Machine learning
Vapnik-Chervonenkis theory

Claire Boyer

September 16th

From $\sup_{f \in \mathcal{C}}$ to $\sup_{A \in \mathcal{A}}$ I

The goal of this chapter is to go on empirical risk minimization guarantees, when the class $\mathcal{C}$ is not finite.
We are given

- an n-sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, i.i.d. copies of (X, Y),
- a class $\mathcal{C}$ of potential classifiers.

Goal here

Control

$$\sup_{f \in \mathcal{C}} \left| \widehat{\mathcal{R}}_n(g) - \mathcal{R}(g) \right|$$

in a more general framework

From $\sup_{f \in \mathcal{C}}$ to $\sup_{A \in \mathcal{A}}$ II

Notation:

- Let ν be the distribution of (X, Y).
- Let ν_n be the empirical measure associated to $\mathcal{D}_n$, i.e. for all $A \in \mathcal{B}(\mathbb{R}^d \times \{0,1\})$,

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i, Y_i) \in A}.$$

- For any rule $f \in \mathcal{C}$, one can associate the borel set

$$A_f = \left\{ (x, y) \in \mathbb{R}^d \times \{0,1\} : f(x) \neq y \right\}.$$

From $\sup_{f \in \mathcal{C}}$ to $\sup_{A \in \mathcal{A}}$ III

Using this notation, it is easy to see that on the one hand,

$$\mathcal{R}(f) = \mathbb{P}\left(f(X) \neq Y\right) = \nu(A_f),$$

and on the other hand,

$$\hat{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{f(X_i) \neq Y_i} = \nu_n(A_f).$$

$$\left\{ \sup_{f \in \mathcal{C}} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) \right| \right\} = \left\{ \sup_{A \in \mathcal{A}} \left| \nu_n(A) - \nu(A) \right| \right\}. \tag{1}$$

with $\mathcal{A} := \{A_f : f \in \mathcal{C}\}$.

In order to control in probability $\sup_{f \in \mathcal{C}} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) \right|$, one has to understand how the empirical measure behaves on a class of a given measurable sets $\mathcal{A}$.

From $\sup_{f \in \mathcal{C}}$ to $\sup_{A \in \mathcal{A}}$ IV

One can already say that for a fixed $A \in \mathcal{A}$,

$$\nu_n(A) - \nu(A) \to 0, \quad a.s.$$

by the law of large numbers (LLN). Also, if $\mathcal{A}$ is finite and its cardinality bounded by N, a similar study as in the previous chapter would lead to, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)|\right) \geqslant 2Ne^{-2n\varepsilon^2}. \tag{2}$$

In particular, Borel-Cantelli lemma implies that for all law ν,

$$\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| \to 0 \quad a.s.$$

From $\sup_{f \in \mathcal{C}}$ to $\sup_{A \in \mathcal{A}}$ V

However, if $\mathcal{A}$ is not finite, bad phenomena can happen: consider $\mathcal{A}$ to be the set of all Borel set in $\mathbb{R}^d \times \{0, 1\}$, then one can find some laws ν such that

$$\sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| \to 1 \quad a.s.$$

Conclusion

The size of $\mathcal{A}$ has to be controlled. To do so, we will need some new combinatorial tools.

Summary

Preliminary on VC theory I

Let $\mathcal{A}$ be a family of subsets of $\mathbb{R}^p$, with cardinality strictly larger than 1 (not necessarily finite). Given n points $(z_1, \ldots, z_n)$ in $\mathbb{R}^p$, one can define

$$\mathcal{N}_{\mathcal{A}}(z_1, \ldots, z_n) = |\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\}| .$$

In words, $\mathcal{N}_{\mathcal{A}}(z_1, \ldots, z_n)$ represents the number of subsets of $\{z_1, \ldots, z_n\}$ that can be obtained by intersecting $\{z_1, \ldots, z_n\}$ with sets of $\mathcal{A}$.

- $\mathcal{N}_{\mathcal{A}}(z_1, \ldots, z_n) \leqslant 2^n$,
- when $\mathcal{N}_{\mathcal{A}}(z_1, \ldots, z_n) = 2^n$, we say that $\mathcal{A}$ shatters the set $\{z_1, \ldots, z_n\}$.

Preliminary on VC theory II

In order to be free from an arbitrary set $\{z_1, \ldots, z_n\}$, define:

Definition (Shattering coefficient)

One can define the n-th shattering coefficient of $\mathcal{A}$ by

$$S_{\mathcal{A}}(n) = \max_{z_1,\ldots,z_n} \left|\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\}\right|.$$

The n-th shattering coefficient is the largest number of subsets of any set of n points that can be formed by intersecting it with the sets in collection $\mathcal{A}$.

- $S_{\mathcal{A}}(n) \leqslant 2^n$, since $\{\{z_1, \ldots, z_n\} \cap A : A \in \mathcal{A}\} \subset P(\{z_1, \ldots, z_n\})$,
- $S_{\mathcal{A}}(1) = 2$, (why?)
- $S_{\mathcal{A}}(k) < 2^k$ for some $k > 1 \Longrightarrow S_{\mathcal{A}}(n) < 2^n$ for all $n \geqslant k$.

Definition (VC dimension)

Let $\mathcal{A}$ be a collection of subsets of $\mathbb{R}^p$. The Vapnik-Chervonenkis dimension of $\mathcal{A}$, denoted by $VC_{\mathcal{A}}$, is the largest integer $n_0 \geqslant 1$ such that $S_{\mathcal{A}}(n_0) = 2^{n_0}$. If $S_{\mathcal{A}}(n) = 2^n$ for all $n \geqslant 1$ then $VC_{\mathcal{A}} = +\infty$.

- It measures in some sense the "size" of the collection $\mathcal{A}$
- It generalizes the notion of cardinality
- Important combinatorial concept, which is central in statistical learning theory
- Note that, by definition, if $VC_{\mathcal{A}} = v$, there exists a set of size v that can be fully shattered. But, this does not imply that all sets of size v or less are fully shattered, in fact, this is typically not the case.

To compute the VC dimension

- Show a lower bound for its value and then a matching upper bound.
- To give a lower bound v on the VC dimension, it suffices to show that a set S of cardinality v can be shattered by $\mathcal{A}$.
- To give an upper bound, we need to prove that no set S of cardinality $v+1$ can be shattered by $\mathcal{A}$, which is typically more difficult.

Here are some examples (proofs are left to the reader):

1. Assume that $|\mathcal{A}| < +\infty$. Trivially $S_{\mathcal{A}}(n) \leqslant |\mathcal{A}|$. By the definition of the VC dimension, one has $S_{\mathcal{A}}(VC_{\mathcal{A}}) = 2^{VC_{\mathcal{A}}}$, then
$$VC_{\mathcal{A}} \leqslant \ln_2 |\mathcal{A}|.$$

Preliminary on VC theory V

2. In dimension $p = 1$,
   - if $\mathcal{A} = \{]-\infty; a] : a \in \mathbb{R}\}$, then
   $$VC_{\mathcal{A}} = 1$$
   with $S_{\mathcal{A}}(n) = n + 1$;
   - if $\mathcal{A} = \left\{[a, b] : (a, b) \in \mathbb{R}^2\right\}$, then
   $$VC_{\mathcal{A}} = 2$$
   with $S_{\mathcal{A}}(n) = n(n+1)/2 + 1$.
3. In dimension p,
   - if $\mathcal{A} = \{]-\infty; a_1] \times \ldots \times]-\infty; a_p] : (a_1, \ldots, a_p) \in \mathbb{R}^p\}$, then
   $$VC_{\mathcal{A}} = d;$$
   - if $\mathcal{A} = \{\text{rectangles of } \mathbb{R}^p\}$, then
   $$VC_{\mathcal{A}} = 2p;$$

- if $\mathcal{A} = \{\text{convex polygons of } \mathbb{R}^2\}$, then

$$VC_{\mathcal{A}} = +\infty.$$

4. **Important!** Let $\mathcal{F}$ be a vector space of functions from $\mathbb{R}^p$ to $\mathbb{R}$, with finite dimension. If

$$\mathcal{A} = \{\{x \in \mathbb{R}^p : f(x) \geqslant 0\} : f \in \mathcal{F}\},$$

then

$$VC_{\mathcal{A}} \leqslant \dim \mathcal{F}.$$

In particular, if $\mathcal{A}$ is the collection of half linear spaces, i.e. subsets of $\mathbb{R}^p$ of the form $\{x \in \mathbb{R}^p : a^T x + b \geqslant 0 : a \in \mathbb{R}^p, b \in \mathbb{R}\}$, then

$$VC_{\mathcal{A}} \leqslant p + 1.$$

Exercise: Show the last point.

Summary

Statement I

Theorem (Vapnik-Chervonenkis)

Let $Z_1, \ldots, Z_n$, independent random variables, of same law ν in $\mathbb{R}^p$, and ν_n its empirical version. Then, for all Borelian collection $\mathcal{A} \subset \mathbb{R}^p$ and for all $\varepsilon > 0$, one has

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \varepsilon\right) \leqslant 8 S_{\mathcal{A}}(n) e^{-n\varepsilon^2/32}.$$

(i) The bound is universal: it does not depend on the law ν.

(ii) VC thm provides a generalization of the inequality (2):

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)|\right) \geqslant 2Ne^{-2n\varepsilon^2}.$$

when $\mathcal{A}$ is finite. Roughly speaking, the cardinality of $\mathcal{A}$ is replaced by the shattering coefficient.

Statement II

(iii) By Borel-Cantelli lemma, one has

$$\sup_{A\in\mathcal{A}} |\nu_n(a) - \nu(A)| \to 0 \quad a.s.$$

provided that the series which terms is $S_{\mathcal{A}}(n)e^{-n\varepsilon^2/32}$ is convergent, which is the case if $|\mathcal{A}| < +\infty$ or if $S_{\mathcal{A}}(n)$ is a polynomial in n. Yet, one cannot conclude if $S_{\mathcal{A}}(n) = 2^n$ for all n (or equivalently if $VC_{\mathcal{A}} = +\infty$).

(iv) The proof of Theorem 3 is not complicated (see the blackboard), it involves key arguments for statistical learning, that we are going to detail. In words, the idea is to put the supremum in front of the probability, by combinatorial arguments on the collection $\mathcal{A}$, described by $S_{\mathcal{A}}(n)$.

One aside application I

Consider an n-sample $Z_1, \ldots, Z_n$ i.i.d. r.v. taking value in $\mathbb{R}$ of same distribution ν. Take $\mathcal{A} = \{]-\infty, a] : a \in \mathbb{R}\}$. Then, for $A =]-\infty, a] \in \mathcal{A}$, $\nu(A) = F(a)$ and $\nu_n(A) = F_n(a)$ where F (resp. F_n) is the distribution function (resp. the empirical distribution function) associated to ν (resp. $Z_1, \ldots, Z_n$). Note that $S_{\mathcal{A}}(n) = n + 1$, then by Vapnik-Chervonenkis theorem, one has

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)| > \varepsilon\right) = \mathbb{P}\left(\sup_{z\in\mathbb{R}} |F_n(z) - F(z)| > \varepsilon\right) \leqslant 8(n+1)e^{-n\varepsilon^2/32}.$$

Borel-Cantelli lemma implies that

$$\sup_{z\in\mathbb{R}} |F_n(z) - F(z)| \to 0 \quad a.s.$$

meaning that the empirical distribution function converges to the distribution function almost surely, in the sense of the function

One aside application II

uniform convergence. This theorem is called the Glivenko-Cantelli theorem.

Summary

Let us admit the following result.

Lemma (Sauer's lemma)

Let $\mathcal{A}$ be a collection of sets with a finite VC dimension $VC_{\mathcal{A}}$. Then, for all $n \geqslant 1$,

$$S_{\mathcal{A}}(n) \leqslant \sum_{i=1}^{VC_{\mathcal{A}}} \binom{n}{i}.$$

Corollary

Let $\mathcal{A}$ be a collection of sets with a finite VC dimension $VC_{\mathcal{A}}$. Then, for all $n \geqslant 1$,

$$S_{\mathcal{A}}(n) \leqslant (n+1)^{VC_{\mathcal{A}}}.$$

Proof

$$(n+1)^{VC_{\mathcal{A}}} = \sum_{i=1}^{VC_{\mathcal{A}}} \binom{VC_{\mathcal{A}}}{i} n^i = \sum_{i=1}^{VC_{\mathcal{A}}} \frac{n^i\, VC_{\mathcal{A}}!}{i!(VC_{\mathcal{A}}-i)!} \geqslant \sum_{i=0}^{VC_{\mathcal{A}}} \frac{n^i}{i!} \geqslant \sum_{i=0}^{VC_{\mathcal{A}}} \binom{n}{i}$$
$$\geqslant S_{\mathcal{A}}(n).$$

The previous corollary shows that

- either, $VC_{\mathcal{A}} = +\infty$, and then $S_{\mathcal{A}}(n) = 2^n$;
- or, $VC_{\mathcal{A}} < +\infty$, and then, $S_{\mathcal{A}}(n) \leqslant (n+1)^{VC_{\mathcal{A}}}$.

Combinatorial aspects III

There is NO intermediate case where for instance $S_{\mathcal{A}}(n) \sim 2^{\sqrt{n}}$.
Combining VC theorem, Lemma ($\mathbb{P}$ to $\mathbb{E}$) and the last lemma, for any collection $\mathcal{A} \subset \mathbb{R}^p$ of measurable sets with a finite VC dimension, one can write

$$\mathbb{E}\left[\sup_{A\in\mathcal{A}} |\nu_n(A) - \nu(A)|\right] \leqslant 8\sqrt{\frac{\ln(8eS_{\mathcal{A}}(n))}{2n}} \leqslant 8\sqrt{\frac{VC_{\mathcal{A}}\ln(n+1)+4}{2n}}$$
$$= O\left(\sqrt{\frac{VC_{\mathcal{A}}\log(n)}{n}}\right).$$

Note that we could get rid of the log term by using chaining techniques, but this is beyond the scope of this lecture.

Summary

ERM & estimation error I

Let us go back to the supervised classification problem. Consider an n-sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. copies of (X, Y). Let $\mathcal{C}$ be a collection of potential decision rules. By denoting $f_n^\star$ the minimizer of the empirical risk over $\mathcal{C}$, we know that

$$\mathcal{R}(f_n^\star) - \inf_{f \in \mathcal{C}} \mathcal{R}(f) \leqslant 2 \sup_{f \in \mathcal{C}} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) \right|,$$

and

$$\left\{ \sup_{f \in \mathcal{C}} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) \right| \right\} = \left\{ \sup_{A \in \mathcal{A}} |\nu_n(A) - \nu(A)| \right\},$$

with $\mathcal{A} = \{A_f : f \in \mathcal{C}\}$ and

$$A_f = \left\{ (x, y) \in \mathbb{R}^d \times \{0, 1\} : f(x) \neq y \right\}.$$

Having the VC theorem in mind,

- we know that the shattering coefficient is going to play a central role for the control $\sup_{f \in \mathcal{C}} \left| \widehat{\mathcal{R}}_n(f) - \mathcal{R}(f) \right|$.
- But, the collection $\mathcal{A}$ including subsets of $\mathbb{R}^d \times \{0, 1\}$ is too complex to be analyzed with combinatorial tools.

Proposition

Let $\mathcal{A} = \{A_f : f \in \mathcal{C}\}$ *and* $\bar{\mathcal{A}} = \left\{ \left\{ x \in \mathbb{R}^d : f(x) = 1 \right\} : f \in \mathcal{C} \right\}$. *Then for all* $n \geqslant 1$,

$$S_{\mathcal{A}}(n) = S_{\bar{\mathcal{A}}}(n) \quad \text{and} \quad VC_{\mathcal{A}} = VC_{\bar{\mathcal{A}}}.$$

Now, we can establish the following result by combining the previous proposition, the technical Lemma ($\mathbb{P}$ to $\mathbb{E}$) and the VC theorem.

Theorem

For all $n \geqslant 1$,

$$\mathbb{P}\left(\left|\mathcal{R}(f_n^\star) - \inf_{f\in\mathcal{C}} \mathcal{R}(f)\right| > \varepsilon\right) \leqslant 8 S_{\bar{\mathcal{A}}}(n) e^{-n\varepsilon^2/128},$$

and

$$\mathbb{E}\mathcal{R}(f_n^\star) - \inf_{f\in\mathcal{C}} \mathcal{R}(f) \leqslant 16\sqrt{\frac{\ln(8eS_{\bar{\mathcal{A}}}(n))}{2n}}.$$

Borel-Cantelli lemma leads to

$$\mathcal{R}(f_n^\star) - \inf_{f\in\mathcal{C}} \mathcal{R}(f) \to 0 \quad a.s.$$

if the series of general term $S_{\bar{\mathcal{A}}}(n)e^{-n\varepsilon^2/128}$ is convergent. This is the case when $VC_{\bar{\mathcal{A}}}$ (or $VC_{\mathcal{A}}$) is finite, since $S_{\bar{\mathcal{A}}}$ has a polynomial growth in n.

ERM & estimation error IV

Take-home message

$VC_{\bar{\mathcal{A}}} < +\infty$ is a sufficient condition to ensure the a.s. convergence of the estimation term to 0 and in this case

$$\mathbb{E}\mathcal{R}(f_n^\star) - \inf_{f \in \mathcal{C}} \mathcal{R}(f) = O\left(\sqrt{\frac{VC_{\bar{\mathcal{A}}} ln(n)}{n}}\right).$$

Examples I

1. **Linear classification** Let $x = (x^{(1)}, \ldots, x^{(d)})$, and consider rules of the form

$$f(x) = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{d} a_j x^{(j)} + a_0 > 0, \\ 0 & \textbf{otherwise}. \end{cases}$$

with $(a_0, \ldots, a_d) \in \mathbb{R}^{d+1}$ be a vector of parameters. In this case,

$$\bar{\mathcal{A}} \subset \left\{ \{x \in \mathbb{R}^d : a^T x + a_0 \geqslant 0 \right\},$$

and given the properties of VC dimension,

$$VC_{\bar{\mathcal{A}}} \leqslant d + 1.$$

Then,

Examples II

$$\mathbb{P}\left(\left|\mathcal{R}(f_n^\star) - \inf_{f\in\mathcal{C}} \mathcal{R}(f)\right| > \varepsilon\right) \leqslant 8(n+1)^{d+1} e^{-n\varepsilon^2/128},$$

and

$$\mathcal{R}(f_n^\star) - \inf_{f\in\mathcal{C}} \mathcal{R}(f) \to 0 \quad a.s.$$

Examples III

2. **Classification by closed balls** The collection $\mathcal{C}$ is composed of indicator functions of closed balls in $\mathbb{R}^d$. Then,

$$\bar{\mathcal{A}} = \left\{ \left\{ x \in \mathbb{R}^d : \sum_{j=1}^{d} |x^{(j)} - a_j|^2 \leqslant a_0 \right\} : (a_0, \ldots, a_d) \in \mathbb{R}^{d+1} \right\}.$$

Noticing that

$$a_0 - \sum_{j=1}^{d} |x^{(j)} - a_j|^2 = a_0 - \sum_{j=1}^{d} (x^{(j)})^2 + 2 \sum_{j=1}^{d} x^{(j)} a_j - \sum_{j=1}^{d} a_j^2,$$

then $\bar{\mathcal{A}}$ is included in a collection of sets $\left\{ \{x \in \mathbb{R}^d : g(x) \geqslant 0\} : g \ in \mathcal{G} \right\}$ with $\mathcal{G}$ a vector space of dimension $d + 2$. Similarly, one can conclude that

$$\mathcal{R}(f_n^\star) - \inf_{f \in \mathcal{C}} \mathcal{R}(f) \to 0 \quad a.s.$$

Examples IV

3. **Classification by convex sets** Here, $\bar{\mathcal{A}}$ is the collection of all convex polygons in $\mathbb{R}^2$, for which we have seen that $VC_{\bar{\mathcal{A}}} = +\infty$. This collection of sets is too "large" so that the estimation error cannot be controlled by the VC theory.
4. **Generalized linear classification** In $\mathbb{R}^d$, consider a fixed number $d^\star$ of measurable functions $\psi_1, \dots \psi_{d^\star}$ of $\mathbb{R}^d \to \mathbb{R}$. The considered classification rules are

$$f(x) = \begin{cases} 1 & \text{if} \quad \sum_{j=1}^{d} a_j \psi_j(x) + a_0 > 0, \\ 0 & \textbf{otherwise}. \end{cases}$$

When $\psi_j(x) = x^{(j)}$ this is again standard linear classification

Examples V

rules. If the $(\psi_j)_j$'s are coordinate functions and product of these coordinates, one can see that $\bar{\mathcal{A}}$ is contained in

$$\left\{ a_0 + \sum_{j=1}^{d} a_j x^{(j)} + \sum_{j=1}^{d} b_j (x^{(j)})^2 + \sum_{1 \leqslant j_1 < j_2 \leqslant d} c_{j_1} c_{j_2} x^{(j_1)} x^{(j_2)} \geqslant 0 \right\}.$$

then $d^\star = 2d + \frac{d(d-1)}{2}$, $VC_{\bar{\mathcal{A}}} \leqslant d^\star + 1$ and

$$\mathcal{R}(f_n^\star) - \inf_{f \in \mathcal{C}} \mathcal{R}(f) \to 0 \quad a.s.$$