Machine Learning

Master 2 Computer Science

Aurélien Garivier

2018-2019



- 1. Before we start
- 2. What is Machine Learning?
- 3. The Learning Models
- 4. Machine Learning Methodology
- 5. Supervised Classification
- 6. Nearest-Neighbor Classification
- 7. Deviation Bound for Bernoulli Variables

Before we start

- 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

- 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

Shai Shaley-Shwartz and Shai Ben-David

UNDERSTANDING MACHINE LEARNING

FROM THEORY TO ALGORITHMS



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







Homework and in-class exercises

and

- \bullet analysis and review of a research article (report + oral presentation)
- or participation in a student ML challenge

(you choose)

What is Machine Learning?

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.
- Al: 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 $probability\mathchar`-$ and $statistics\mathchar`-based methods \rightarrow Data Science (Michael Jordan)$
- 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

ML conferences

The Learning Models

What ML is composed of

- (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

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)

- Inside R: package datasets
- Inside scikitlearn: package sklearn.datasets
- UCI Machine Learning Repository

• Challenges: Kaggle, etc.

- 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

Adakis Tinks -			н- +; Q -
	Minato		
Fig. Full Look New Plate Servery Build Datase Tech. Hole			
Restriction and the second			K front thinks
14. Courts from consisters a factorists in strength, Chains of the mather of the state of the	Industrie Balance		1.71
Instantes (1.7.1)	of al (mentlesset. /	8	UR+
254 plot T7, 8014.1, trans 17, babed, its itself, lagree, size cannot T1, and T11, plan + 0.0, and 901	da initial francoscowie-		
200 ARCOMETT, MITCH 20011 (CARTING), VT. MITCH 2017001 (CARTING), LANSING,	Velues		
267 Laws W. Wij. 1. Ind 4. Hy-Inst 11	2011	0.3	
200 F v In W(), Thiosophic lagerting in the second standard in the TVT shift, and Sund' the free bits	Gatrin	List of a	
270 . skpes[] - ricseffscients[]*4	Gatcal	Last of #	
379 Privati(11, 181), [-913(1, [930(1)001), 17, 391), [e1700]], [engTime1, [length=1,], heise, called, in 273	an	List of 6	
379 Stor D. D. LUG, J. J. A. MAL, Langer U. Sting (T19) 2041/0. (Distribution in Index). 3	Unydata	Large last (17 elements, 224.3 00)	
378 28 defectT, d'Mill.1/bejTT: tops/17. heid. Desider[1], Lapris', slawshiefT1, westTT1: slaws.	0.00	1.315-07-7	
276 serves(1), #(1),) 25011, Journance, 11, 46(r.)-2*m(r.)/mpromes, Suppose, 1, Saba, Lanca, Sep-	Liedietata -	facilities city, mich	
270 + DirectTT, d'Mily 1/LegiTT), Ledist, Hyr.Start(12)	pethota	function ()	12
279 Marcanette and [-]-1*20] [-]regeneration, W. W[], [-]*00] [-]regeneration, longthed to be a cannot, be	SoudData	fanction <)	
281	and governments.	function (F. 1. a)	
383 Kames v. ET HG-ETC', "SHIT-ETC', "BHC-ETC', "B-COP', "BCB') #"TO-REP. "BCB")		and the second sec	
384 Jepent 30, 84, samply order. American(b) sorte (same)[b], "(", round slopesth), [10], hty - htypiords	Files Plate Paillages Hele	. Weinst	#0
700	A State al Dest.	4. C. W.	P. Martin
207 # T == 123.5, 14, 18, 18, 14, 181, 190, 204, 191, 401, 193, 191, 190, 1100, 1100, 1100, 1004, 1004, 1004, 1000			
200 / T			
320 # T = 1(400, 800, 100(1097) Her = 10000 createdata(T, Merc): 300			
31 Detroit 1 Store	8 -		/
The second s	1.5		/
111 8.01118 8.00808 1.80081 8.08088 8.00818			-
11 10000	Faeto	9.65	
111 B. DELT D. GOLDER I. BOODER B. BOODER B. BOODER B. GOLDER	RA-ETC	0.266	
111 12540	- UCB (10 10 1 10	
11 8.00718.0.00177 1.00081 8.10021 1.00081	- D-UC8	077	
111 YOON			
111 8.08548 0.00140 1.80080 8.08018 1.00272		and contraction of the second s	
111 20040 111 PS 12040 BD RDARF ER STORA 25 RODET 44 NOTAL		and the second s	
11 #.06438 8.00107 1.80088 8.8818 8.00210	8		
111 2010 30 54017 00 2010 00 2010 00 2010 00 2010 00 000 0	1000	and the second se	
117 8.08375 8.00888 1.80088 8.18012 0.00188			
111 20100 111 86.86017 42.81518 72.85720 32.54084 49.33025	0 1		
11 B. CEZIN H. JORNI I. BOORI B. BOORI B. JOLIT	60 107	500 600 1000 5000 8000	10000 20000 60000
11 102.77100 42.75521 77.29644 30.54520 51.36304			
11] 8.08177 0.00800 1.80080 8.08008 0.00800			

Machine Learning tools: python

Classification

fzer 25 rg is with eaugery an extenbelo spe as Applications: Coart deletting, Image ****** Also there Sale country there mader f. Con. -Division-

Dimensionality reduction

Excluding the humble of random wante cal 12 (2/1 20)

And rations Weinknight or should diam'r. Alto think PCS subjects and mail INVESTIGATION PARTY

Algorithme SVE, sugar . Moriel selection

encounted all an an an an

Repression

** W.

aman

Comparing workfulling and encounty samelers are worked. Goal: Incontrast approximity do paratritizer. 1000 Monifest and ward, and while a 151.26

field to a success washing the

Annihurings Date seconds Mark-

1. 1. 10. Preprocessing

Clustering

ash.

Pair # +

- Family

heading coll action and normal particip Application Insolantic ht...... such as test to use with most the entries. also three Nodaks an extent of the 37433 N 1

Active of egreciality of whit for expansions

And rations Contract terraristics.

Also there is bloom a mellikes in

Grouping event percentages

PARS 8

-Dans is

Nette-

Co-going development: What's new

Community

About un Soort normality of the ring. Note Webling Learning Frankel

27

Knime, Weka and co: integrated environments

Supervised Classification

- Domain set \mathcal{X}
- \bullet Label set ${\mathcal Y}$
- Statistical Model: $\{D \text{ probability over } \mathcal{X} \times \mathcal{Y}\}$
- Training data: pairs (X_i, Y_i) ∈ X × Y, 1 ≤ i ≤ 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

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

Example: Character Recognition

Domain set \mathcal{X}	64×64 images
Label set ${\mathcal Y}$	$\{0,1,\ldots,9\}$
Joint distribution ${\cal D}$?
Prediction function $h \in \mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$	
$Risk \ R(h) = P_{X,Y}(h(X) \neq Y)$	
Sample $S = \{(x_i, y_i)\}_{i=1}^m$	MNIST dataset
Empirical risk	
$L_{\mathcal{S}}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq y_i\}$	
Learning algorithm	
$\mathcal{A}=(\mathcal{A}_n)_n,~\mathcal{A}_n:(\mathcal{X} imes\mathcal{Y})^n ightarrow\mathcal{H}$	neural nets, boosting
Expected risk $R_n(\mathcal{A}) = \mathbb{E}_n[R(\mathcal{A}_n(D_n)))]$	
Empirical risk minimizer	
$\hat{h}_n = { m argmin}_{h\in \mathcal{H}}\hat{R}_n(h)$	
Regularized empirical risk minimizer	
$\hat{h}_n = { m argmin}_{h\in \mathcal{H}}\hat{R}_n(h) + \lambda \mathcal{C}(h)$	

One usually distinguishes

- the *realizable case:* there exists $h : \mathcal{X} \to \mathcal{Y}$ such that $\mathbb{P}_{(X,Y)\sim \mathcal{D}}(h(X) = Y) = 1$,
- 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.

Statistical Learning

One can have 2 visions of *D*: As a pair (D_x, k) , where

- for A ⊂ X, D_x(A) = D(A × Y) is the marginal distribution of X,
- and for x ∈ X and B ⊂ Y,
 k(B|x) = P(Y ∈ B|X = x) is (a version of)
 the conditional distribution of Y given X.

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

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

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

- for y ∈ 𝔅, D_y(y) = D(𝔅 × y) is the marginal distribution of Y,
- and for $A \subset \mathcal{X}$ and $y \in \mathcal{Y}$, $D(A|y) = \mathbb{P}(X \in A|Y = y)$ 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 $h^*(x) = \mathbb{1}\{\eta(x) \ge 1/2\} = \mathbb{1}\{\eta(x) \ge 1 - \eta(x)\} = \mathbb{1}\{2\eta(x) - 1 \ge 0\}.$ 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,

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

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

Nearest-Neighbor Classification

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

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

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

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:

$$orall x, x' \in \mathcal{X}, ig| \eta(x) - \eta(x') ig| \leq c ig\| x - x' \|$$
 .

Theorem

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

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

Proof Outline

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

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

- $Y \sim \mathcal{B}(p), Y' \sim \mathcal{B}(q) \implies \mathbb{P}(Y \neq Y') \leq 2\min(p, 1-p) + |p-q|,$ $\mathbb{E}\Big[\mathbb{1}\{Y \neq Y_{I(X)}\}|X, X_1, \dots, X_m\Big] \leq 2\min(\eta(X), 1-\eta(X)) + c ||X - X_{I(X)}||.$
- 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:

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

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

- 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

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

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

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

(i) Chernoffs' bound for Bernoulli variables:

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

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

(ii) Hoeffding's bound for Bernoulli variables: since $kl(p,q) \ge 2(p-q)^2$,

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

(iii) Inequalities (1) and (2) hold for arbitrary independent random variables with range [0, 1] and expectation μ.
 Reason: exp(λx) ≤ (1 − x) exp(0) + x exp(λ).

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 $KL(P, Q) = +\infty;$
- otherwise, let $\frac{dP}{dQ}$ be the Radon-Nikodym derivative of P with respect to Q. Then

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

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

Examples:

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

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

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

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

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

Let
$$\mu \in (0,1)$$
. $X_1, \ldots, X_n \stackrel{\text{iid}}{\sim} \mathcal{B}(\mu)$, and let $x \in (\mu, 1]$. Then
$$\liminf_m \frac{1}{m} \log \mathbb{P}(\bar{X}_m > x) \ge - \operatorname{kl}(x, \mu)$$

"Chernoff's bound is asymptotically almost tight"