Algorithmic Foundations of Learning

Lecture 4
VC Dimension. Covering and Packing Numbers

Patrick Rebeschini

Department of Statistics
University of Oxford

Recap: Regression

- $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$. $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \mathbb{R}\}$. $\ell(a, (x, y)) = \phi(a(x), y)$
- Goal: $$\boxed{\texttt{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\}) \leq \frac{f(\text{dimension}, \text{complexity of } \mathcal{A})}{n^\alpha}}$$

SVM (Proposition 3.2)

Let $\mathcal{A}_2 := \{x \in \mathbb{R}^d \to w^\top x : \|w\|_2 \leq c\}$. Then

$$\boxed{\texttt{Rad}(\mathcal{A}_2 \circ \{x_1, \ldots, x_n\}) \leq \max_i \|x_i\|_\infty c \frac{\sqrt{d}}{\sqrt{n}}}$$

Boosting (Proposition 3.6)

Let $\mathcal{A}_\Delta := \{x \in \mathbb{R}^d \to w^\top x : \|w\|_1 = c, w_1, \ldots, w_d \geq 0\}$. Then

$$\boxed{\texttt{Rad}(\mathcal{A}_\Delta \circ \{x_1, \ldots, x_n\} \leq \max_i \|x_i\|_\infty c \frac{\sqrt{2 \log d}}{\sqrt{n}}}$$

Difference between d and $\log d$ related to difference between ℓ_2 and ℓ_1 ball, resp.

Today: Classification (binary)

- $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$
- Admissible action set $\mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\}$
- Loss function $\ell(a, (x, y)) = \phi(a(x), y)$, for $\phi : \{-1, 1\}^2 \to \mathbb{R}_+$
- Today we consider $\phi(\hat{y}, y) = 1_{\hat{y} \neq y} = (1 - y\hat{y})/2$, a.k.a. the **true loss**

Recall. For regression we used:

(Proposition 3.1)

If the function $\hat{y} \to \phi(\hat{y}, y)$ is γ-Lipschitz for any $y \in \mathcal{Y}$, then

$$\boxed{\mathtt{Rad}(\mathcal{L} \circ \{z_1, \ldots, z_n\}) \leq \gamma \, \mathtt{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\})}$$

For classification with the true loss we can use:

(Proposition 4.1)

If ϕ is the true loss, then $\boxed{\mathtt{Rad}(\mathcal{L} \circ \{z_1, \ldots, z_n\}) = \frac{1}{2} \mathtt{Rad}(\mathcal{A} \circ \{x_1, \ldots, x_n\})}$

Growth Function

- $\mathcal{A} \circ \{x_1, \ldots, x_n\} = \{(a(x_1), \ldots, a(x_n)) \in \{-1, 1\}^n : a \in \mathcal{A}\}$
- $|\mathcal{A} \circ \{x_1, \ldots, x_n\}| \leq 2^n$ **even if the class $\mathcal{A}$ is infinite**
- **Important:** It can growth **polynomially** with n

Growth function (Definition 4.2)

The *growth function* of $\mathcal{A}$ is defined as

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

Max number of labelings of n vectors that we can obtain using classifiers in $\mathcal{A}$

Yields "data-independent" bound on Rademacher complexity (Massart's lemma)

(Proposition 4.3)

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

Note: To drive convergence to 0 as n grows, we need $\tau_{\mathcal{A}}$ to grow **polynomially**

Growth Function: Examples

- **Half spaces over the real line** $\mathcal{A} = \{a(x) = 2 1_{x \le w} - 1 : w \in \mathbb{R}\}$

$$\begin{array}{l} 0000\cdots0 \\ 1000\cdots0 \\ 1100\cdots0 \\ \vdots \\ 1111\cdots1 \end{array}$$

$$\boxed{\tau_{\mathcal{A}}(n) = n + 1}$$

- **Intervals over the real line** $\mathcal{A} = \{a(x) = 2 1_{w^- \le x \le w^+} - 1 : w^- \le w^+\}$

$$\begin{array}{llllll} 0000\cdots00 & & & & & \\ 1000\cdots00 & 0100\cdots00 & 0010\cdots00 & \cdots & 00000\cdots10 & 00000\cdots01 \\ 1100\cdots00 & 0110\cdots00 & 0011\cdots00 & \cdots & 0000\cdots11 & \\ \vdots & & & & & \\ 1111\cdots11 & & & & & \end{array}$$

$$\boxed{\tau_{\mathcal{A}}(n) = 1 + n(n+1)/2}$$

Problem: not always easy to compute! **Solution:** VC dimension

VC Dimension

VC dimension (Definition 4.6)

$$\boxed{\mathtt{VC}(\mathcal{A}) := \max\{n \in \mathbb{N} : \tau_{\mathcal{A}}(n) = 2^n\}}$$

If $\tau_{\mathcal{A}}(n) = 2^n$ for all integer n, then $\mathtt{VC}(\mathcal{A}) = \infty$

- **Half spaces over the real line** $\mathcal{A} = \{a(x) = 21_{x \leq w} - 1 : w \in \mathbb{R}\}$
  $\boxed{\tau_{\mathcal{A}}(n) = n + 1}$ $\tau_{\mathcal{A}}(1) = 2^1$ and $\tau_{\mathcal{A}}(2) = 3 < 2^2 \implies \mathtt{VC}(\mathcal{A}) = 1$

- **Intervals over the real line** $\mathcal{A} = \{a(x) = 21_{w^- \leq x \leq w^+} - 1 : w^- \leq w^+\}$
  $\boxed{\tau_{\mathcal{A}}(n) = 1 + n(n+1)/2}$ $\tau_{\mathcal{A}}(2) = 2^2$ and $\tau_{\mathcal{A}}(3) = 7 < 2^3 \implies \mathtt{VC}(\mathcal{A}) = 2$

Key point: We can compute the VC dimension without computing $\tau_{\mathcal{A}}$

- **Sufficient** to find k such that $\tau_{\mathcal{A}}(k) = 2^k$ and $\tau_{\mathcal{A}}(k+1) < 2^{k+1}$
- This can be done without computing $\tau_{\mathcal{A}}$. **Sufficient** to:
  - Find distinct $x_1, \dots, x_k$ that are "shattered" by $\mathcal{A} \Rightarrow \mathtt{VC}(\mathcal{A}) \geq k$
    (i.e., classifiers in $\mathcal{A}$ can assign all possible 2^k labelings to these points)
  - Show that no set of $k + 1$ points can be "shattered" by $\mathcal{A} \Rightarrow \mathtt{VC}(\mathcal{A}) < k + 1$
    (i.e., for any set of $k + 1$ points there is a label that can **not** be assigned)

Bounds using VC Dimension

If $\mathtt{VC}(\mathcal{A})$ is **finite**, then $\tau_{\mathcal{A}}$ eventually grows **polynomially**

Sauer-Shelah's Lemma (Lemma 4.11)

$$\tau_{\mathcal{A}}(n) \begin{cases} = 2^n & \text{if } n \leq \mathtt{VC}(\mathcal{A}) \\ \leq \left(\frac{en}{\mathtt{VC}(\mathcal{A})}\right)^{\mathtt{VC}(\mathcal{A})} & \text{if } n > \mathtt{VC}(\mathcal{A}) \end{cases}$$

(Proposition 4.13)

For any $x_1, \ldots, x_n \in \mathbb{R}^d$ we have

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

- This bound is "data-independent" as it holds for any $x_1, \ldots, x_n$ (as such, it does not allow to exploit the *statistical* nature of the data)
- We will remove the $\log$-term using covering numbers and chaining

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 4.13)

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 $\texttt{Cov}(\mathcal{S}, \rho, \varepsilon)$
- 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 $\texttt{Pack}(\mathcal{S}, \rho, \varepsilon)$

Covering and Packing Numbers. Properties

Duality (Proposition 4.14)

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

Covering and packing numbers *typically* grow **exponentially** with the dimension (in so-called "Logarithmic metric entropy" spaces)

Bounded Balls (Proposition 4.15)

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

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

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