Machine Learning

Master 2 Computer Science

Aurélien Garivier

2018-2019

UNIVERSITÉ DE LYON | ENS DE LYON

Table of contents

Before we start

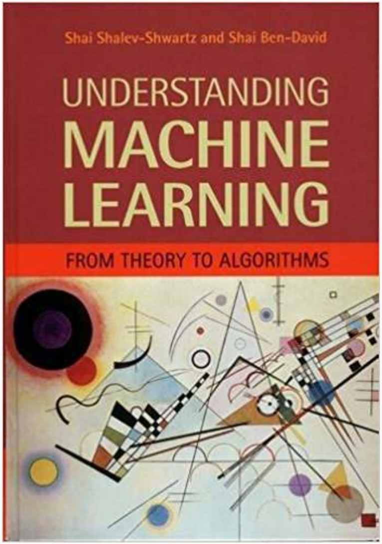
Outline (1/2)

- 1. 09.10 Introduction, nearest-neighbor classification
- 09.17 Pot du DI at IFE Descartes (meet your tutor)
- 2. 09.24 ML methodology, k-nearest neighbors, decision trees
- 3. 10.01 PAC Learning Theory, no-free-lunch theorem
- 4. 10.8 VC dimension, empirical risk minimization
- 5. 10.15 Linear separators, Support Vector Machines
- 6. 10.22 Kernels, regularization
- 10.29 holidays

Outline (2/2)

- 7. 11.5 Boosting, Bagging, Random Forests
- 11.12 winter school
- 8. 11.19 Neural networks and stochastic gradient descent
- 11.26 winter school
- 12.03 no lecture
- 9. 12.10 Regression, model selection
- 10. 12.17 Dimension reduction
- 11. 01.07 Clustering
- 12. 01.14 Online learning

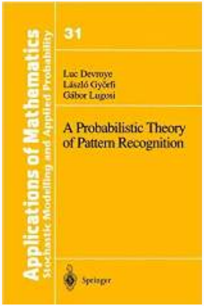
Reference textbook

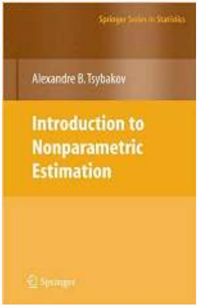


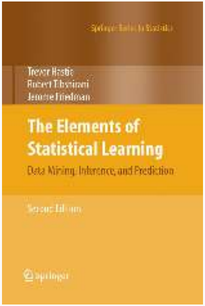
General introduction to Machine Learning theory, by two leading researchers of the field.

Covers most of the content of this course (the converse is also almost true).

Additional References







Evaluation

Homework and in-class exercises

and

- analysis and review of a research article (report + oral presentation)
- <u>or</u> participation in a student ML challenge

(you choose)

What is Machine Learning?

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.

ML: Learn from and make predictions on data

- 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".

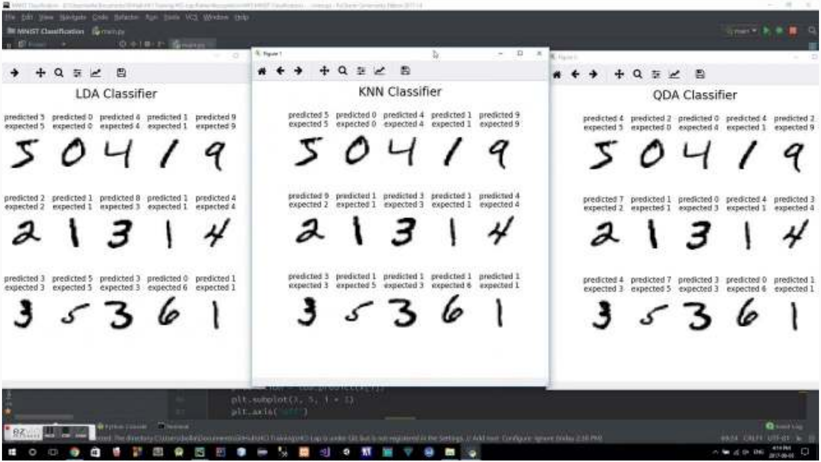
Machine Learning: Typical Problems

- 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.

Example: MNIST dataset



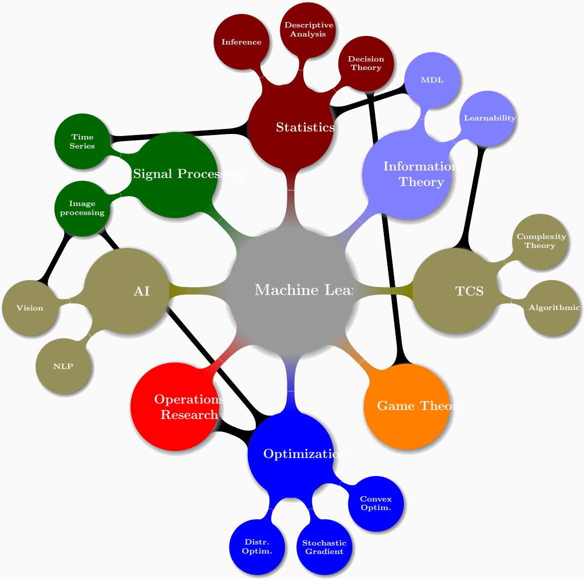
Related Fields

- **Computational Statistics**: focuses in prediction-making through the use of computers together with statistical models (ex: Bayesian methods).
- **Statistical Learning**: ML by statistical methods, with statistical point of view (probabilistic guarantees: consistency, oracle inequalities, minimax)
- **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.
- Importance of **probability**- and **statistics**-based methods $\rightarrow$ **Data Science** (Michael Jordan)
- Strong ties to **Mathematical Optimization**, which delivers methods, theory and application domains to the field

Machine Learning and Statistics

- Data analysis (inference, description) is the goal of statistics for long.
- Machine Learning has more **operational** goals (ex: consistency is important the statistics literature, but often makes little sense in ML).
  Models (if any) are *instrumental*.
  Ex: linear model (nice mathematical theory) vs Random Forests.
- Machine Learning/big data: no seperation between statistical modelling and optimization (in contrast to the statistics tradition).
- In ML, data is often here before (unfortunately).
- ML more focused on *correlation*, less on *causality*.
- Algorithmic considerations play a major role in ML.
- No clear separation (statistics evolves as well), but different hypotheses focus of interest. Ex: model-free versus model-based, asymptotic consistency versus finite sample bounds.

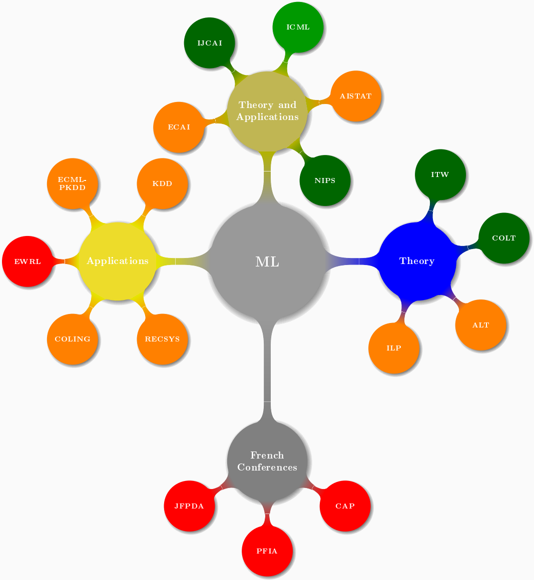
ML and its neighbors



ML journals

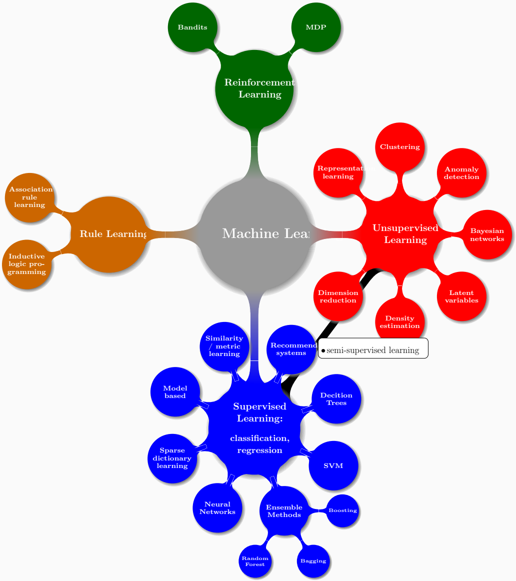
Machine Learning
ML & statistics
JMLR
Machine Learning
Artificial Intelligence
Mathematics, Statistics and Learning
Statistics Theory
Annals of Statistics
Bernoulli
IEEE-IT
Electronic Journal of Statistics
ESAIM-PS
Computer science
TCS
Optimization
Optimization series of Imperial College Press
Journal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

ML conferences



The Learning Models

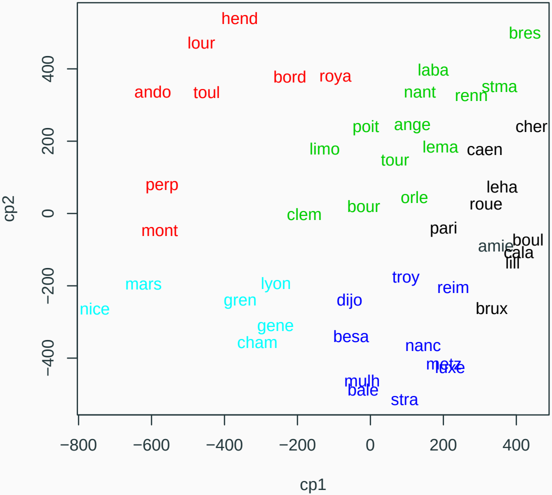
What ML is composed of



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



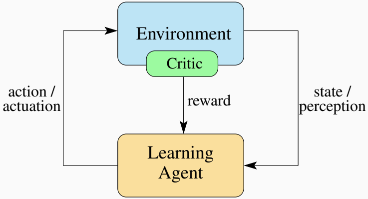
Supervised Learning

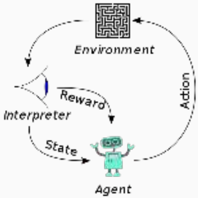
- Observations = pairs (X_i, Y_i)
- Goal = learn to *predict* Y_i *given* X_i
- Regression (when Y is continuous)
- Classification (when Y is discrete)

Examples:

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

Reinforcement Learning

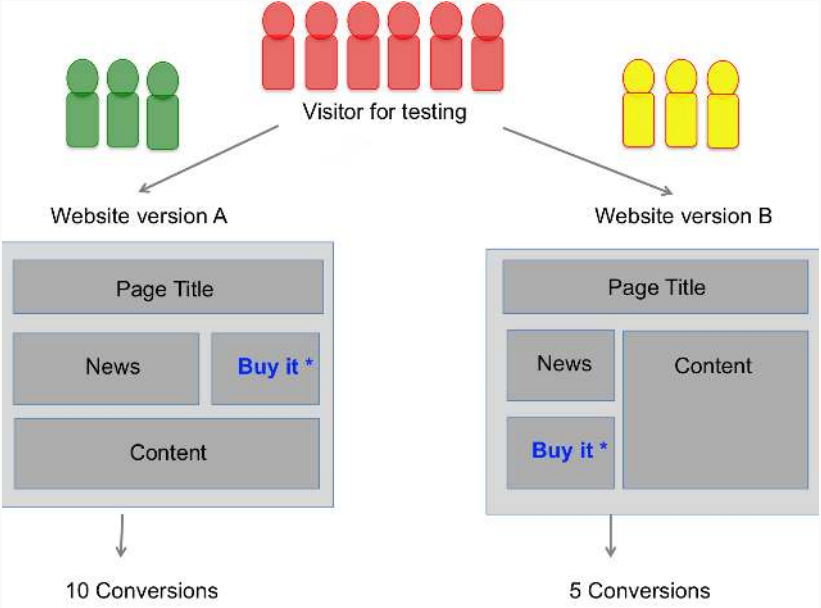




[Src: https://en.wikipedia.org/wiki/Reinforcement_learning]

- 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

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 (n, p) often called

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

Data repositories

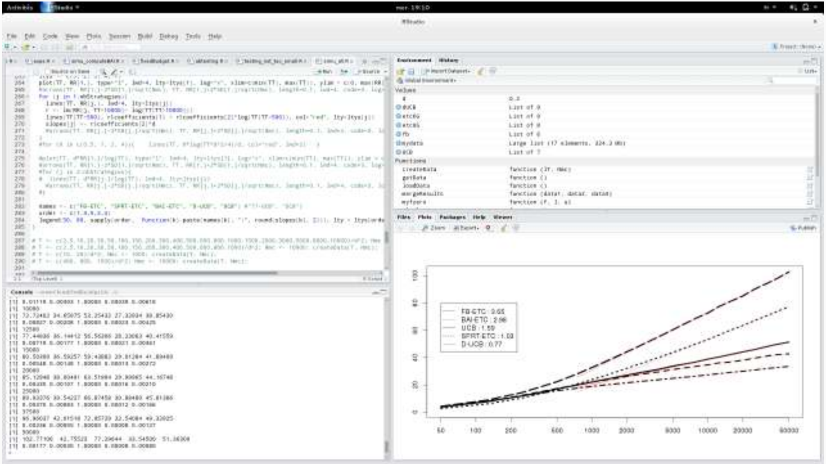
- Inside R: package `datasets`
- Inside scikitlearn: package `sklearn.datasets`
- UCI Machine Learning Repository
- Challenges: Kaggle, etc.

The big steps of data analysis

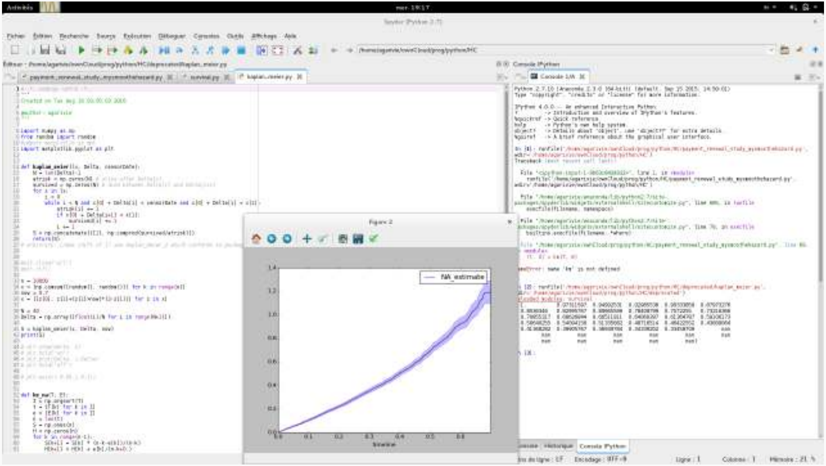
1. Extracting the data to expected format
2. Exploring the data
    - detection of outliers, of inconsistencies
    - descriptive exploration of the distributions, of correlations
    - data transformations

    - 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

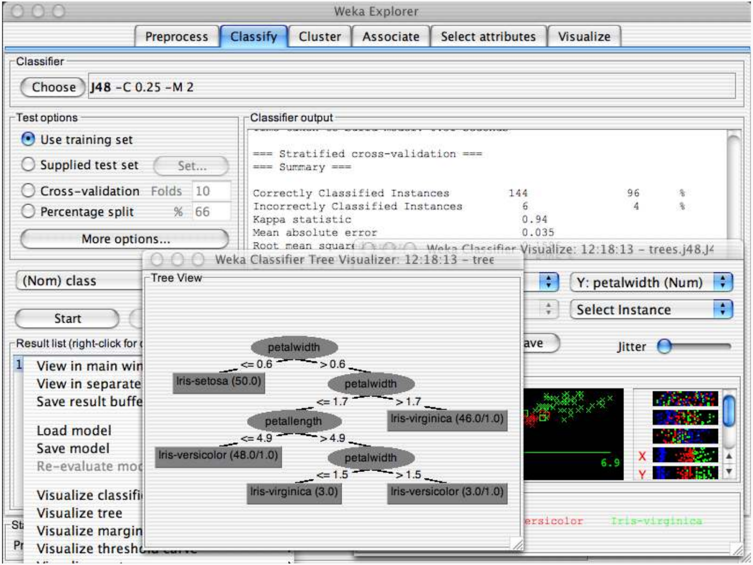


Machine Learning tools: python



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

Knime, Weka and co: integrated environments



Supervised Classification

Statistical Learning Framework

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

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

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 $\mathcal{A} = (\mathcal{A}_n)_n$, $\mathcal{A}_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ Expected risk $R_n(\mathcal{A}) = \mathbb{E}_n[R(\mathcal{A}_n(D_n)))]$	neural nets, boosting...
Empirical risk minimizer $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$ Regularized empirical risk minimizer $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h) + \lambda C(h)$	

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:

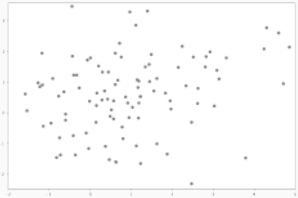
- spam filtering, character recognition
- credit risk, heart disease prediction

We generally focus on the agnostic case.

One can have 2 visions of D:

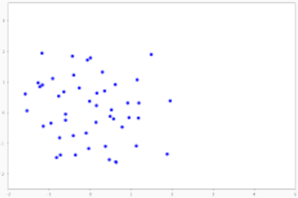
As a pair (D_x, k), where

- 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

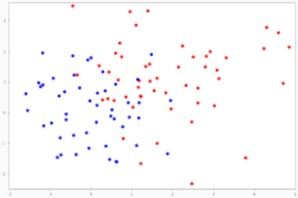
- for $y \in \mathcal{Y}$, $D_y(y) = D\big(\mathcal{X} \times y\big)$ is the marginal distribution of Y,
- and for $A \subset \mathcal{X}$ and $y \in \mathcal{Y}$, $D(A|y) = \mathbb{P}\big(X \in A | Y = y\big)$ is the conditional distribution of X given $Y = y$.



One can have 2 visions of D:

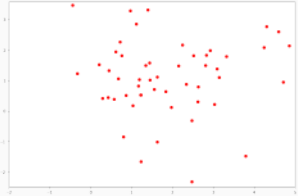
As a pair (D_x, k), where

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

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

Nearest-Neighbor Classification

The Nearest-Neighbor Classifier

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

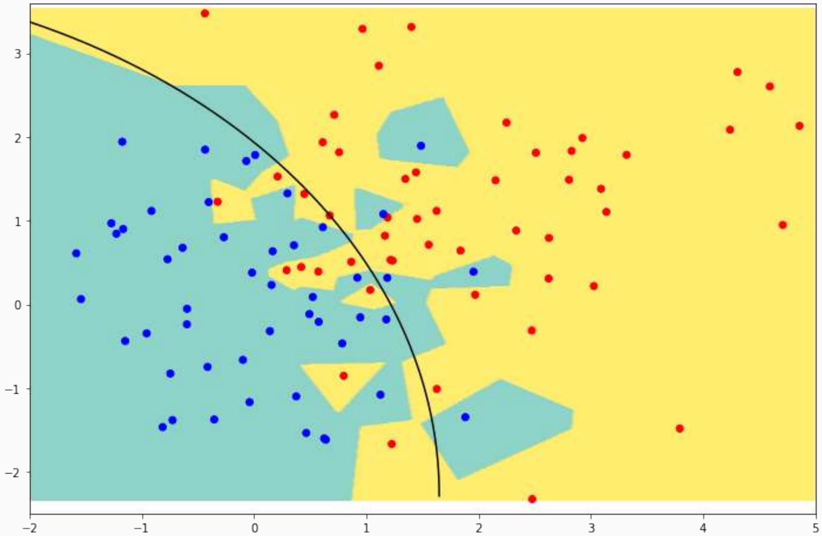
The nearest-neighbor classifier $\hat{h}_m^{NN} : \mathcal{X} \to \mathcal{Y}$ is defined as

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

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



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 $m \geq 1$

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

Proof Outline

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

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

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

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

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

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

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

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

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

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?

Deviation Bound for Bernoulli Variables

Remember: Jensen's Inequality

Basic version: if $\phi : \mathcal{X} \to \mathbb{R}$ is convex and $t \in (0, 1)$ then for all $x, x' \in \mathcal{X}$, $f(tx + (1-t)x') \le tf(x) + (1-t)f(x')$.

Probabilistic version: If $\phi : \mathcal{X} \to \mathbb{R}$ is convex and if X is a random variable with range in $\mathcal{X}$, then $\phi\big(\mathbb{E}[X]\big) \le \mathbb{E}\big[\phi(X =)\big]$.

Example: For a real-valued random variable X $\mathbb{E}[X^2] \ge \mathbb{E}[X]^2$ and thus $\mathbb{V}ar[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \ge 0$.

Think about equality case.

Chernoff's Bound

Theorem (Chernoff-Hoeffding Deviation Bound)

Let $\mu \in (0,1)$. $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{B}(\mu)$, *and let* $x \in (\mu, 1]$.

(i) *Chernoffs' bound for Bernoulli variables:*

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp\left(- n\, \mathrm{kl}(x, \mu)\right) , \tag{1}$$

where $\mathrm{kl}(p,q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$.

(ii) *Hoeffding's bound for Bernoulli variables: since* $\mathrm{kl}(p,q) \geq 2(p-q)^2$,

$$\mathbb{P}(\bar{X}_n \geq x) \leq \exp\left(- 2n(x,\mu)^2 \right) . \tag{2}$$

(iii) *Inequalities* (1) *and* (2) *hold for arbitrary independent random variables with range* $[0,1]$ *and expectation* μ.

Reason: $\exp(\lambda x) \leq (1-x)\exp(0) + x \exp(\lambda)$.

Kullback-Leibler Divergence

Definition

Let P and Q be two probability distributions on a set $\mathcal{X}$. The Kullback-Leibler divergence from Q to P is defined as follows:

- if P is not absolutely continuous with respect to Q, then $\mathrm{KL}(P, Q) = +\infty$;
- otherwise, let $\frac{dP}{dQ}$ be the Radon-Nikodym derivative of P with respect to Q. Then

$$\mathrm{KL}(P, Q) = \int_{\mathcal{X}} \log \frac{dP}{dQ}\, dP = \int_{\mathcal{X}} \frac{dP}{dQ} \log \frac{dP}{dQ}\, dQ\ .$$

  The integral always exists but may be equal to $+\infty$.

Examples:

- $\mathrm{KL}\left(\mathcal{B}(p), \mathcal{B}(q)\right) = \mathrm{kl}(p, q)$,
- $\mathrm{KL}\left(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)\right) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.

Properties

Tensorization of entropy: If $P = P_1 \otimes P_2$ and $Q = Q_1 \otimes Q_2$, then

$$\mathrm{KL}(P, Q) = \mathrm{KL}(P_1, Q_1) + \mathrm{KL}(P_2, Q_2) \; .$$

Contraction of entropy aka **data-processing inequality:**
Let P and Q be probability distributions on $\mathcal{X}$, and let $X \sim P$ and $Y \sim Q$. If $f : \mathcal{X} \to \mathcal{X}'$ is a measurable function and if $\tilde{P}$ (resp. $\tilde{Q}$) is the distribution of $f(X)$ (resp. $f(Y)$), then

$$\mathrm{KL}(\tilde{P}, \tilde{Q}) \leq \mathrm{KL}(P, Q) \; .$$

Application: Lower bound

Let $\mu \in (0,1)$. $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{B}(\mu)$, and let $x \in (\mu, 1]$. Then

$$\liminf_m \frac{1}{m} \log \mathbb{P}(\bar{X}_m > x) \geq -\,\mathrm{kl}(x, \mu)$$

"Chernoff's bound is asymptotically almost tight"