Machine Learning 1
Introduction, nearest neighbor classifier

Master 2 Computer Science

Aurélien Garivier

2019-2020

UNIVERSITÉ DE LYON | ENS DE LYON

Table of contents

Before we start

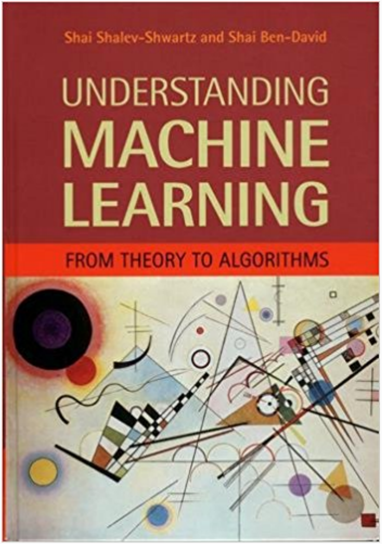
Outline (1/2)

- 1. 09.09 Introduction, nearest-neighbor classification
- 2. 09.16 ML methodology, k-nearest neighbors, decision trees
- 3. 09.23 PAC Learning Theory, no-free-lunch theorem
- 4. 09.30 Dimensionality Reduction: PCA, random projections
- 5. 10.07 VC dimension, empirical risk minimization
- 6. 10.14 Linear separators, Support Vector Machines
- 7. 10.21 Kernels, regularization
- 10.28 holidays
- 8. 11.4 Boosting, Bagging, Random Forests
- 11.11 bank holiday

Outline (2/2)

- 9. 11.18 Neural networks and stochastic gradient descent
- 10. 11.25 Regression, model selection
- 11. 12.02 Clustering
- 12. 12.09 Online Learning
- 13. 12.16 Reinforcement Learning
- 12.23 holidays
- 12.30 holidays
- 14. 01.06 Questions / Exercises
- 15. 01.13 Final Exam

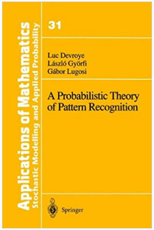
Reference textbook

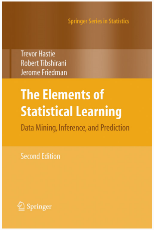


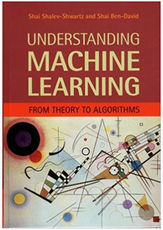
General introduction to Machine Learning theory, by two leading researchers of the field.

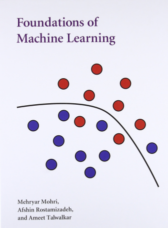
Covers a good part of the content of this course (other references will be provided for specific topics).

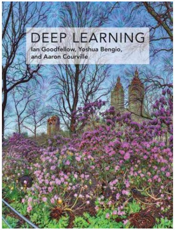
Additional References











WikiStat

scikit learn

Evaluation

Homework and in-class exercises

and

- analysis and review of a research article
  - report + oral presentation
  - articles will be proposed along the lectures
- or participation in a ML student challenge:
  - topic: anomaly detection
  - data: Airbus sensors
  - teams: 4 participants
  - start: October 10th
  - see `https://defi-ia.insa-toulouse.fr/`

(you choose)

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

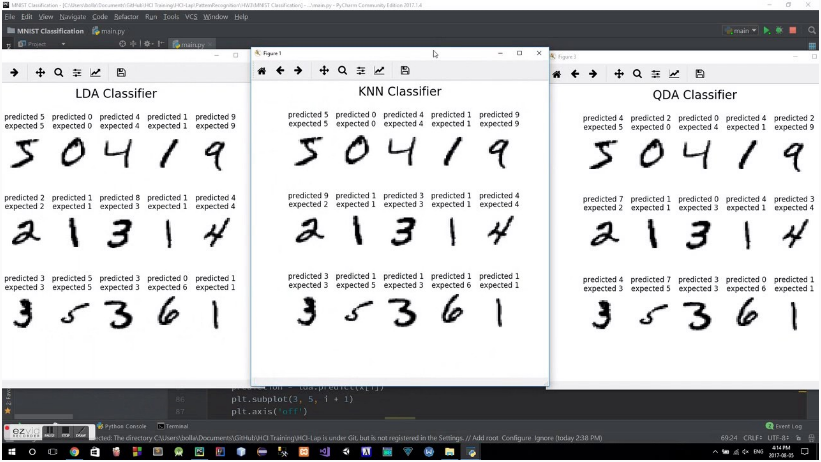
What is Machine Learning?

- 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.

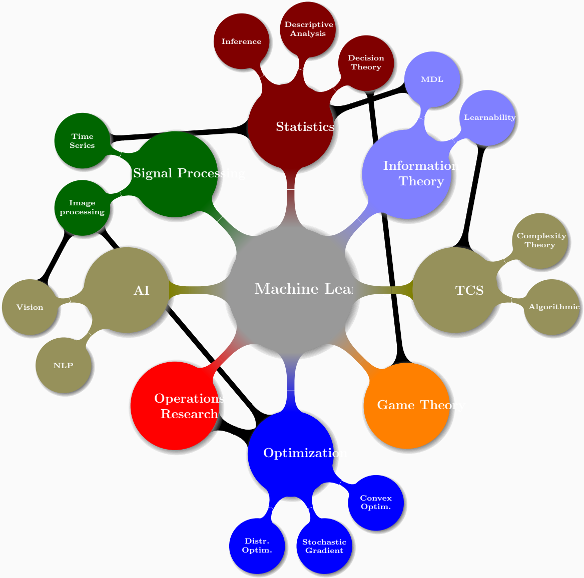
It proves much more succesful to have a machine infer what the good decision rules are.

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

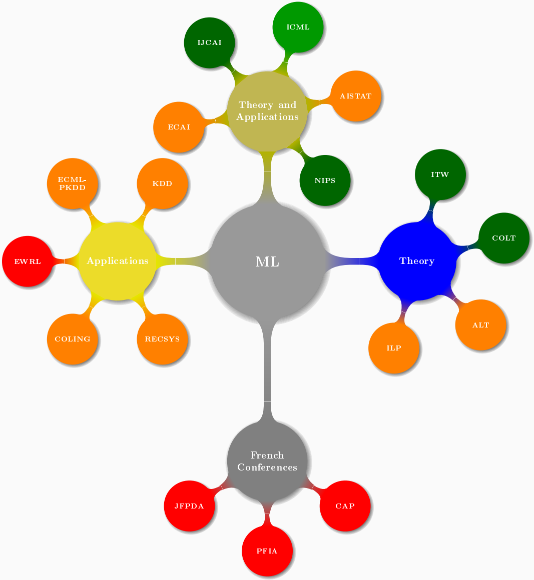
ML and its neighbors



ML journals

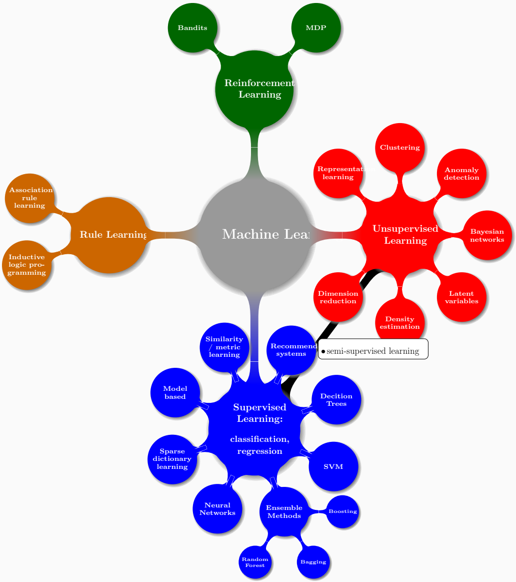

Machine Learning
ML & statistics
JMLR
Machine Learning
Artificial Intelligence
Mathematics Statistics and Learning
Computer science
TCS
Statistics Theory
Annals of Statistics
Bernoulli
IEEE-IT
Electronic Journal of Statistics
ESAIM-PS
Optimization
Optimization series of Imperial College Press
Journal of Optimization Theory and Applications
SIAM journal on optimization
Mathematics of Operations Research

ML conferences



The Learning Models

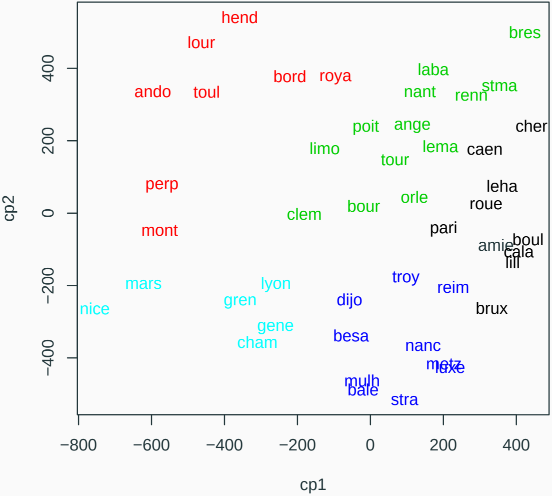
What ML is composed of



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



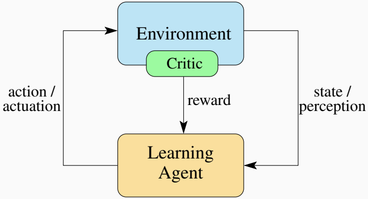
Supervised Learning

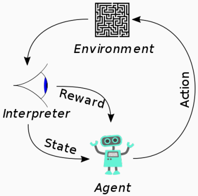
- Observations = pairs (X_i, Y_i)
- Goal = learn to *predict* Y_i *given* X_i
- Regression (when Y is continuous)
- Classification (when Y is discrete)

Examples:

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

Reinforcement Learning

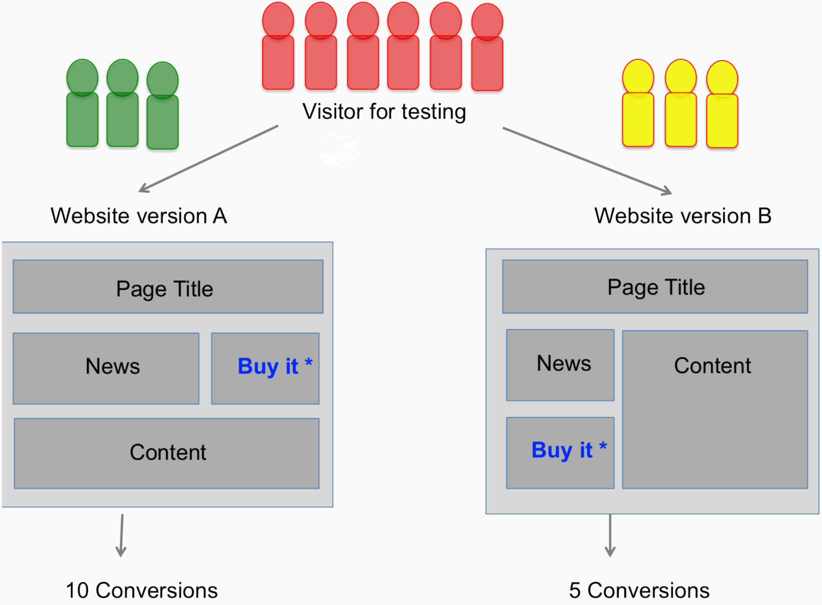




[Src: https://en.wikipedia.org/wiki/Reinforcement_learning]

- 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

m-by-p matrix X

- m 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

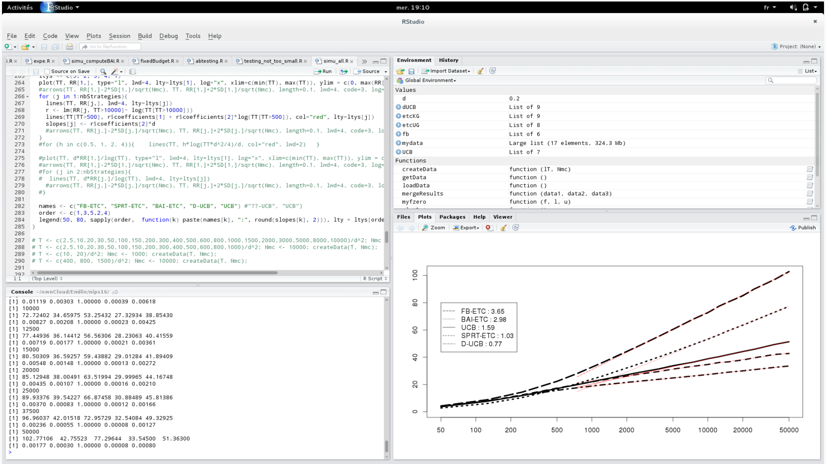
- Inside R: package `datasets`
- Inside scikitlearn: package `sklearn.datasets`
- UCI Machine Learning Repository
- Challenges: Kaggle, etc.

The big steps of data analysis

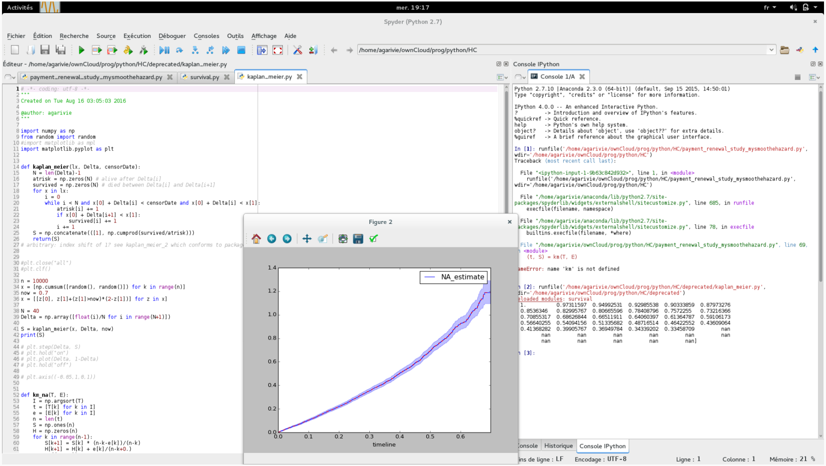
1. Extracting the data to expected format
2. Exploring the data
    - detection of outliers, of inconsistencies
    - descriptive exploration of the distributions, of correlations
    - data transformations

    - 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

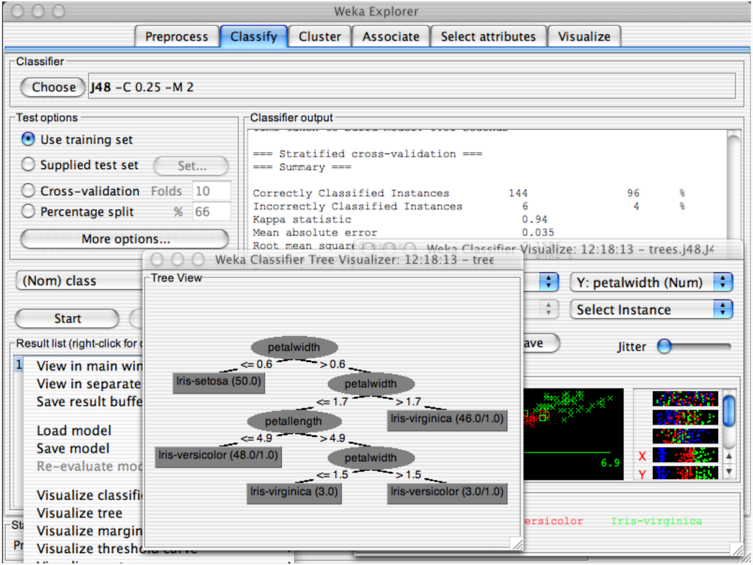


Machine Learning tools: python



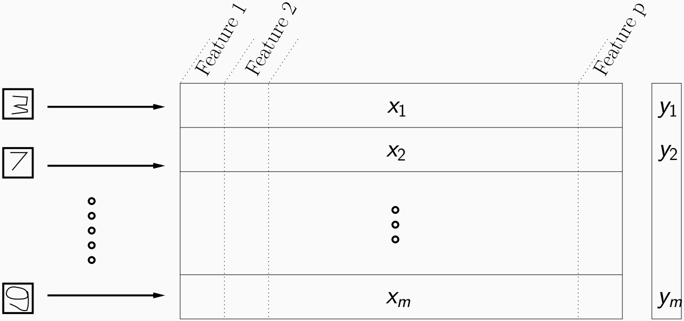
scikitlearn: `http://scikit-learn.org/stable/index.html`

Knime, Weka and co: integrated environments



Supervised Classification

What is a classifier?



$X \in \mathcal{M}_{m,p}(\mathbb{R})$ $\qquad$ $Y \in \mathcal{Y}^m$

Data: m-by-p matrix X

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

Classifier $\mathcal{A}_m$

$h_m : \mathcal{X} \to \mathcal{Y}$

 $\mapsto$ 6

Statistical Learning Hypothesis

Assumption

- The examples $(X_i, Y_i)_{1\leq i\leq m}$ are iid samples of an unknown joint distribution $\mathcal{D}$;
- The points to classify later are also independent draws of the *same* distribution $\mathcal{D}$.

Hence, for every *decision rule* $h : \mathcal{X} \to \mathcal{Y}$ we can define the *risk*

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\big(h(X) \neq Y\big) = \mathcal{D}\Big(\big\{(x,y) : h(x) \neq y\big\}\Big) .$$

The goal of the learning algorithm is to *minimize the expected risk*:

$$R_m(\mathcal{A}_m) = \mathbb{E}_{\mathcal{D}^{\otimes m}}\left[L_{\mathcal{D}}\Big(\underbrace{\mathcal{A}_m\big((X_1, Y_1), \dots, (X_m, Y_m)\big)}_{\hat{h}_m}\Big)\right]$$

for *every* distribution $\mathcal{D}$, using only the examples.

Realizable case vs agnostic learning

One usually distinguishes

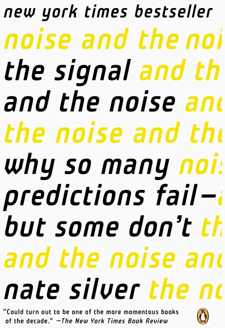
- the *realizable case:* there exists $h : \mathcal{X} \to \mathcal{Y}$ such that $\mathbb{P}_{(X,Y)\sim\mathcal{D}}\big(h(X) = Y\big) = 1$,
- and the *agnostic case* otherwise (x does not permit to predict y with certainty).

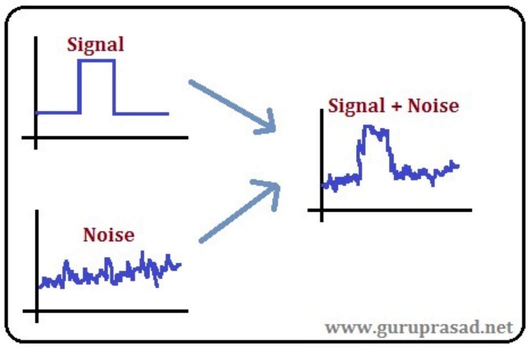
Examples:

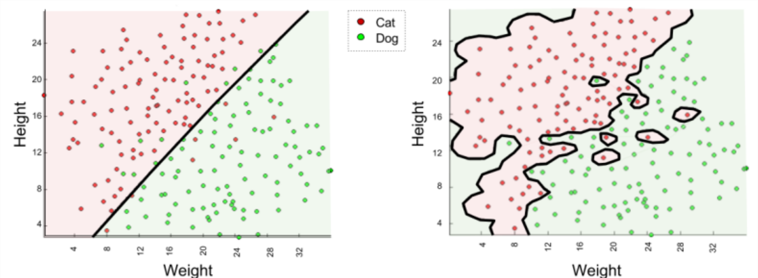
- spam filtering, character recognition
- credit risk, heart disease prediction

We generally focus on the agnostic case.

Signal and Noise







Statistical Learning Framework

- Domain set $\mathcal{X}$
- Label set $\mathcal{Y}$
- Statistical Model: $\{D \text{ probability over } \mathcal{X} \times \mathcal{Y}\}$
- Training data: pairs $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$, $1 \leq i \leq m$
  $m =$ *sample size*
- Learner's output: $\hat{h}_m : \mathcal{X} \to \mathcal{Y}$. Possibly $\hat{h}_m \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$.
- Measures of success: risk measure of hypothesis $h \in \mathcal{H}$

$$L_{\mathcal{D}}(h) = \mathbb{P}_{(X,Y)\sim\mathcal{D}}\big(h(X) \neq Y\big) = D\Big(\big\{(x, y) : h(x) \neq y\big\}\Big) .$$

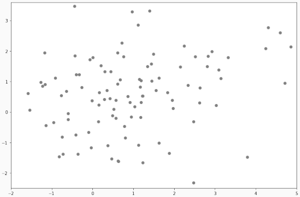
Example: Character Recognition

Domain set $\mathcal{X}$ Label set $\mathcal{Y}$ Joint distribution $\mathcal{D}$	64×64 images $\{0, 1, \ldots, 9\}$?
Prediction function $h \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ Risk $R(h) = P_{X,Y}(h(X) \neq Y)$	
Sample $S = \{(x_i, y_i)\}_{i=1}^m$ Empirical risk $L_S(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}\{h(x_i) \neq y_i\}$	MNIST dataset
Learning algorithm $\mathcal{A} = (\mathcal{A}_m)_m$, $\mathcal{A}_m : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ Expected risk $R_m(\mathcal{A}) = \mathbb{E}_m[L(\mathcal{A}_m(S_m)))]$	neural nets, boosting...
Empirical risk minimizer $\hat{h}_m = \arg\min_{h \in \mathcal{H}} L_S(h)$ Regularized empirical risk minimizer $\hat{h}_m = \arg\min_{h \in \mathcal{H}} L_S(h) + \lambda C(h)$	

Statistical Learning

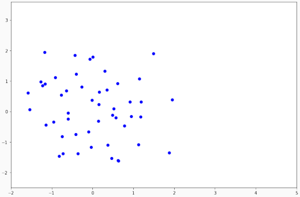
One can have 2 visions of D:

As a pair (D_x, k), where



- for $A \subset \mathcal{X}$, $D_x(A) = D\big(A \times \mathcal{Y}\big)$ is the marginal distribution of X,
- and for $x \in \mathcal{X}$ and $B \subset \mathcal{Y}$, $k(B|x) = \mathbb{P}\big(Y \in B|X = x\big)$ is (a version of) the conditional distribution of Y given X.

As a pair $\Big(D_y, \big(D(\cdot|y)\big)_y\Big)$, where



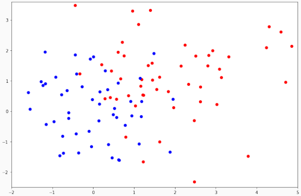
- for $y \in \mathcal{Y}$, $D_y(y) = D\big(\mathcal{X} \times y\big)$ is the marginal distribution of Y,
- and for $A \subset \mathcal{X}$ and $y \in \mathcal{Y}$, $D(A|y) = \mathbb{P}\big(X \in A|Y = y\big)$ is the conditional distribution of X given $Y = y$.

Statistical Learning

One can have 2 visions of D:

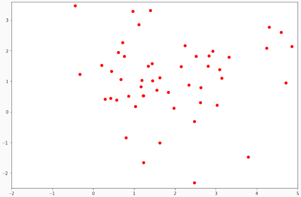
As a pair (D_x, k), where

- for $A \subset \mathcal{X}$, $D_x(A) = D\big(A \times \mathcal{Y}\big)$ is the marginal distribution of X,
- and for $x \in \mathcal{X}$ and $B \subset \mathcal{Y}$, $k(B|x) = \mathbb{P}\big(Y \in B|X = x\big)$ is (a version of) the conditional distribution of Y given X.



As a pair $\Big(D_y, \big(D(\cdot|y)\big)_y\Big)$, where

- for $y \in \mathcal{Y}$, $D_y(y) = D\big(\mathcal{X} \times y\big)$ is the marginal distribution of Y,
- and for $A \subset \mathcal{X}$ and $y \in \mathcal{Y}$, $D(A|y) = \mathbb{P}\big(X \in A|Y = y\big)$ is the conditional distribution of X given $Y = y$.



Bayes Classifier

Consider binary classification $\mathcal{Y} = \{0, 1\}$, let $\eta(x) = \mathbb{P}(Y = 1 | X = x)$.

Theorem

The Bayes classifier is defined by
$h^*(x) = \mathbb{1}\{\eta(x) \geq 1/2\} = \mathbb{1}\{\eta(x) \geq 1 - \eta(x)\} = \mathbb{1}\{2\eta(x) - 1 \geq 0\}$.
For every classifier $h : \mathcal{X} \to \mathcal{Y} = \{0, 1\}$,

$$L_{\mathcal{D}}(h) \geq L_D(h^*) = \mathbb{E}\Big[\min\big(\eta(X), 1 - \eta(X)\big)\Big].$$

The Bayes risk $L_D^* = L_D(h^*)$ *is called the* **noise** *of the problem.*

More precisely,

$$L_D(h) - L_D(h^*) = \mathbb{E}\Big[\big|2\eta(X) - 1\big|\ \mathbb{1}\{h(X) \neq h^*(X)\}\Big].$$

Extends to $|\mathcal{Y}| > 2$.

Proof

$$
\begin{aligned}
L_D(h) - L_D(h^*) &= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(\\
&\quad \mathbb{1}\{Y = 1\}\Big(\mathbb{1}\{h^*(X) = 1\} - \mathbb{1}\{h^*(X) = 0\}) \\
&\quad + \mathbb{1}\{Y = 0\}\Big(\mathbb{1}\{h^*(X) = 0\} - \mathbb{1}\{h^*(X) = 1\})\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{Y = 1\} - 1\Big)\Big(2\mathbb{1}\{h^*(X) = 1\} - 1\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{Y = 1\} - 1\Big)\Big(2\mathbb{1}\{\eta(X) \geq \frac{1}{2}\} - 1\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{\eta(X) \geq \frac{1}{2}\} - 1\Big)\mathbb{E}\Big[2\mathbb{1}\{Y = 1\} - 1 \mid X\Big]\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\Big(2\mathbb{1}\{\eta(X) \geq \frac{1}{2}\} - 1\Big)\Big(2\mathbb{E}[\mathbb{1}\{Y = 1\} \mid X] - 1\Big)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\,\mathrm{sign}\Big(\eta(X) - \frac{1}{2}\Big)(2\eta(X) - 1)\Big] \\
&= \mathbb{E}\Big[\mathbb{1}\{h(X) \neq h^*(X)\}\big|2\eta(X) - 1\big|\Big]
\end{aligned}
$$

Nearest-Neighbor Classification

The Nearest-Neighbor Classifier

We assume that $\mathcal{X}$ is a metric space with distance d.

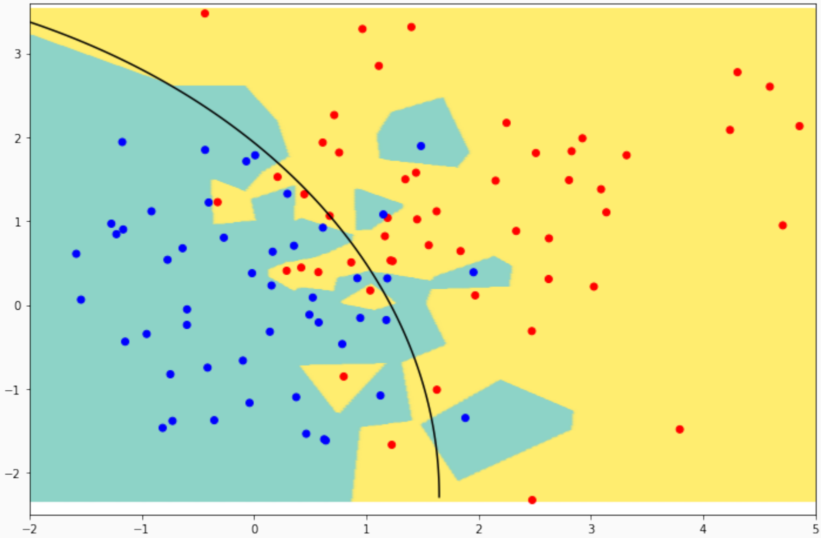
The nearest-neighbor classifier $\hat{h}_m^{NN} : \mathcal{X} \to \mathcal{Y}$ is defined as

$$\hat{h}_m^{NN}(x) = Y_I \quad \text{where} \quad I \in \operatorname*{arg\,min}_{1 \le i \le m} \; d(x - X_i)\;.$$

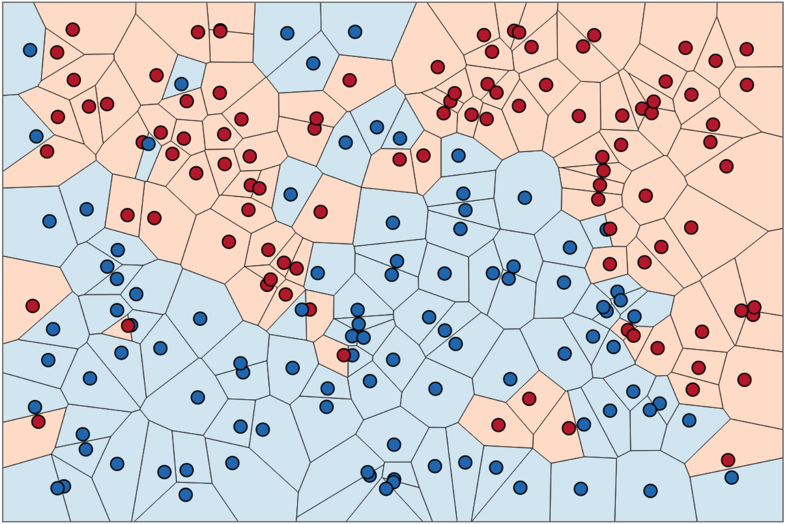
Typical distance: L^2 norm on $\mathbb{R}^d$ $\|x - x'\| = \sqrt{\sum_{j=1}^d (x_i - x_i')^2}$.

Buts many other possibilities: Hamming distance on $\{0, 1\}^d$, etc.

Numerically



Numerically



Analysis

A1. $\mathcal{Y} = \{0, 1\}$.

A2. $\mathcal{X} = [0, 1[^d$.

A3. η is c-Lipschitz continuous:

$$\forall x, x' \in \mathcal{X}, \left|\eta(x) - \eta(x')\right| \leq c\left\|x - x'\right\| .$$

.

Theorem

Under the previous assumptions, for all distributions D and all $m \geq 1$

$$R_m\left(\hat{h}_m^{NN}\right) \leq 2L_D^* + \frac{3c\sqrt{d}}{m^{1/(d+1)}}$$

Proof Outline

- Conditioning: as $I(x) = \arg\min_{1\leq i\leq m} \|x - X_i\|$,

$$R_m\big(\hat{h}_m^{NN}\big) = \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}\{Y \neq Y_{I(X)}\}\big|X, X_1, \dots, X_m\big]\Big] \ .$$

- $Y \sim \mathcal{B}(p),\ Y' \sim \mathcal{B}(q) \implies \mathbb{P}(Y \neq Y') \leq 2\min(p, 1-p) + |p - q|$,

$$\mathbb{E}\Big[\mathbb{1}\{Y \neq Y_{I(X)}\}|X, X_1, \dots, X_m\Big] \leq 2\min\big(\eta(X), 1-\eta(X)\big) + c\,\big\|X - X_{I(X)}\big\| \ .$$

- Partition $\mathcal{X}$ into $|\mathcal{C}| = T^d$ cells of diameter $\sqrt{d}/T$:

$$\mathcal{C} = \left\{\left[\frac{j_1 - 1}{T}, \frac{j_1}{T}\right[\times \cdots \times \left[\frac{j_d - 1}{T}, \frac{j_d}{T}\right[, \quad 1 \leq j_1, \dots, j_d \leq T\right\} \ .$$

- 2 cases: either the cell of X is occupied by a sample point, or not:

$$\big\|X - X_{I(X)}\big\| \leq \sum_{c\in\mathcal{C}} \mathbb{1}\{X \in c\}\left(\frac{\sqrt{d}}{T}\mathbb{1}\bigcup_{i=1}^{m}\{X_i \in c\} + \sqrt{d}\,\mathbb{1}\bigcap_{i=1}^{m}\{X_i \notin c\}\right) \ .$$

- $\implies \mathbb{E}\big[\|X - X_{I(X)}\|\big] \leq \frac{\sqrt{d}}{T} + \frac{\sqrt{d}\,T^d}{e\,m}$ and choose $T = \Big\lfloor m^{\frac{1}{d+1}}\Big\rfloor$.

What does the analysis say?

- Is it loose? (sanity check: uniform $\mathcal{D}_X$)
- Non-asympototic (finite sample bound)
- The second term $\frac{3c\sqrt{d}}{m^{1/(d+1)}}$ is distribution independent
- Does not give the trajectorial decrease of risk
- Exponential bound d (cannot be avoided...)

  $\implies$ *curse of dimensionality*

- How to improve the classifier?