



école —
normale —
supérieure —
paris-saclay —

université
PARIS-SACLAY

Introduction to Statistical Learning

Nicolas Vayatis

Lecture 2 - Probability inequalities, complexity measures

Course overview

- Chapter 1 : Optimality in statistical learning
Probabilistic view / Performance criteria / Optimal elements
- **Chapter 2 : Mathematical foundations of statistical learning**
Concentration inequality / Complexity measures /
Regularization
- Chapter 3 : Consistency of mainstream machine learning methods
Boosting, SVM, Neural networks / Bagging, Random forests

Main messages of Chapter 1 (so far)

- To account for the uncertainty of evaluation, data are assumed to be sampled according to a *fixed* but *unknown* **probability distribution**.
- A prediction objective is characterized by an **error measure**, e.g. the classification *problem* is characterized by misclassification rate as an *error measure* in the case of predicting binary labels using supervised classification *data*.
- The nature of **optimal elements** does tell something about the difficulty of the prediction objective.
- Empirical Risk Minimization (ERM) can be seen as a **generic inference principle** accounting for global optimization methods (e.g. based on convex losses in the case of convex risk minimization)

Chapter 2 - Mathematical tools

Probability inequalities

Complexity measures

Regularization and stability

Chapter 2

Introduction

Learning like the twenty-question game

- Assume Nature has picked one function among K and we want to reveal this function
- Assume we have an oracle answering YES or NO when we ask a question about this function
- What is the optimal number n of questions to ask to find the unknown function?

Brute force learning

Finite case

- ISSUE : How many questions with answers YES or NO one has to ask the oracle to find THE function among K functions?
- STRATEGY : Proceed recursively by splitting the set of functions in two groups and asking whether THE function is the first group and removing the group which does not contain the function. This leads to the identification of the desired function with about $\log K$ questions.
- ANSWER : Number of questions $n = \left\lceil \frac{\log K}{\log 2} \right\rceil$
- NB : this quantity represents the number of bits of information characterizing the function in the set of K functions

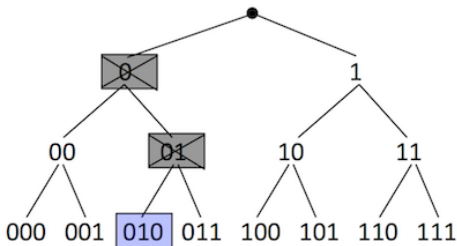
Shannon's Information theory

The origin of the $\log K$

- Related to the entropy of a distribution P in information

theory :
$$H(P) = - \sum_{k=1}^K P(k) \log P(k)$$

- The entropy is the number of bits to encode a collection of K symbols (functions)



From questions to data

Zero error case

- Notations : Domain space \mathcal{X} and label space $\mathcal{Y} = \{0, 1\}$
- ISSUE : How many examples $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are required to find among a *finite* collection (size K) of indicator functions $f : \mathcal{X} \rightarrow \{0, 1\}$ the desired one?
- SAME ANSWER : Number of examples $n = \left\lceil \frac{\log K}{\log 2} \right\rceil$
- STRATEGY : One has to find a vector x_i such that half of the functions take value 1 and the other half take value 0 and ask the oracle whether the desired function takes value 1 or 0 on this vector and discard those functions taking the opposite value. Apply this n times.

Probably approximately correct learning

Zero error case

- REMARK : it may be hard to find such an x_i which splits the collection of functions in two.
- NEW MODEL : Assume X_1, \dots, X_n is an IID sample
- QUESTION : How many examples (X_i, Y_i) are required to find among a finite collection of indicator functions $f : \mathcal{X} \rightarrow \{0, 1\}$ the one that with probability $1 - \delta$ is ε -close to the desired one?
- ANSWER : Number of examples

$$n = \left\lceil \frac{\log K + \log \left(\frac{1}{\delta} \right)}{\varepsilon} \right\rceil$$

Probably approximately correct learning

General case

- ASSUME : among K functions, NONE of them commits zero error on the sample (X_i, Y_i) .
- SAME ISSUE AS BEFORE
- ANSWER : Number of examples on average

$$n = \left\lceil \frac{\log K + \log \left(\frac{1}{\delta} \right)}{\epsilon^2} \right\rceil$$

Same dependency on K , the only change is in the constant.

(Proof coming next)

Finite case (the "log K")

Proposition (Uniform bound for finite classes)

Consider a finite family \mathcal{H} of classifiers. We have, for any $\delta > 0$, with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H}, \quad L(h) \leq \hat{L}_n(h) + \sqrt{\frac{\log |\mathcal{H}| + \log\left(\frac{1}{\delta}\right)}{2n}}$$

Proof relies on : Hoeffding's inequality (see later) + union bound ($\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$)

Chapter 2

Probability inequalities

Historical perspective on probability inequalities

- Kolmogorov, Smirnov (1936) : convergence of empirical cdf to their expectations
- Dvoretzky, Kiefer, Wolfowitz (1956) : nonasymptotic version of Kolmogorov-Smirnov
- Hoeffding (1963) : deviation inequality (average of IID from its expectation)
- Vapnik-Chervonenkis (1968) : equivalent of DKW for general measures (not only 1D on half lines)
- Mc Diarmid (1981) : first concentration inequality
- Massart (1990) : exact constant in DKW
- Talagrand (1996) : new concentration inequalities

Topics : uniform law of large numbers (and central limit theorem), empirical processes, large deviations, convex geometry, high dimensional probability

Reference : book by Boucheron-Lugosi-Massart (2013)

Hoeffding's lemma

Proposition

Consider Z a random variable such that :

- $\mathbb{E}(Z) = 0$
- $Z \in [a, b]$ almost surely

Then, for any $s > 0$, we have :

$$\mathbb{E}(e^{sZ}) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Interpretation : the Laplace transform of bounded random variables exhibits subgaussian behavior.

Hoeffding's inequality

Proposition

Consider Z_1, \dots, Z_n IID over $[0, 1]$ and $\bar{Z}_n = \frac{1}{n} \sum_{i=1}^n Z_i$. We then have, for any $t > 0$

$$\mathbb{P}\{\bar{Z}_n - \mathbb{E}(Z_1) > t\} \leq \exp(-2nt^2)$$

and

$$\mathbb{P}\{\bar{Z}_n - \mathbb{E}(Z_1) < -t\} \leq \exp(-2nt^2)$$

Consequence : This bound implies the strong law of large numbers for bounded random variables (by Borel-Cantelli lemma)

Proof technique : Chernoff's bounding method

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}(Z_1) > t\right) \leq \inf_{s>0} \exp\left(-nst + n \log \mathbb{E}(e^{s(Z_1 - \mathbb{E}(Z_1))})\right)$$

Beyond IID sequences

Definition. (Martingale difference)

Consider $V = (V_1, \dots, V_n, \dots)$ and $Z = (Z_1, \dots, Z_n, \dots)$ two sequences of random variables. We call V a martingale difference sequence wrt Z if, for any n we have :

- V_n is a function of Z_1, \dots, Z_n
- $\mathbb{E}(V_{n+1} \mid Z_1, \dots, Z_n) = 0$

A martingale inequality

Theorem. (Azuma's inequality)

Consider V a a martingale difference sequence wrt Z . Assume that, for any n , there exists U_n a function of Z_1, \dots, Z_{n-1} and $c_n \geq 0$ such that :

$$U_n \leq V_n \leq U_n + c_n$$

We then have, for any $t > 0$

$$\mathbb{P} \left(\sum_{i=1}^n V_i > t \right) \leq \exp \left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right)$$

and

$$\mathbb{P} \left(\sum_{i=1}^n V_i < -t \right) \leq \exp \left(-\frac{2t^2}{\sum_{i=1}^n c_i^2} \right)$$

A basic concentration inequality

Theorem. (McDiarmid's inequality)

Consider Z_1, \dots, Z_n IID. Under a regularity assumption on the function f called the bounded difference assumption with constant c/n , we have, for any $t > 0$

$$\mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) > t) \leq \exp(-2nt^2/c^2)$$

and

$$\mathbb{P}(f(Z_1, \dots, Z_n) - \mathbb{E}(f(Z_1, \dots, Z_n)) < -t) \leq \exp(-2nt^2/c^2)$$

- Here the average of IID random variables is replaced by a general function of these IID variables.
- Take-home message : **Independence is more important/general than averaging**

Bounded difference assumption

- Consider a function f of n variables. We say that f has bounded differences if the variations along each variables are uniformly bounded.
- Here we need to have : for some $c > 0$

$$\sup_{z_1, \dots, z_n, z'_i} |f(z_1, \dots, z_n) - f(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n)| \leq \frac{c}{n}$$

B. Complexity measures : From finite to infinite sets of functions

1. Metric complexities
2. Combinatorial complexities
3. Geometric complexities

Historical perspective

- Kolmogorov (1950's) : developed metric concepts such as covering numbers, metric entropy... in mathematical analysis.
- Vapnik and Chervonenkis (1970's) : discovered combinatorial concepts such as VC entropy, VC dimension and growth function in probability theory.
- Koltchinskii and Panchenko (2000) then Bartlett and Mendelson (2002) : baptized a geometry-related quantity Rademacher complexity which was a variation of gaussian complexity in the continuous case to solve some technical issues in machine learning theory.

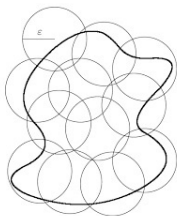
B. Complexity measures

1. Complexity measures based on metric concepts
(from dots to balls)

Covering numbers

Definition

- Consider a general space \mathcal{H} (possibly space of functions) with a metric $\| \cdot \|$
- An ε -cover \mathcal{T} is a set of elements of \mathcal{H} such that for any $h \in \mathcal{H}$ there exists an element $t \in \mathcal{T}$ such that t is ε -close to h (i.e. $\|h - t\| \leq \varepsilon$)



- The covering number $N(\varepsilon)$ is the cardinality of the smallest ε -cover of \mathcal{H}
- The metric entropy of \mathcal{H} is the function $\varepsilon \mapsto \log N(\varepsilon)$

Covering numbers

Example

- Result : for \mathcal{H} being the unit ball in \mathbb{R}^d , we have :

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^d$$

Covering numbers

Upper bound on the error

Result by D. Pollard (1984)

- Notations : n sample size, ℓ loss function
- For bounded loss functions ($\|\ell(\cdot, \cdot)\| \leq M$), we have :

$$\mathbb{P} \left(\sup_{h \in \mathcal{H}} |L(h) - \hat{L}_n(h)| > \varepsilon \right) \leq N \left(\frac{\varepsilon}{8M} \right) \exp \left(-\frac{n\varepsilon^2}{2M^2} \right)$$

- How "tight" this inequality is ?

B. Complexity measures

2. Complexity concepts based on combinatorics

Vapnik-Chervonenkis (VC) entropy

- For a given sample (X_1, \dots, X_n) and for a given $+1/-1$ classifier h , denote by $X_n(h)$ the $+1/-1$ (classification) vector :

$$X_n(h) = (h(X_1), \dots, h(X_n))^T \in \{-1, 1\}^n$$

- For this sample (X_1, \dots, X_n) , denote by $\hat{N}(\mathcal{H})$ the cardinality of such vectors ("**colorings of the data**") induced by the set of functions $h \in \mathcal{H}$ (this set of vectors is sometimes called the *trace* of the set of functions on the sample). Note that there are at most 2^n vectors but can be less than 2^n since some vectors ('colorings') may be unreachable with functions in \mathcal{H} .
- VC entropy :

$$\mathcal{E}(\mathcal{H}) = \mathbb{E}(\log \hat{N}(\mathcal{H}))$$

Sufficient condition for the estimation error to go to zero

- Finite case (reminder) : convergence to zero of the estimation error if

$$\frac{\log |\mathcal{H}|}{n} \rightarrow 0, \quad n \rightarrow \infty$$

- Similar role for the VC entropy : convergence to zero of the estimation error if

$$\frac{\mathcal{E}(\mathcal{H})}{n} = \frac{\mathbb{E}(\log \hat{N}(\mathcal{H}))}{n} \rightarrow 0, \quad n \rightarrow \infty$$

- Questions : are there weaker conditions ? Which sets of functions fulfill such a condition ? What are the rates of convergence ?

VC dimension Definition

- The VC dimension is the largest integer such that there exists a sample of n points in \mathbb{R}^d for which all its "colorings" (separations in +1/-1 classes) can be achieved by elements of \mathcal{H} , *i.e.*

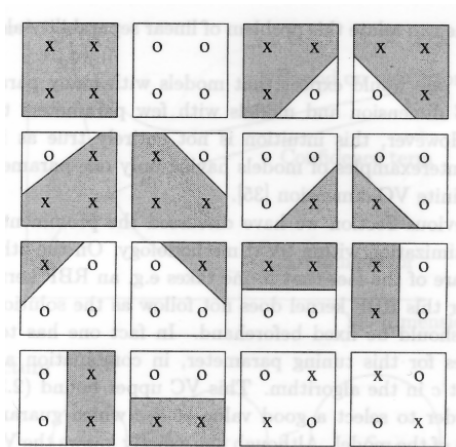
$$V(\mathcal{C}) = \max\{n \text{ integer} : \exists \text{ sample s.t. } |\hat{\mathcal{N}}(\mathcal{H})| = 2^n\}$$

- By comparison to the VC entropy, the VC dimension corresponds to the "worst" sample since the expectation is replaced by a maximum over all possible training samples.

VC dimension Examples

- Halfspaces in \mathbb{R}^d : $V = d + 1$
- Axis-aligned rectangles in \mathbb{R}^2 : $V = 4$
- Just any rectangles in \mathbb{R}^2 : $V = 7$
- Triangles in \mathbb{R}^2 : $V = 7$
- Convex polygons in \mathbb{R}^2 : $V = +\infty$

VC dimension Halfplanes



Observation : Number of parameters is irrelevant

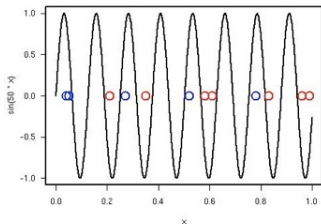
- Set of indicator functions parameterized by a single parameter ω :

$$h(x) = \mathbb{I}\{x : \sin(\omega x) > 0\} , \text{ where } \omega \in [0, 2\pi)$$

- VC dimension of this set is infinite, using :

$$\omega = \frac{1}{2} \left(1 + \sum_{i=1}^n \left(\frac{1 - y_i}{2} \right) 10^i \right)$$

for a set of points $x_j = 2\pi 10^{-j}$



Application : VC bound on classification error

- Assume \mathcal{H} has finite VC dimension V . Then, we have, for any δ , with probability at least $1 - \delta$:

$$L(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L(h) + \sqrt{\frac{2V \log\left(\frac{en}{V}\right)}{n}} + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}$$

- Behavior of the bound wrt V : as VC dimension V increases, the estimation error increases, but at the same time, it is expected that the approximation error goes down since the hypothesis space gets larger.

B. Complexity measures

3. Rademacher complexity : the modern approach to complexity

Learning Theory : Pre vs Post 2000

- Combinatorial complexity concepts (like VC-Dimension) were leading to loose bounds and raised technical difficulties. Cucker and Smale (2001); Evgeniou et al. (1999; 2000); Bartlett et al (1998) resolved various issues.
- Those complexity concepts also accounted for worst-case situations in terms of sample configuration. There was a challenge to develop data-dependent complexity measures.
- Two new approaches started in the late 1990s / early 2000s : **Stability** and **Rademacher complexity**.

Rademacher complexity

Why another concept ?

- The concept was already there in 1968 (Vapnik-Chervonenkis paper) but was not identified as a key quantity except used in an intermediate step of a proof which had to be simplified in later stages of the proof.
- It was rediscovered in 2000 by Koltchinskii and Panchenko and led to neater bounds and theory to encompass all state-of-the-art methods such as SVM, boosting and bagging, as well as neural nets.

A data-dependent view on complexity

- VC entropy is about counting ("coloring") vectors *on average* wrt the training data in the hypercube of \mathbb{R}^n defined by vectors of the form :

$$X_n(h) = (h(X_1), \dots, h(X_n))^T \in \{-1, 1\}^n, \quad \text{for all } h \in \mathcal{H}$$

- Rademacher complexity is about estimating the average of the maximal correlation between a random binary-valued vector and the classification vector for a **fixed** training data set over the class of candidate classifiers.

Definition of Rademacher complexity

- Consider a sample of $D_n = (X_1, \dots, X_n)$ of IID random variables, and a vector of Rademacher random variables : $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ with ε_i 's IID and independent of the training data such that $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$
- Then the Rademacher complexity of the set of functions \mathcal{H} is the sample-dependent quantity :

$$\begin{aligned}\hat{R}_n(\mathcal{H}) &= \mathbb{E} \left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(X_i) \middle| D_n \right) \\ &= \frac{1}{n} \mathbb{E} \left(\sup_{h \in \mathcal{H}} (\varepsilon^T X_n(h)) \middle| D_n \right)\end{aligned}$$

Exercise : Rademacher complexity for linear classes

- Consider a sample x_1, \dots, x_n which are all contained in a ball with radius R
- Denote by \mathcal{H} the hypothesis space of linear functions such that $h(x) = \beta^T x$ where $\|\beta\|_2 \leq M$
- We then have :

$$\hat{R}_n(\mathcal{H}) \leq \frac{MR}{\sqrt{n}}$$

Exercise : concentration of Rademacher complexity

- Set

$$f(X_1, \dots, X_n) = \hat{R}_n(\mathcal{H}) = \mathbb{E} \left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i h(X_i) \middle| D_n \right)$$

- The function f satisfies the bounded differences assumption (why?)
- Therefore, we have, by McDiarmid's inequality, with probability at least $1 - \delta$:

$$\mathbb{E}(\hat{R}_n(\mathcal{H})) \leq \hat{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Application to ERM

Theorem.

Expected error of ERM Let \mathcal{H} be a class of classifiers from \mathbb{R}^d to $\{-1, +1\}$

Consider \hat{h}_n the ERM classifier :

$$\hat{h}_n = \arg \min_{h \in \mathcal{H}} \hat{L}_n(h)$$

Then, with probability at least $1 - \delta$:

$$L(\hat{h}_n) \leq \inf_{g \in \mathcal{H}} L(g) + \mathbb{E}(\hat{R}_n(\mathcal{H})) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$L(\hat{h}_n) \leq \inf_{h \in \mathcal{H}} L(h) + \hat{R}_n(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Link between Rademacher complexity and VC dimension

- Rademacher bounded by VC dimension

$$\hat{R}_n(\mathcal{H}) \leq \sqrt{\frac{2(1 + \log(n/V))}{(n/V)}}$$

- This means that finite VC dimension implies that Rademacher complexity is of the order of $\sqrt{(\log n)/n}$
- However, there are classes with infinite VC dimension which have Rademacher complexity of the same order of magnitude $\sqrt{(\log n)/n}$

Coming next

- Next lectures :
 - Regularization and stability
 - Analysis of ML algorithms