Machine learning lecture III : Results from empirical processes theory

Bartłomiej Błaszczyszyn ([1]), Mohamed Kadhem Karray ([2])

(1) INRIA ; (2) Orange Labs

5 juillet 2022

Introduction

- We present some results from empirical processes theory which are useful for data science.
- These results, together with the Vapnik-Chervonenkis theory (previous lecture) will permit to
  - establish a uniform bound of the deviation of the empirical loss $L_{S^{(n)}}(h)$ from the true loss $\mathcal{L}_Q(h)$ for h within an infinite hypothesis class $\mathcal{H}$.
- We shall
  - show the measurability of the supremum such as $\sup_{h\in\mathcal{H}} |L_{S^{(n)}}(h) - \mathcal{L}_Q(h)|$ (cf. first lecture),
  - give upper bounds for $\mathbf{P}\left(\sup_{h\in\mathcal{H}} |L_{S^{(n)}}(h) - \mathcal{L}_Q(h)| > \varepsilon\right)$.
- This lecture relies on [1].
  - More details, and in particular references and detailed proofs may be found there.
- We are particularly grateful to the authors [5], [6], and [4], who are our first source of inspiration for the present work.

Outline

Empirical processes : Motivation

- Consider a general learning framework as in Lecture I.
  - Remind that the empirical loss is defined as

$$L_{s^{(n)}}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(h, s_i\right)$$

    is the expectation of the $\ell(h, \cdot)$'s with respect to the empirical distribution (which puts a probability mass $1/n$ at each s_i) ;
  - whereas the true loss

$$\mathcal{L}_Q(h) = \mathbf{E}\left[\ell\left(h, Z\right)\right], \quad \text{where } Z \overset{\text{dist.}}{\sim} Q.$$

    is the expectation of $\ell(h, Z)$ with respect to the true distribution $Z \overset{\text{dist.}}{\sim} Q$.
- Controlling the supremum $\sup_{h\in\mathcal{H}} |L_{S^{(n)}}(h) - \mathcal{L}_Q(h)|$ falls in the scope of empirical processes theory.

Empirical processes : Notation

	Machine learning	**Empirical processes**
Data space	$(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$	$(\mathbb{D}, \mathcal{D})$
Learning samples	$S_1, S_2, \ldots$	$X_1, X_2, \ldots$
Hypothesis	h	$f = \ell\,(h, \cdot)$
Data distribution	Q	P
Empirical loss	$L_{S^{(n)}}(h) = \frac{\sum_{i=1}^{n} \ell(h, S_i)}{n}$	$\mathbb{P}_n f = \frac{\sum_{i=1}^{n} f(X_i)}{n}$
True loss	$\mathcal{L}_Q(h) = \int \ell\,(h, \cdot)\,\mathrm{d}Q$	$Pf = \int f \mathrm{d}P$

Empirical processes

- **Definition** : *Empirical measure and process*. Let P be a probability measure on some measurable space $(\mathbb{D}, \mathcal{D})$, let $X_1, X_2, \ldots$ be i.i.d $\mathbb{D}$-valued random variables with common probability distribution P, and let $n \in \mathbb{N}^*$.
  - The n^{th} *empirical measure* associated to P, denoted $\mathbb{P}_n$, is defined by

    $$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

    where δ_x is the Dirac measure at x.
  - Given a collection $\mathcal{F}$ of measurable functions $\mathbb{D} \to \mathbb{R}$, the n^{th} *empirical process* is the real-valued stochastic process $\mathbb{G}_n$ indexed by $\mathcal{F}$ defined by

    $$\mathbb{G}_n f = \sqrt{n} \left(\mathbb{P}_n - P \right) f, \quad f \in \mathcal{F}, \tag{1}$$

    where we use the notation $\mu f = \int f(x) \mu(\mathrm{d}x) = \int f \mathrm{d}\mu$ for Lebesgue integral.

Empirical processes : Law of large numbers and central limit theorem

- **Lemma** : Let P be a probability measure on some measurable space $(\mathbb{D}, \mathcal{D})$, let $\mathbb{P}_n$ and $\mathbb{G}_n$ be the n^{th} associated empirical measure and process respectively ($n \in \mathbb{N}^*$), and let $f : \mathbb{D} \to \mathbb{R}$ be a measurable function.
  - *Law of large numbers* : If Pf exists, then $\mathbb{P}_n f \overset{\text{a.s.}}{\to} Pf$ as $n \to \infty$.
  - *Central limit theorem* : If $Pf^2 < \infty$, then $\mathbb{G}_n f \overset{\text{w}}{\to} \mathcal{N}(0, P(f - Pf)^2)$ as $n \to \infty$.
- **Reminder** : We say that a sequence X_n of real-valued random variables converges *weakly* to a measure μ on $\mathbb{R}$, and write $X_n \overset{\text{w}}{\to} \mu$, if

$$\mathbf{E}\left[f\left(X_n\right)\right] \to \int f \mathrm{d}\mu,$$

for any bounded and continuous function $f : \mathbb{R} \to \mathbb{R}$.

Measurability of the supremum

- **Definition** : *Pointwise separable class of functions*. Let $\mathbb{D}$ be a nonempty set, and $\mathcal{F}$ be a collection of functions $\mathbb{D} \to \mathbb{R}$. We say that $\mathcal{F}$ is *pointwise separable* if there is a countable subcollection $\mathcal{F}_0 \subset \mathcal{F}$ such that every $f \in \mathcal{F}$ *is* the pointwise limit of a sequence f_m in $\mathcal{F}_0$; i.e. $f_m(x) \to f(x)$ for every $x \in \mathbb{D}$.
- **Lemma** : In the context of the above definition, assume that $\mathcal{F}$ is pointwise separable with countable dense subset $\mathcal{F}_0$ (w.r.t pointwise convergence). Let $D(\mathcal{F})$ be the set of all functions $z : \mathcal{F} \to \mathbb{R}$ with the property

$$z(f_m) \to z(f),$$

for every $f \in \mathcal{F}$ and every sequence f_m in $\mathcal{F}_0$ such that $f_m \to f$ pointwise. Then for any $z \in D(\mathcal{F})$,

$$\sup_{f \in \mathcal{F}} z(f) = \sup_{f \in \mathcal{F}_0} z(f).$$

Tail bounds : Measurability of the supremum of the empirical process

- We aim to show the measurability of the supremum $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$ of the empirical process
- **Definition** : *Envelope function*. Let $\mathbb{D}$ be a set, and let $\mathcal{F}$ be a class of functions $\mathbb{D} \to \mathbb{R}$. An *envelope function* of $\mathcal{F}$ is any function $F : \mathbb{D} \to \mathbb{R}_+$ such that $|f(x)| \leq F(x)$, for every $x \in \mathbb{D}$ and $f \in \mathcal{F}$.
- **Lemma** : Assume that $\mathcal{F}$ is pointwise separable and let $\mathcal{F}_0$ be a countable dense subset of $\mathcal{F}$ w.r.t pointwise convergence. Assume moreover that $\mathcal{F}$ has a measurable envelope function F satisfying $PF < \infty$. Then $||\mathbb{G}_n||_{\mathcal{F}}$ is measurable, and

$$||\mathbb{G}_n||_{\mathcal{F}} = ||\mathbb{G}_n||_{\mathcal{F}_0}.$$

- We aim now to derive tail bounds of the supremum $\|\mathbb{G}_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\mathbb{G}_n f|$ of the empirical process.

Tail bounds : Bracketing number

- **Definition** : *Bracketing number*. Let $\mathbb{D}$ be a given set, let $\mathcal{M}$ be the class of all functions $\mathbb{D} \to \mathbb{R}$, let $\varphi : \mathcal{M} \to \bar{\mathbb{R}}_+$, let $\mathcal{F} \subset \mathcal{M}$, and let $\varepsilon \in \mathbb{R}_+^*$.
  - Given two functions $l, u : \mathbb{D} \to \mathbb{R}$, the *bracket* $[l, u]$ is the set of all functions f with $l \leq f \leq u$.
  - An ε-bracket is a bracket $[l, u]$ such that $\varphi(l) < \infty$, $\varphi(u) < \infty$, and $\varphi(u - l) < \varepsilon$.
  - The *bracketing number* $\mathcal{N}_\varphi^{[]}(\varepsilon, \mathcal{F})$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$. (The lower and upper bounds of the ε-brackets are not necessarily in $\mathcal{F}$.)
- **Example** : *Bracketing number w.r.t L^q-norm*. Oftenly, we shall consider a probability space $(\mathbb{D}, \mathcal{D}, P)$, and consider a class $\mathcal{F} \subset \mathcal{L}_{\mathbb{R}}^q(P, \mathbb{D})$ (for some $q \in [1, \infty]$), and φ as the $L^q(P)$-norm. In this case, we shall denote $\mathcal{N}_\varphi^{[]}(\varepsilon, \mathcal{F})$ as $\mathcal{N}_{L^q(P)}^{[]}(\varepsilon, \mathcal{F})$.

Tail probability of the empirical process

- **Theorem** : *Uniformly bounded class of functions*. Assume that $\mathcal{F}$ is pointwise separable and that any $f \in \mathcal{F}$ has range in $[0, 1]$. Assume moreover that for some constants v and K, either

$$\sup_Q \mathcal{N}_{L^2(Q)}(\varepsilon, \mathcal{F}) \leq \left(\frac{K}{\varepsilon}\right)^v, \quad \forall \varepsilon \in \,]0, K[\,,$$

or

$$\mathcal{N}^{[]}_{L^2(P)}(\varepsilon, \mathcal{F}) \leq \left(\frac{K}{\varepsilon}\right)^v, \quad \forall \varepsilon \in \,]0, K[\,,$$

Then $||\mathbb{G}_n||_{\mathcal{F}}$ is measurable and

$$\mathbf{P}\left(\|\mathbb{G}_n\|_{\mathcal{F}} > t\right) \leq \left(\frac{Dt}{\sqrt{v}}\right)^v e^{-2t^2}, \quad \forall t \in \mathbb{R}_+^*,$$

for a constant D that depends only on K.

Tail probability of the empirical process

- **Theorem** : *Class of sets*. Let $\mathcal{C} \subset \mathcal{D}$ and assume that $\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\}$ is pointwise separable. Assume moreover that for some constants v and K, either

$$\sup_Q \mathcal{N}_{L^1(Q)}(\varepsilon, \mathcal{F}) \leq \left(\frac{K}{\varepsilon}\right)^v, \quad \forall \varepsilon \in \left]0, K\right[, \tag{2}$$

or

$$\mathcal{N}^{[]}_{L^1(P)}(\varepsilon, \mathcal{F}) \leq \left(\frac{K}{\varepsilon}\right)^v, \quad \forall \varepsilon \in \left]0, K\right[, \tag{3}$$

Then $||\mathbb{G}_n||_{\mathcal{F}}$ is measurable and

$$\mathbf{P}\left(\|\mathbb{G}_n\|_{\mathcal{F}} > t\right) \leq \frac{D}{t}\left(\frac{Dt^2}{v}\right)^v e^{-2t^2}, \quad \forall t \in \mathbb{R}_+^*,$$

for a constant D that depends only on K.

Tail probability of the empirical process

- **Theorem** : *Class of sets (refinement)*. Let $\mathcal{C} \subset \mathcal{D}$ and assume that $\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\}$ is pointwise separable and satisfies either (2) or (3) for some constants v and K. Assume moreover that for some constants v', w and K',

$$\mathcal{N}_{L^1(P)}(\varepsilon, \mathcal{F}_\delta) \leq K' \delta^w \varepsilon^{-v'}, \quad \text{for every } \delta \geq \varepsilon > 0, \tag{4}$$

where $\mathcal{F}_\delta = \{\mathbf{1}_C : C \in \mathcal{C}, |P(C) - 1/2| \leq \delta\}$. Then $||\mathbb{G}_n||_\mathcal{F}$ is measurable and

$$\mathbf{P}\left(\|\mathbb{G}_n\|_\mathcal{F} > t\right) \leq D t^{2v' - 2w} e^{-2t^2}, \quad \forall t > K\sqrt{w},$$

for a constant D that depends only on K, K', w, v, *and* v'.

Tail probability of the empirical process

- **Corollary** : *Empirical CDF; tail bound*. Let $X_1, X_2, \ldots$ be i.i.d real-valued random variables with common cumulative distribution function F. Let $\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \leq x\}$ be the empirical cumulative distribution function. Then $||\mathbb{F}_n - F||_{\mathbb{R}}$ is measurable and

$$\mathbf{P}\left(||\mathbb{F}_n - F||_{\mathbb{R}} > t\right) \leq De^{-2nt^2}, \quad \forall t \in \mathbb{R}_+^*,$$

  for some universal constant D.
- The result in the above Corollary is due originally to [2, Lemma 2 p.646] and has been refined by [3, Corollary 1 p.1270] who shows that $D = 2$.

Bibliography

[1] B. Błaszczyszyn and M. K. Karray.
Data science : From multivariate statistics to machine and deep learning.
Book in preparation, 2022.

[2] A. Dvoretzky, J. Kiefer, and J. Wolfowitz.
Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator.
Annals of Mathematical Statistics, 27, 1956.

[3] P. Massart.
The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality.
Annals of Probability, 18, 1990.

[4] S. Shalev-Shwartz and S. Ben-David.
Understanding machine learning : From theory to algorithms.
Cambridge University Press, 2014.

[5] M. Talagrand.
Sharper bounds for Gaussian and empirical processes.
The Annals of Probability, 22(1), 1994.

[6] A. W. van der Vaart and J. A. Wellner.
Weak convergence and empirical processes with applications to statistics.
Springer, 1996.