CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 25**

Covering Numbers and Pseudo-dimension

Lecturer: Peter Bartlett *Scribe: Pascal Michaillat*

1 Recapitulation

Theorem 1.1. (*Dudley*)

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(x_i) \leq c \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon$$

For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and $\dim(\mathcal{F}) \leq d$:

$$\hat{R}_n(\mathcal{F}) \leq c\sqrt{\frac{d}{n}} \ .$$

Example. (*VC-classes*) Let $\mathcal{F} \in \{-1,+1\}^{\mathcal{X}}$, $d_{VC}(\mathcal{F}) = d$:

$$\begin{aligned} \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) &\leq |\mathcal{F}_{|(x_1,\dots,x_n)}| \\ &\leq \Pi_{\mathcal{F}}(n) \\ &\leq \left(\frac{en}{d}\right)^d \end{aligned}$$

Therefore, we obtain:

$$\sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n))}{n}} \leq \sqrt{\frac{d\log(\frac{en}{d})}{n}} \ .$$

This can be improved to:

$$\begin{aligned} \log(\mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n))) &\leq 2d\log(\frac{1}{\epsilon}) + cd \\ \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) &\leq (\frac{c}{\epsilon})^{2d} \end{aligned}$$

which yields:

$$\hat{R}_n(\mathcal{F}) \leq c\sqrt{\frac{d}{n}} \ .$$

2 Packing Number

As we will see, packing numbers are tightly related to covering numbers. Particularly, bounds on packing numbers imply bounds on covering numbers and vice versa.

Definition. (*Packing number*) For a metric space $(\mathcal{S}, \rho)$ and $T \subseteq \mathcal{S}$, we say that $T' \subseteq T$ is an ϵ-packing if

$$\forall\ (t_1, t_2) \in T' \times T',\ \ \rho(t_1, t_2) \geq \epsilon\ .$$

The ϵ-packing number of T is defined as

$$\mathcal{M}(\epsilon, T, \rho) = \max\left\{|T'| : T' \subseteq T \text{ is an } \epsilon\text{-packing}\right\}$$

Lemma 2.1. $\mathcal{M}(2\epsilon, T, \rho) \leq \mathcal{N}(\epsilon, T, \rho) \leq \mathcal{M}(\epsilon, T, \rho)$ where $\mathcal{M}$ denotes the packing number and $\mathcal{N}$ denotes the covering number.

Proof.
$[\mathcal{M}(2\epsilon) \leq \mathcal{N}(\epsilon)]$
Let $C \subseteq T$ be an ϵ-cover and $P \subseteq T$ a 2ϵ-packing. We have

$$\forall p \in P,\ \exists c \in C,\ \ \rho(c, p) < \epsilon\ .$$

Now suppose $|P| > |C|$. Then some $c \in C$ has $\rho(c, p_1) < \epsilon$ and $\rho(c, p_2) < \epsilon$ for distinct $p_1, p_2 \in P$. Using the triangular inequality on the metric ρ, we obtain $\rho(p_1, p_2) < 2\epsilon$, which in turn implies that P is not a 2ϵ-packing. We reach a contradiction and $|P| \leq |C|$.

$[\mathcal{N}(\epsilon) \leq \mathcal{M}(\epsilon)]$
Suppose $P \subseteq T$ is an ϵ-packing, but not an ϵ-cover. Then $\exists t \in T,\ \forall p \in P, \rho(t, p) \geq \epsilon$. Hence $P \cup \{t\}$ is an ϵ-packing. To conclude, any maximal ϵ-packing is an ϵ-cover. □

Recall. For $V \subseteq \{0,1\}^n$, we defined $d_{VC}(V) = \max\left\{|I| :\ I \subseteq \{1, \ldots, n\}, |V_{|I}| = 2^{|I|}\right\}$

Theorem 2.2. If $V \subseteq \{0,1\}^n$ has $d_{VC} = d$ and $d_1(a, b) = \frac{1}{n}\sum_{i=1}^{n} |a_i - b_i|$, then:

$$\mathcal{M}(\epsilon, V, d_1) \leq 2(\frac{4e}{\epsilon})^d\ .$$

We also have:

$$\mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) \leq (\frac{c}{\epsilon})^{2d}\ .$$

3 Pseudo-dimension

The pseudo-dimension is a generalization of the VC-dimension to real valued functions. It is not limited to functions with values in $\{0,1\}$.

Definition. (*Pseudo-dimension*) The pseudo-dimension of $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is:

$$d_P(\mathcal{F}) = d_{VC}(\{(x, y) \mapsto \text{sign}(f(x) - y) : f \in \mathcal{F}\})$$

.

Note. $\{(x, y) \mapsto \text{sign}(f(x) - y),\ f \in \mathcal{F}\}$ is the set of indicators of the epigraphs[1] of the functions $f \in \mathcal{F}$.

Example. If $\mathcal{F}$ is the linear space of functions of dimension d, then $d_P(\mathcal{F}) = d$. Basically, combinatorial dimensions such as the VC-dimension or the pseudo-dimension are generalizations of the usual linear dimension.

[1]Epigraph of f is the set $\{(x, y) : y \geq f(x)\}$.

Example. Let $\mathcal{F}$ be the set of non-decreasing functions: $\mathbb{R} \to [0,1]$. Then $d_P(\mathcal{F}) = +\infty$, although $\mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) \leq n^{\frac{1}{\epsilon}}$.

The idea is, for any d, to pick an increasing sequence $\{y_1, \ldots, y_d\}$ and set $x_i = i$. Then $\{(x_i, y_i)\}$ is shattered by $\{(x, y) \mapsto \text{sign}(f(x) - y),\ f \in \mathcal{F}\}$.

Note. d_P does not characterize the rate of growth of the covering number as the VC-dimension characterizes the rate of growth of $\Pi_n(\mathcal{F})$. What is true for the VC-dimension is that:

$$\Pi_n(\mathcal{F}) \text{ is polynomial in n } \Leftrightarrow d_{VC}(\mathcal{F}) < +\infty$$

whereas for the pseudo-dimension we only have:

$$\sup_{P_n} \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) \text{ polynomial in n } \not\Rightarrow d_p(\mathcal{F}) < +\infty$$

$$\sup_{P_n} \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n)) \text{ polynomial in n } \Leftarrow d_p(\mathcal{F}) < +\infty$$

Theorem 3.1. For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and for any probability distribution $\mathcal{P}$:

$$\mathcal{M}(\epsilon,\ \mathcal{F},\ ,L_2(\mathcal{P})) \leq 2(\frac{4\ e}{\epsilon^2})^{d_P(\mathcal{F})}\ .$$

Proof. Let $f,\ g : \mathcal{X} \to [0,\ 1]$. Let $X \sim \mathcal{P}$ and $U \sim \mathcal{U}$ and $\mathcal{U} \equiv \text{Uniform}([0,\ 1])$.

$$\begin{aligned}
||f - g||^2 &= \mathbb{E}\left[(f(X) - g(x))^2\right] \\
&= \mathbb{E}(\mathbb{P}\{f(X) \geq U|X\} - \mathbb{P}\{g(X) \geq U|X\})^2 \\
&= \mathbb{E}(\mathbb{E}[\mathbb{1}[f(X) \geq U] - \mathbb{1}[g(X) \geq U]|X])^2 \\
&\leq \mathbb{E}(\mathbb{1}[f(X) \geq U] - \mathbb{1}[g(X) \geq U])^2 \\
&= \mathbb{E}\,|\mathbb{1}[f(X) \geq U] - \mathbb{1}[g(X) \geq U]|
\end{aligned}$$

Now let $\mathcal{G} = \{(x, y) \mapsto \mathbb{1}[f(x) - y \geq 0] : f \in \mathcal{F}\}$ We obtain

$$\mathcal{M}(\epsilon, \mathcal{F}, L_2(\mathcal{P}) \leq \mathcal{M}(\epsilon^2,\ \mathcal{G},\ L_1(\mathcal{P} \times \mathcal{U})))\ ,$$

and finally

$$\begin{aligned}
\mathcal{M}(\epsilon^2, \mathcal{G}, L_1(\mathcal{P} \times \mathcal{U}) &\leq 2\left(\frac{4e}{\epsilon^2}\right)^{d_{\mathrm{VC}}(\mathcal{G})} \\
&= 2\left(\frac{4e}{\epsilon^2}\right)^{d_P(\mathcal{F})} \quad .
\end{aligned}$$

□