Algorithmic Foundations of Learning

Lecture 5
Covering Numbers Bounds for Rademacher Complexity. Chaining

Patrick Rebeschini

Department of Statistics
University of Oxford

Recap: Binary Classification

Only a finite number of elements in $\mathcal{A}$ matter: those giving different labels

Growth function (Definition 4.2)

$$\tau_{\mathcal{A}}(n) := \sup_{x_1,\ldots,x_n \in \mathbb{R}^d} |\mathcal{A} \circ \{x_1,\ldots,x_n\}|$$

(Proposition 4.3)

$$\texttt{Rad}(\mathcal{A} \circ \{x_1,\ldots,x_n\}) \le \sqrt{\frac{2\log \tau_{\mathcal{A}}(n)}{n}}$$

(Proposition 4.13)

$$\texttt{Rad}(\mathcal{A} \circ \{x_1,\ldots,x_n\}) \le \sqrt{\frac{2\,\texttt{VC}(\mathcal{A})\log(en/\texttt{VC}(\mathcal{A}))}{n}}$$

Question: Can we use same idea in regression, isolating elements that matter?

Yes! We need covering/packing numbers, metric arguments (no combinatorics)

NB: This will also help in classification, allowing to remove $\log(en/\texttt{VC}(\mathcal{A}))$

Covering and Packing Numbers

A pseudometric space $(\mathcal{S}, \rho)$ is a set $\mathcal{S}$ and a function $\rho : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_+$ (called a *pseudometric*) such that, for any $x, y, z \in \mathcal{S}$ we have:

- $\rho(x, y) = \rho(y, x)$ (symmetry)
- $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (triangle inequality)
- $\rho(x, x) = 0$

A metric space is obtained if one further assumes that $\rho(x, y) = 0$ implies $x = y$

Covering and Packing Numbers (Definition 4.14)

Let $(\mathcal{S}, \rho)$ be a pseudometric space, $\varepsilon > 0$

- The set $\mathcal{C} \subseteq \mathcal{S}$ is a *ε-cover* of $(\mathcal{S}, \rho)$ if for every $x \in \mathcal{S}$ there exists $y \in \mathcal{C}$ such that $\rho(x, y) \leq \varepsilon$.The set $\mathcal{C} \subseteq \mathcal{S}$ is a *minimal ε-cover* if there is no other ε-cover with lower cardinality. The cardinality of any minimal ε-cover is the *ε-covering number*, denoted by $\mathtt{Cov}(\mathcal{S}, \rho, \varepsilon)$
- The set $\mathcal{P} \subseteq \mathcal{S}$ is a *ε-packing* of $(\mathcal{S}, \rho)$ if for every $x, x' \in \mathcal{P}$ we have $\rho(x, x') > \varepsilon$.The set $\mathcal{P} \subseteq \mathcal{S}$ is a *maximal ε-packing* if there is no other ε-packing with greater cardinality.The cardinality of any maximal ε-packing is the *ε-packing number*, denoted by $\mathtt{Pack}(\mathcal{S}, \rho, \varepsilon)$

Covering and Packing Numbers. Properties

Duality (Proposition 4.15)

$$\boxed{\texttt{Cov}(\mathcal{S}, \rho, \varepsilon) \leq \texttt{Pack}(\mathcal{S}, \rho, \varepsilon) \leq \texttt{Cov}(\mathcal{S}, \rho, \varepsilon/2)}$$

Covering and packing numbers typically grow **exponentially** with the dimension

Bounded Balls (Proposition 4.16)

$\mathcal{B}_r^d := \{y \in \mathbb{R}^d : \|y\| \leq r\}$ be the d-dim. ball with radius $r \geq 0$. If $\varepsilon \leq r$, then

$$\boxed{\left(\frac{r}{\varepsilon}\right)^d \leq \texttt{Cov}(\mathcal{B}_r^d, \|\cdot\|, \varepsilon) \leq \texttt{Pack}(\mathcal{B}_r^d, \|\cdot\|, \varepsilon) \leq \left(\frac{3r}{\varepsilon}\right)^d}$$

Proof: Volume argument

Covering and packing numbers grow exponentially also w.r.t. the **VC dimension**. This, along with chaining, will allow us to remove the $\log$-term in Prop. 4.13

Back to Regression...

- $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathcal{X} = \mathbb{R}^d \to \mathbb{R}\}$
- Given data $x = \{x_1, \dots, x_n\} \in \mathcal{X}^n$, define data-dependent pseudonorms on $\mathcal{A}$:

$$\|a\|_{p,x} := \left(\frac{1}{n}\sum_{i=1}^{n} |a(x_i)|^p\right)^{1/p} \qquad \|a\|_{\infty,x} := \max_i |a(x_i)|$$

- The pseudonorms induce the following pseudometrics:

$$\|a-b\|_{p,x} := \left(\frac{1}{n}\sum_{i=1}^{n} |a(x_i)-b(x_i)|^p\right)^{1/p} \qquad \|a-b\|_{\infty,x} := \max_i |a(x_i)-b(x_i)|$$

(Proposition 5.1)

For any $x = \{x_1, \dots, x_n\} \in \mathcal{X}^n$, $1 \leq p \leq q$, and $\varepsilon > 0$, we have

$$\boxed{\mathtt{Cov}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \leq \mathtt{Cov}(\mathcal{A}, \|\cdot\|_{q,x}, \varepsilon)}$$

$$\boxed{\mathtt{Pack}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \leq \mathtt{Pack}(\mathcal{A}, \|\cdot\|_{q,x}, \varepsilon)}$$

Thus, in what follows it is enough to prove results for small values of p

Bound on Rademacher Complexity via Covering Numbers

(Proposition 5.2)

For any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$, let $\sup_{a \in \mathcal{A}} \|a\|_{2,x} \le c_x$. Then,

$$\texttt{Rad}(\mathcal{A} \circ x) \le \inf_{\varepsilon > 0} \left\{ \varepsilon + \frac{\sqrt{2}\, c_x}{\sqrt{n}} \sqrt{\log \texttt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)} \right\}$$

Proof:
Fix $x \in \mathcal{X}^n$, $\varepsilon > 0$. Let $\mathcal{C} \subseteq \mathcal{A}$ be a minimal ε-cover of $(\mathcal{A}, \|\cdot\|_{1,x})$
For any $a \in \mathcal{A}$ let $\tilde{a} \in \mathcal{C}$ be such that $\|a - \tilde{a}\|_{1,x} \le \varepsilon$

$$\texttt{Rad}(\mathcal{A} \circ x) \le \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i (a(x_i) - \tilde{a}(x_i)) + \mathbf{E} \sup_{a \in \mathcal{A}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i \tilde{a}(x_i)$$

$$\le \varepsilon + \mathbf{E} \max_{a \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \Omega_i a(x_i)$$

$$\le \varepsilon + \max_{a \in \mathcal{C}} \sqrt{\sum_{i=1}^{n} a(x_i)^2}\, \frac{\sqrt{2 \log |\mathcal{C}|}}{n}$$ by Massart's lemma

$$\le \varepsilon + c_x \sqrt{\frac{2 \log \texttt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)}{n}}$$ as $|\mathcal{C}| = \texttt{Cov}(\mathcal{A}, \|\cdot\|_{1,x}, \varepsilon)$

Improved Result using Chaining

(Proposition 5.3)

For any $x = \{x_1, \ldots, x_n\} \in \mathcal{X}^n$ and $\sup_{a\in\mathcal{A}} \|a\|_{2,x} \leq c_x$ we have

$$\texttt{Rad}(\mathcal{A} \circ x) \leq \inf_{\varepsilon\in[0,c_x/2]} \left\{ 4\varepsilon + \frac{12}{\sqrt{n}} \int_{\varepsilon}^{c_x/2} \mathrm{d}\nu \sqrt{\log \texttt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \nu)} \right\}$$

Proof (main ideas):

Fix $x \in \mathcal{X}^n$. Define family of covers: let $\varepsilon_j = \frac{c_x}{2^j}$ and $\mathcal{C}_j \subseteq \mathcal{A}$ be a minimal ε_j-cover of $(\mathcal{A}, \|\cdot\|_{2,x})$

For any $a \in \mathcal{A}, j \geq 1$ let $a_j \in \mathcal{C}_j$ s.t. $\|a - a_j\|_{2,x} \leq \varepsilon_j$. Use $a = a - a_m + \sum_{j=1}^m (a_j - a_{j-1})$ (chain)

$$\texttt{Rad}(\mathcal{A} \circ x) \leq \mathbf{E} \sup_{a\in\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i (a(x_i) - a_m(x_i)) + \mathbf{E} \sup_{a\in\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i \sum_{j=1}^m (a_j(x_i) - a_{j-1}(x_i))$$

First term:

$$\sum_{i=1}^n \Omega_i (a(x_i) - a_m(x_i)) \leq \sum_{i=1}^n |a(x_i) - a_m(x_i)| = n\|a - a_m\|_{1,x} \leq n\|a - a_m\|_{2,x} \leq n\varepsilon_m$$

Second term: $\mathbf{E} \sup_{a\in\mathcal{A}} \frac{1}{n} \sum_{i=1}^n \Omega_i (a_j(x_i) - a_{j-1}(x_i)) \leq \sup_{a\in\mathcal{A}} \|a_j - a_{j-1}\|_{2,x} \frac{\sqrt{2\log|\mathcal{C}_j||\mathcal{C}_{j-1}|}}{\sqrt{n}}$

We get $\texttt{Rad}(\mathcal{A} \circ x) \leq \varepsilon_m + \frac{12}{\sqrt{n}} \sum_{j=1}^m (\varepsilon_j - \varepsilon_{j+1}) \sqrt{\log \texttt{Cov}(\mathcal{A}, \|\cdot\|_{2,x}, \varepsilon_j)} \leq$ integral

Back to Classification...

(Proposition 5.5)

$$\texttt{Pack}(\mathcal{A}, \|\cdot\|_{p,x}, \varepsilon) \le \left(\frac{10}{\varepsilon^p}\log\frac{2e}{\varepsilon^p}\right)^{\texttt{VC}(\mathcal{A})}$$

$\Longrightarrow$

(Theorem 5.6)

$$\texttt{Rad}(\mathcal{A}\circ x) \lesssim \sqrt{\frac{\texttt{VC}(\mathcal{A})}{n}}$$

Proof of Proposition 5.5 (main ideas):

W.l.o.g. $p = 1$. Fix $x \in \mathcal{X}^n$ and $\varepsilon > 0$. Let $\mathcal{P} \subseteq \mathcal{A}$ be a maximal ε-packing. For any $a, b \in \mathcal{P}$

$$\varepsilon < \|a - b\|_{1,x} = \frac{1}{n}\sum_{i=1}^{n}|a(x_i) - b(x_i)| = \frac{1}{n}\sum_{i=1}^{n} 1_{a(x_i)\neq b(x_i)} = \mathbf{P}(a(Z) \neq b(Z))$$

Let $Z_1, \ldots, Z_m$ be m i.i.d. random variables distributed as Z (uniform in $\{x_1, \ldots, x_n\}$):

$$\begin{aligned}
&\mathbf{P}(|\mathcal{P}\circ\{Z_1,\ldots,Z_m\}| = |\mathcal{P}|) \\
&\quad = \mathbf{P}(\text{For every } a,b\in\mathcal{P}, a\neq b, \text{ we have } a\circ\{Z_1,\ldots,Z_m\} \neq b\circ\{Z_1,\ldots,Z_m\}) \\
&\quad = 1 - \mathbf{P}(\text{There exists } a,b\in\mathcal{P}, a\neq b, \text{ such that } a\circ\{Z_1,\ldots,Z_m\} = b\circ\{Z_1,\ldots,Z_m\}) \\
&\quad > 1 - |\mathcal{P}|^2(1-\varepsilon)^m > 1 - |\mathcal{P}|^2 e^{-m\varepsilon}
\end{aligned}$$

by union bound, independence, and packing property

Bound > 0 for $m = \frac{2}{\varepsilon}\log|\mathcal{P}| \Rightarrow$ there exists $z_1, \ldots, z_m$ (probabilistic method)

$$|\mathcal{P}| = |\mathcal{P}\circ z| \le |\mathcal{A}\circ z| \le \tau_{\mathcal{A}}(m) = \tau_{\mathcal{A}}\left(\frac{2}{\varepsilon}\log|\mathcal{P}|\right)$$

Proof follows by using Sauer-Shelah's lemma and computing an upper bound for the recursion