Machine Learning 5: VC dimension, Sauer's Lemma, Fundamental Theorem of Statistical Learning

Master 2 Computer Science

Aurélien Garivier

2018-2019

UNIVERSITÉ DE LYON
ENS DE LYON

Table of contents

VC dimension and Sauer's lemma

Shattering

Definition

Let $\mathcal{H}$ be a class of functions $\mathcal{X} \to \{0,1\}$ and let $C = \{c_1, \ldots, c_m\} \subset \mathcal{X}$. The *restriction* of $\mathcal{H}$ to C is the set of functions $C \to \{0,1\}$ that can be derived from $\mathcal{H}$:

$$\mathcal{H}_C = \Big\{(c_1, \ldots, c_m) \to \big(h(c_1), \ldots, h(c_m)\big) : h \in \mathcal{H}\Big\} .$$

Shattering

A hypothesis class $\mathcal{H}$ *shatters* a finite set $C \subset \mathcal{X}$ if $\mathcal{H}_C = \{0,1\}^C$.

Example:

- $\mathcal{H} = \big\{h_a : a \in \mathbb{R}\big\}$.
- $\mathcal{H}^2_{\mathrm{rec}} = \big\{h_{(a_1,b_1,a_2,b_2)} : a_1 \le b_1 \text{ and } a_2 \le b_2\big\}$ where

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \le x_1 \le b_1 \text{ and } a_2 \le x_2 \le b_2 \,; \\ 0 & \text{otherwise} . \end{cases}$$

VC dimension

Definition

The *Vapnik Chervonenkis dimension* $\mathrm{VCdim}(\mathcal{H})$ of a hypothesis class $\mathcal{H}$ is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathrm{VCdim}(\mathcal{H}) = \infty$.

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then $\mathcal{H}$ is not PAC-learnable.

Proof: for every training size m, there exists a set C of size $2m$ that is shattered by $\mathcal{H}$. By the NFL theorem, for every learning algorithm A there exists a probability distribution D over $\mathcal{X} \times \{0, 1\}$ such that $L_D(h) = 0$ but with probability at least $1/7$ over the training set, we have $L_D\big(A(S)\big) \geq 1/8$.

Fundamental theorem of PAC learning

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Then the following propositions are equivalent:

1. $\mathcal{H}$ has the uniform convergence property,
2. any ERM ruel is a successful agnostic PAC learner for $\mathcal{H}$,
3. $\mathcal{H}$ is agnostic PAC learnable,
4. $\mathcal{H}$ is PAC learnable,
5. any ERM rule is a sucessful PAC learner for $\mathcal{H}$,
6. $\mathcal{H}$ has finite VC-dimension.

Fundamental theorem of PAC learning (quantitative version)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function of $0 - 1$ loss. Assume thatVCdim$(\mathcal{H}) < \infty$. Then there exist constants C_1, C_2 such that:

1. $\mathcal{H}$ has the uniform convergence property with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2} \ ,$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log(1/\delta)}{\epsilon^2} \ ,$$

3. $\mathcal{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon} \ .$$

Sauer's lemma

Definition

Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as the maximal number of different functions that can be obtained by restricting $\mathcal{H}$ to a set of size m:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset X : |C| = m} \left| \mathcal{H}_C \right| .$$

Note: if $\mathrm{VCdim}(\mathcal{H}) = d$, then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$.

Sauer's lemma

Let $\mathcal{H}$ be a hypothesis class with $\mathrm{VCdim}(\mathcal{H}) \leq d < \infty$. Then, for all $m \geq d$,

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d} \right)^d .$$

Proof of Sauer's lemma 1/2

In fact we prove the stronger claim:

$$\left|\mathcal{H}_c\right| \leq \left|\{B \subset C : \mathcal{H} \text{ shatters } B\}\right| \leq \sum_{i=0}^{d} \binom{m}{i} .$$

where the last inequality holds since no set of size larger than d is shattered by $\mathcal{H}$. The proof is by induction.

m=1: The empty set is always considered to be shattered by $\mathcal{H}$. Hence, either $|\mathcal{H}(C)| = 1$ and $d = 0$, inequality $1 \leq 1$, or $d \geq 1$ and the inequality is $2 \leq 2$.

Induction: Let $C = \{c_1, \ldots, c_m\}$, and let $C' = \{c_2, \ldots, c_m\}$. We note functions like vectors, and we define

$$Y_0 = \left\{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ or } (1, y_2, \ldots, y_m) \in \mathcal{H}_C\right\}, \quad \text{and}$$

$$Y_1 = \left\{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_C \text{ and } (1, y_2, \ldots, y_m) \in \mathcal{H}_C\right\} .$$

Then $|\mathcal{H}_C| = |Y_0| + |Y_1|$. Moreover, $Y_0 = \mathcal{H}_{C'}$ and hence by the induction hypothesis:

$$|Y_0| \leq |\mathcal{H}_{C'}| \leq \left|\{B \subset C' : \mathcal{H} \text{ shatters } B\}\right| = \left|\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}\right|$$

Next, define

$$\mathcal{H}' = \left\{ h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } h'(c) = \begin{cases} 1 - h(c) \text{ if } c = c_1 \\ h(c) \text{ otherwise} \end{cases} \right\}$$

Note that $\mathcal{H}'$ shatters $B \subset C'$ iff $\mathcal{H}'$ shatters $B \cup \{c_1\}$, and that $Y_1 = \mathcal{H}'_{C'}$. Hence, by the induction hypothesis,

$$|Y_1| = |\mathcal{H}'_{C'}| \leq \left|\{B \subset C' : \mathcal{H}' \text{ shatters } B\}\right| = \left|\{B \subset C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}\right|$$

$$= \left|\{B \subset C : c_1 \in B \text{ and } \mathcal{H}' \text{ shatters } B\}\right| \leq \left|\{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}\right| .$$

Overall,

$$|\mathcal{H}_C| = |Y_0| + |Y_1| \leq \left|\{B \subset C : c_1 \notin B \text{ and } \mathcal{H} \text{ shatters } B\}\right| + \left|\{B \subset C : c_1 \in B \text{ and } \mathcal{H} \text{ shatters } B\}\right| = \left|\{B \subset C : \mathcal{H} \text{ shatters } B\}\right| .$$

Proof of Sauer's lemma 2/2

For the last inequality, one may observe that if $m \geq 2d$, defining $N \sim \mathcal{B}(m, 1/2)$, Chernoff's inequality and inequality $\log(u) \geq (u-1)/u$ yield

$$\begin{aligned}
-\log \mathbb{P}(N \leq d) \geq m\,\mathrm{kl}\left(\frac{d}{m}, \frac{1}{2}\right) &\geq d \log \frac{2d}{m} + (m-d) \log \frac{2(m-d)}{m} \\
&\geq m \log(2) + d \log \frac{d}{m} + (m-d) \frac{-d/m}{(m-d)/m} \\
&= m \log(2) + d \log \frac{d}{em} \ ,
\end{aligned}$$

and hence

$$\sum_{i=0}^{d} \binom{m}{i} = 2^d \mathbb{P}(N \leq d) \leq \exp\left(-d \log \frac{d}{em}\right) = \left(\frac{em}{d}\right)^d \ .$$

Besides, for the case $d \leq m \leq 2d$, the inequality is obvious since $(em/d)^d \geq 2^m$: indeed, function $f : x \mapsto -x \log(x/e)$ is increasing on $[0, 1]$, and hence for all $d \leq m \leq 2d$:

$$\frac{d}{m} \log \frac{em}{d} = f(d/m) \geq f(1/2) = \frac{1}{2} \log(2e) \geq \log(2) \ ,$$

which implies

$$\left(\frac{em}{d}\right)^d = \exp\left(d \log \frac{em}{d}\right) \geq \exp(m \log(2)) = 2^m \ .$$

Alternately, you may simply observe that for all $m \geq d$,

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \leq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} = \left(1 + \frac{d}{m}\right)^m \leq e^d \ .$$

Finite VC dimension implies Uniform Convergence

Finite VC dimension implies Uniform Convergence

Theorem

Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then, for every distribution D dans for every $\delta \in (0,1)$, with probability at least $1-\delta$ over the choice of the sample $S \sim D^{\otimes m}$ we have

$$\sup_{h \in \mathcal{H}} \left| L_D(h) - L_S(h) \right| \leq \frac{1 + \sqrt{\log\left(\tau_{\mathcal{H}}(2m)\right)}}{\delta \sqrt{m/2}} \,.$$

Note: this result is sufficient to prove that finite VC-dim $\implies$ learnable, but the dependency in δ is not correct at all: roughly speaking, the factor $1/\delta$ can be replaced by $\log(1/\delta)$.

Proof: symmetrization and Rademacher complexity (1/2)

We consider the 0-1 loss, or any $[0,1]-$valued loss. Observe that $L_D(h) = \mathbb{E}[L_{S'}(h)]$ where $S' = z_1', \dots, z_m'$ is another iid sample of D. Hence,

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|L_D(h) - L_S(h)\right|\right] &= \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\right] \leq \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left|\mathbb{E}_{S'}\left[L_{S'}(h) - L_S(h)\right]\right|\right] \\
&\leq \mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\mathbb{E}_{S'}\left[\left|L_{S'}(h) - L_S(h)\right|\right]\right] \leq \mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}}\left|L_{S'}(h) - L_S(h)\right|\right]\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \ell(h,z_i') - \ell(h,z_i)\right|\right] \\
&= \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h,z_i') - \ell(h,z_i)\big)\right|\right] \quad \text{for all } \sigma\in\{\pm1\}^m \\
&= \mathbb{E}_\Sigma\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \Sigma_i\big(\ell(h,z_i') - \ell(h,z_i)\big)\right|\right] \quad \text{if } \Sigma\sim\mathcal{U}\big(\{\pm1\}^m\big) \\
&= \mathbb{E}_{S,S'}\mathbb{E}_\Sigma\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \Sigma_i\big(\ell(h,z_i') - \ell(h,z_i)\big)\right|\right] .
\end{aligned}$$

Now, for every S, S', let $C = C_{S,S'}$ be the instances appearing in S and S'. Then

$$\sup_{h\in\mathcal{H}}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h,z_i') - \ell(h,z_i)\big)\right| = \max_{h\in\mathcal{H}_C}\frac{1}{m}\left|\sum_{i=1}^m \sigma_i\big(\ell(h,z_i') - \ell(h,z_i)\big)\right| .$$

Proof: symmetrization and Rademacher complexity (2/2)

Moreover, for every $h \in \mathcal{H}_C$ let $Z_h = \frac{1}{m}\sum_{i=1}^m \Sigma_i\big(\ell(h, z_i') - \ell(h, z_i)\big)$. Then $\mathbb{E}_\Sigma[Z_h] = 0$, each summand belongs to $[-1, 1]$ and by Hoeffding's inequality, for every $\epsilon > 0$:

$$\mathbb{P}_\Sigma\big[|Z_h| \geq \epsilon\big] \leq 2\exp\left(-\frac{m\epsilon^2}{2}\right) .$$

Hence, by the union bound,

$$\mathbb{P}_\Sigma\big[\max_{h \in \mathcal{H}_C} |Z_h| \geq \epsilon\big] \leq 2\big|\mathcal{H}_C\big| \exp\left(-\frac{m\epsilon^2}{2}\right) .$$

The following lemma permits to deduce that

$$\mathbb{E}_\Sigma\left[\max_{h \in \mathcal{H}_C} Z_h|\right] \leq \frac{1 + \sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{m/2}} \leq \frac{1 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{m/2}} .$$

Hence,

$$\mathbb{E}_S\left[\sup_{h \in \mathcal{H}} \big|L_D(h) - L_S(h)\big|\right] \leq \mathbb{E}_{S,S'}\mathbb{E}_\Sigma\left[\sup_{h \in \mathcal{H}} \frac{1}{m}\left|\sum_{i=1}^m \Sigma_i\big(\ell(h, z_i') - \ell(h, z_i)\big)\right|\right] \leq \frac{1 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\sqrt{m/2}} ,$$

and we conclude by using Markov's inequality (poor idea! Better: McDiarmid's inequality).

Technical Lemma

Lemma

Let $a > 0$, $b > 1$, and let Z be a real-valued random variable such that for all $t \geq 0$, $\mathbb{P}(Z \geq t) \leq 2b \exp\left(-\frac{t^2}{a^2}\right)$. Then

$$\mathbb{E}[Z] \leq a\left(\sqrt{\log(b)} + \frac{1}{\sqrt{\log(b)}}\right) \ .$$

Proof:

$$\begin{aligned}
\mathbb{E}[Z] &\leq \int_0^\infty \mathbb{P}(Z \geq t) dt \leq a\sqrt{\log(b)} + \int_{a\sqrt{\log(b)}}^\infty 2b \exp\left(-\frac{t^2}{a^2}\right) \\
&\leq a\sqrt{\log(b)} + 2b \int_{a\sqrt{\log(b)}}^\infty \frac{t}{a\sqrt{\log(b)}} \exp\left(-\frac{t^2}{a^2}\right) \\
&= a\sqrt{\log(b)} + \frac{2b}{a\sqrt{\log(b)}} \times \frac{a^2}{2} \exp\left(-\frac{\left(a\sqrt{\log(b)}\right)^2}{a^2}\right) \\
&= a\sqrt{\log(b)} + \frac{a}{\sqrt{\log(b)}} \ .
\end{aligned}$$

NB: cutting at $a\sqrt{\log(2b)}$ gives a better but less nice inequality for our use.

Finite VC-dimension implies learnability

Application: Finite VC-dim classes are agnostically learnable

It suffices to prove that finite VC-dim implies the uniform convergence property. From Sauer's lemma, for all $m \geq d/2$ we have $\tau_{\mathcal{H}}(2m) \leq (2em/d)^d$. With the previous theorem, this yields that with probability at least $1-\delta$:

$$\sup_{h\in\mathcal{H}} \left| L_D(h) - L_S(h) \right| \frac{1+\sqrt{d\log\left(2em/d\right)}}{\delta\sqrt{m/2}} \leq \frac{1}{\delta}\sqrt{\frac{8d\log(2em/d)}{m}}$$

as soon as $\sqrt{d\log\left(2em/d\right)} \geq 1$. To ensure that this is at most ϵ, one may choose

$$m \geq \frac{8d\log(m)}{(\delta\epsilon)^2} + \frac{8d\log(2e/d)}{(\delta\epsilon)^2}\ .$$

By the following lemma, it is sufficient that

$$m \geq \frac{32d\log\left(\frac{4d}{(\delta\epsilon)^2}\right)}{(\delta\epsilon)^2} + \frac{16d\log\left(\frac{2e}{d}\right)}{(\delta\epsilon)^2}\ .$$

Technical Lemma

Lemma

Let $a > 0$. Then

$$x \geq 2a\log(a) \quad \Longrightarrow \quad x \geq a\log(x)\ .$$

Proof: For $a \leq e$, true for every $x > 0$. Otherwise, for $a \geq \sqrt{e}$ we have $2a\log(a) \geq a$ and thus for every $t \geq 2a\log(a)$, as $f : t \mapsto t - a\log(t)$ is increasing on $[a, \infty)$, $f(t) \geq f(2a\log(a)) = a\log(a) - a\log(2\log(a)) \geq 0$, since for every $a > 0$ it holds that $a \geq 2\log(a)$.

Lemma

Let $a \geq 1$, $b > 0$. Then

$$x \geq 4a\log(2a) + 2b \quad \Longrightarrow \quad x \geq a\log(x) + b\ .$$

Proof: It suffices to check that $x \geq 2a\log(x)$ (given by the above lemma) and that $x \geq 2b$ (obvious since $4a\log(2a) \geq 0$).