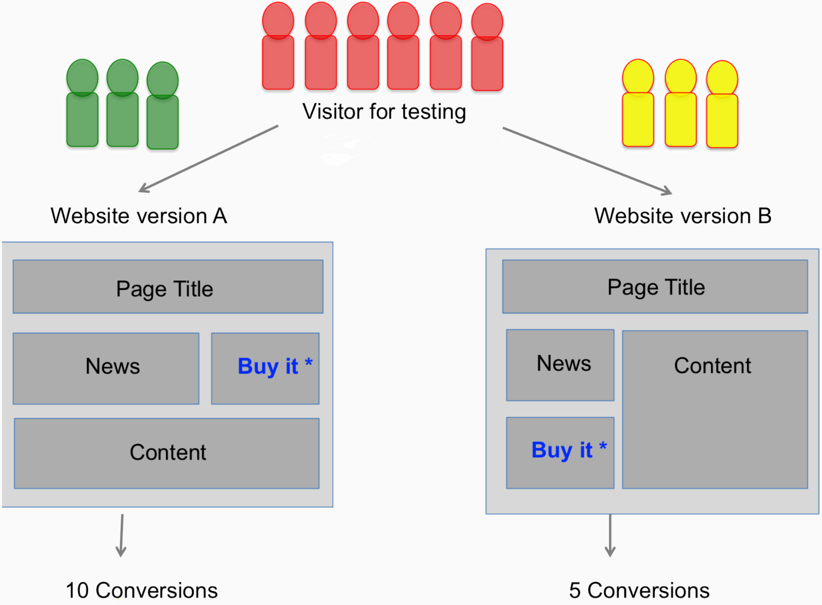




[Src: https://en.wikipedia.org/wiki/Reinforcement_learning]

- area of machine learning inspired by behaviourist psychology
- how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.
- Model: random system (typically : Markov Decision Process)
    - agent
    - state
    - actions
    - rewards
- sometimes called approximate dynamic programming, or neuro-dynamic programming

Example: A/B testing



Machine Learning Methodology

ML Data

n-by-p matrix X

- n examples = points of observations
- p features = characteristics measured for each example

Questions to consider:

- Are the features centered?
- Are the features normalized? bounded?

In `scikitlearn`, all methods expect a 2D array of shape (m, p) often called

```
X (n_samples, n_features)
```

Data repositories

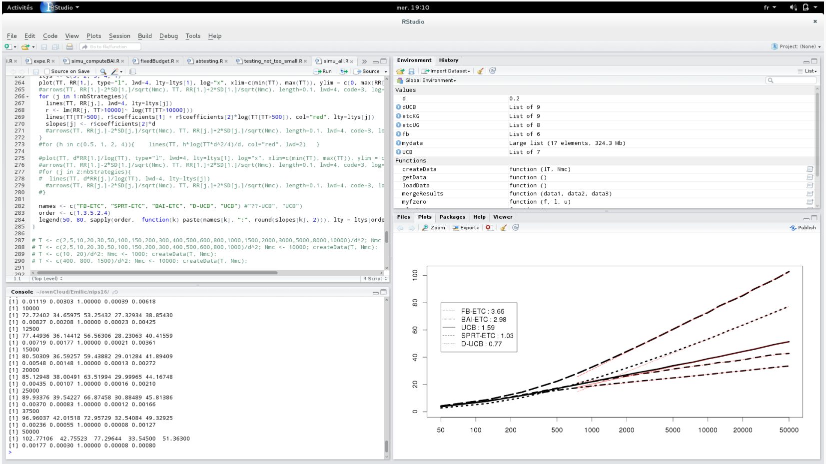
- Inside R: package `datasets`
- Inside scikitlearn: package `sklearn.datasets`
- UCI Machine Learning Repository
- Challenges: Kaggle, etc.

The big steps of data analysis

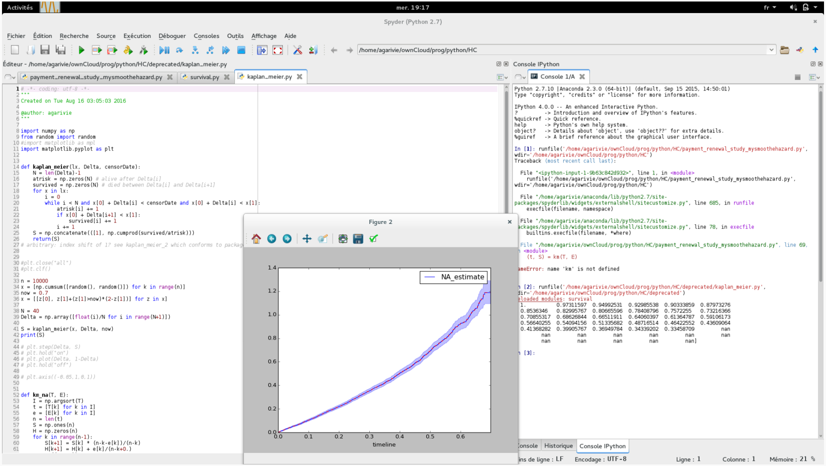
1. Extracting the data to expected format
2. Exploring the data
    - detection of outliers, of inconsistencies
    - descriptive exploration of the distributions, of correlations
    - data transformations

    - learning sample
    - validation sample
    - test sample
3. For each algorithm: parameter estimation using training and validation samples
4. Choice of final algorithm using testing sample, risk estimation

Machine Learning tools: R

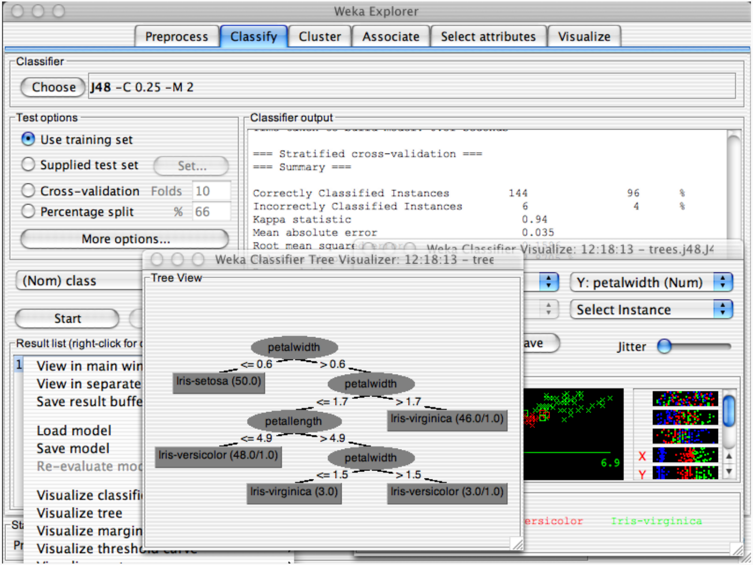


Machine Learning tools: python



scikitlearn: `http://scikit-learn.org/stable/index.html`

Knime, Weka and co: integrated environments



Statistics 101

Statistical model

- Sample size: n
- Observation space: $\mathcal{X}$
- Statistical model = pair $(\mathcal{X}^n, \mathcal{P})$, where $\mathcal{P}$ is a family of probability distributions on $\mathcal{X}^n$
- Observation: $(X_1, \ldots, X_n) \sim P$ where $P \in \mathcal{P}$
- Parametric model : $\mathcal{P} = \left\{ P_\theta : \theta \in \Theta \subset \mathbb{R}^d \right\}$
- Product model: $\mathcal{P} = \left\{ Q^{\otimes n} : Q \in \mathcal{Q} \right\} \overset{param}{=} \left\{ Q_\theta^{\otimes n} : \theta \in \Theta \right\}$
- Bernoulli model: parametric product model with $Q = \mathcal{B}(\theta), \Theta = [0, 1]$

Estimator

- Statistic = any function of $(X_1, \ldots, X_n)$ (and not θ!)
- Estimator of $g(\theta)$ = any statistic; a good estimator tries to be "close" to $g(\theta)$ "whatever the value of θ.
- Ex: Bernoulli model: $\hat{\theta}_n = \bar{X}_n$, $\tilde{\theta}_n = X_1$, $\check{\theta}_n = 2\big(X_1 + \cdots + X_{n/2}\big)/n$
- Bias of T_n: $\theta \mapsto \mathbb{E}_\theta\big[T_n - g(\theta)\big]$
- Consistant: $T_n \xrightarrow{P} g(\theta)$ when $n \to \infty$
- Quadratic risk: $\theta \mapsto \mathbb{E}_\theta\left[\big(T_n - g(\theta)\big)^2\right]$
- Minimax risk:
$$\inf_{T_n} \sup_{\theta \in \Theta} \mathbb{E}_\theta\left[\big(T_n - g(\theta)\big)^2\right]$$
  Minimax estimator: reaches the minimax risk

Moment Estimators

Definition: if $\theta = \phi\Big(E_\theta[X_1], \dots, E_\theta[X_1^d]\Big)$, then

$$\hat{g}_n = \phi\Big(\frac{1}{n}\sum_{i=1}^{n} X_i, \dots, \frac{1}{n}\sum_{i=1}^{n} X_i^d\Big)$$

Prop: if $E_\theta[X_1^d] < \infty$ and if ϕ is continuous, then $\hat{g}_n$ is consistent

Ex: Bernoulli model $\theta = E[X_1] \longrightarrow \hat{\theta}_n = \bar{X}_n$
More generally: if $g(\theta) = \mathbb{E}_\theta[X_1]$, then $\hat{g}_n = \bar{X}_n$

Remark: best constant guess = expectation

Ex: Gaussian model

Maximum Likelihood Estimator

Definition the likelihood function in a parametric model is

$$\ell(\theta, X_1, \dots, X_n) = \begin{cases} P_\theta(X_1, \dots, X_n) & \text{in a discret model} \\ f_\theta(X_1, \dots, X_n) & \text{in a continuous model} \end{cases}$$

Definition The maximum likelihood estimator of θ is defined by

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta} \ell(\theta, X_1, \dots, X_n)$$

Ex: Bernoulli model:

$$\ell(\theta, X_1, \dots, X_n) = \prod_{i=1}^{n} p^{X_1}(1-p)^{1-X_i}$$

and $\hat{\theta}_n = \bar{X}_n$

Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$$

with $\mathbb{E}[\epsilon_i] = 0, \mathbb{V}ar[\epsilon_i] = \sigma^2$ and (ϵ_i) independent.

Matrix form: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, with $X_{i,0} = 1, \mathbf{X} = \left(X_{ij}\right) \in \mathcal{M}_{n,p+1}(\mathbb{R})$ and $Y \in \mathbb{R}^n$ and ϵ random vector with range in $\mathbb{R}^n$.

Least Mean Square estimator:

$$\hat{\beta}_n = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left\|\mathbf{Y} - \mathbf{X}\beta\right\| = \left(X^T X\right)^{-1} X^T Y$$

if $\mathrm{rank}(X) = p + 1$.

- if $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, the ML estimator is the LMS estimator and
$$\hat{\beta}_n \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right)$$
- simple regression: $p = 1$, $\hat{\beta}_{n,1} = \dfrac{\mathbb{C}\mathrm{ov}_n(X, Y)}{\mathbb{V}ar_n(X)}$