CS281B/Stat241B. Statistical Learning Theory. Lecture 9.

Peter Bartlett

- Covering numbers
  - Approximating real-valued functions
  - Chaining and Dudley's entropy integral
  - Sudakov's lower bound

ERM and uniform laws of large numbers

Empirical risk minimization:

Choose $f_n \in F$ to minimize $\hat{R}$.

$$
\begin{aligned}
R(f_n) &\leq \inf_{f \in F} R(f) + \sup_{f \in F} \left| R(f) - \hat{R}(f) \right| + O(1/\sqrt{n}) \\
&= \inf_{f \in F} R(f) + O\left(\mathbb{E} \| R_n \|_F \right).
\end{aligned}
$$

Covering and packing numbers

Definition: A **pseudometric space** (S, d) is a set S and a function $d : S \times S \to [0, \infty)$ satisfying

1. $d(x, x) = 0$,
2. $d(x, y) = d(y, x)$,
3. $d(x, z) \leq d(x, y) + d(y, z)$.

Covering numbers

Definition: An ϵ-cover of a subset T of a pseudometric space (S, d) is a set $\hat{T} \subset T$ such that for each $t \in T$ there is a $\hat{t} \in \hat{T}$ such that $d(t, \hat{t}) \leq \epsilon$. The ϵ-covering number of T is

$$\mathcal{N}(\epsilon, T, d) = \min\{|\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-cover of } T\}.$$

A set T is **totally bounded** if, for all $\epsilon > 0$, $\mathcal{N}(\epsilon, T, d) < \infty$.
The function $\epsilon \mapsto \log \mathcal{N}(\epsilon, T, d)$ is the **metric entropy** of T.
If $\lim_{\epsilon \to 0} \log \mathcal{N}(\epsilon) / \log(1/\epsilon)$ exists, it is called the **metric dimension**.

- Entropy: number of bits to approximately specify an element of T.
- Example: $([0,1]^d, l_\infty)$ has $\mathcal{N}(\epsilon) = \Theta(1/\epsilon^d)$.
  Intuition: A d-dimensional set has metric dimension d.

Covering numbers

Theorem: For $F \subseteq [-1,1]^{\mathcal{X}}$ and $x_1, \ldots, x_n \in \mathcal{X}$, consider the $L_2(P_n)$ pseudometric on F,

$$d_n(f,g)^2 = P_n(f-g)^2.$$

Then

$$\mathbb{E}\|R_n\|_F \leq \inf_{\alpha>0} \left(\mathbb{E}\sqrt{\frac{2\log(2\mathcal{N}(\alpha, F, d_n))}{n}} + \alpha \right).$$

Covering numbers

Proof:

For a sample $X_1, \ldots, X_n$, fix a minimal α-cover $\hat{F}$ of F.

$$\begin{aligned}
\mathbb{E}\|R_n\|_F &= \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \\
&= \mathbb{E} \sup_{\hat{f} \in \hat{F}} \sup_{f \in F \cap B_\alpha(\hat{f})} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \hat{f}(X_i) + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(X_i) - \hat{f}(X_i)) \right| \\
&\leq \mathbb{E} \sqrt{\frac{2 \log(2 \mathcal{N}(\alpha, F, d_n))}{n}} + \alpha.
\end{aligned}$$

Covering numbers

Example: If $\mathcal{N}(\alpha, F, d_n) = \alpha^{-d}$, we can choose $\alpha = 1/\sqrt{n}$ to get

$$\mathbb{E}\|R_n\|_F = O\left(\sqrt{\frac{d \log n}{n}}\right).$$

Packing numbers

Definition: An ϵ-packing of a subset T of a pseudometric space (S, d) is a subset $\hat{T} \subset T$ such that each pair $s, t \in \hat{T}$ satisfies $d(s, t) > \epsilon$. The ϵ-packing number of T is

$$\mathcal{M}(\epsilon, T, d) = \max\{|\hat{T}| : \hat{T} \text{ is an } \epsilon\text{-packing of } T\}.$$

Covering and packing numbers

Theorem: For all $\epsilon > 0$, $\mathcal{M}(2\epsilon) \leq \mathcal{N}(\epsilon) \leq \mathcal{M}(\epsilon)$.

Thus, the scaling of the covering and packing numbers is the same.

Covering and packing numbers: Proof

For the first inequality, consider a minimal ϵ-cover $\hat{T}$. Any two elements of a 2ϵ-packing of T cannot be within ϵ of the same element of $\hat{T}$. (Otherwise, the triangle inequality shows that they are within 2ϵ of each other.) Thus, there can be no more than one element of a 2ϵ packing for each of the $\mathcal{N}(\epsilon)$ elements of $\hat{T}$. That is, $\mathcal{M}(2\epsilon) \leq \mathcal{N}(\epsilon)$.

For the second inequality, consider an ϵ-packing $\hat{T}$ of size $\mathcal{M}(\epsilon)$. Since it is maximal, no other point $s \in T$ can be added for which some $t \in \hat{T}$ has $d(s,t) > \epsilon$. Thus, $\hat{T}$ is an ϵ-cover. So the minimal ϵ-cover has size $\mathcal{N}(\epsilon) \leq \mathcal{M}(\epsilon)$.

Example: smoothly parameterized functions

Let F be a parameterized class of functions,

$$F = \{f(\theta, \cdot) : \theta \in \Theta\}.$$

Let $\|\cdot\|_\Theta$ be a norm on Θ and let $\|\cdot\|_F$ be a norm on F. Suppose that the mapping $\theta \mapsto f(\theta, \cdot)$ is L-Lipschitz, that is,

$$\|f(\theta, \cdot) - f(\theta', \cdot)\|_F \leq L\|\theta - \theta'\|_\Theta.$$

Then $\mathcal{N}(\epsilon, F, \|\cdot\|_F) \leq \mathcal{N}(\epsilon/L, \Theta, \|\cdot\|_\Theta)$.

Example: smoothly parameterized functions

A Lipschitz parameterization allows us to translates a cover of the parameter space into a cover of the function space.

Example: If F is smoothly parameterized by a (compact set of) d parameters, then $\mathcal{N}(\epsilon, F) = O(1/\epsilon^d)$.

Example: non-decreasing functions

Example: For the class F of non-decreasing functions from $\mathbb{R}$ to $[0,1]$, and the random pseudometric d_n on F,

$$d_n(f,g)^2 = P_n(f-g)^2.$$

we have

$$\mathcal{N}(\epsilon, F, d_n) = n^{O(1/\epsilon)}.$$

For this class, the metric dimension is infinite.

Example: non-decreasing functions

To see this, notice that we need only approximate restrictions of functions in this class to $X_1, \ldots, X_n$. We can replace the range $[0, 1]$ by a discretization $\hat{\mathcal{Y}} := \{0, \epsilon, 2\epsilon, \ldots, \lfloor 1/\epsilon \rfloor \epsilon, 1\}$. Then for any $f \in F$ there is a $\hat{f} : \{X_1, \ldots, X_n\} \to \hat{\mathcal{Y}}$ that has $d_n(f, \hat{f}) \leq \epsilon$. So we just need to count the number of non-decreasing $\hat{f}$'s.

We can specify a non-decreasing function $\hat{f}$ by specifying, for each value in $\hat{\mathcal{Y}}$, the smallest X_i at which it lies on or above that value.

Example: non-decreasing functions

$$\mathcal{N}(\epsilon, F, d_n) = n^{O(1/\epsilon)}.$$

Two consequences of this covering number bound:

- We can write the class of functions of total variation no more than 1 as $G = \{(f-g)/2 : f, g \in F\}$, so it has $\mathcal{N}(\epsilon, G, d_n) = n^{O(1/\epsilon)}$.
- The discretization theorem implies

$$\mathbb{E}\|R_n\|_F \leq \inf_{\alpha>0}\left(c\sqrt{\frac{\log n}{\alpha n}} + \alpha\right) = O\left(\left(\frac{\log n}{n}\right)^{1/3}\right).$$

(But we know that this result is loose. Why?)

Overview

- Covering numbers
  - Approximating real-valued functions
  - Chaining and Dudley's entropy integral
  - Sudakov's lower bound

Chaining and Dudley's entropy integral

Theorem: For some universal constant c, if $F \subseteq [0,1]^{\mathcal{X}}$,

$$\mathbb{E}\|R_n\|_F \le c\mathbb{E} \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\alpha, F, d_n)}{n}}\, d\alpha.$$

Proof of Dudley's entropy integral

Rather than choosing a fixed value of α, we approximate an element of F at progressively finer scales:

$$
\begin{aligned}
f &= \hat{f}_N + f - \hat{f}_N = \hat{f}_0 + \sum_{i=1}^{N} (\hat{f}_i - \hat{f}_{i-1}) + f - \hat{f}_N, \\
\hat{f}_i &\in \hat{F}_i, \qquad d_n(\hat{f}_i, \hat{f}_{i-1}) \leq \alpha_i, \\
\alpha_i &= 2^{-i} \mathrm{diam}(F), \\
\hat{F}_i &= \ \alpha_i\text{-cover of } F.
\end{aligned}
$$

We can set $\hat{f}_0 = 0$ and notice that $\mathrm{diam}(F) \leq 1$.

Proof of Dudley's entropy integral

$$
\begin{aligned}
\mathbb{E}\|R_n\|_F &= \mathbb{E}\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right| \\
&= \mathbb{E}\sup_{f\in F}\left|\left\langle \epsilon, \sum_{j=1}^{N}(\hat{f}_j - \hat{f}_{j-1}) + f - \hat{f}_N \right\rangle\right| \\
&\le \mathbb{E}\sum_{j=1}^{N}\sup_{\hat{f}_j\in\hat{F}_j,\hat{f}_{j-1}\in\hat{F}_{j-1}}\left|\left\langle \epsilon, \hat{f}_j - \hat{f}_{j-1}\right\rangle\right| + \mathbb{E}\sup_{f\in F}\left|\left\langle \epsilon, f - \hat{f}_N\right\rangle\right| \\
&\le \mathbb{E}\sum_{j=1}^{N}\alpha_j\sqrt{\frac{2\log(2|\hat{F}_j|\,|\hat{F}_{j-1}|)}{n}} + \alpha_N
\end{aligned}
$$

Proof of Dudley's entropy integral

Now, $|F_{j-1}| \leq |F_j| = \mathcal{N}(\alpha_j, F, d_n)$ and $\alpha_j = 2\alpha_{j+1} = 2(\alpha_j - \alpha_{j+1})$:

$$\begin{aligned}
\mathbb{E}\|R_n\|_F &\leq c\mathbb{E}\left(\sum_{j=1}^{N}(\alpha_j - \alpha_{j+1})\sqrt{\frac{\log \mathcal{N}(\alpha_j, F, d_n)}{n}}\right) + \alpha_N \\
&\leq c\mathbb{E}\int_{\alpha_{N+1}}^{\alpha_0}\sqrt{\frac{\log \mathcal{N}(\alpha, F, d_n)}{n}}\,d\alpha + \alpha_N,
\end{aligned}$$

where at the last step we've lower bounded the integral by the piecewise constant function.

Chaining and Dudley's entropy integral

Theorem: For some universal constant c, if $F \subseteq [0,1]^{\mathcal{X}}$,

$$\mathbb{E}\|R_n\|_F \leq c\mathbb{E} \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\alpha, F, d_n)}{n}}\, d\alpha.$$

Applications of chaining and Dudley's entropy integral

Example: For a subset F of a d-dimensional linear space, $\log \mathcal{N}(\epsilon, F, d_n) \sim d \log(1/\epsilon)$.
A single discretization gives

$$\mathbb{E}\|R_n\|_F \leq c\sqrt{\frac{d \log n}{n}}.$$

Chaining gives

$$\mathbb{E}\|R_n\|_F \leq c\sqrt{\frac{d}{n}} \int_0^1 \sqrt{\log 1/\epsilon}\, d\epsilon = c'\sqrt{\frac{d}{n}}.$$

(To calculate the integral, notice that $y = \sqrt{\log(1/x)}$ means $x = e^{-y^2}$.)

Applications of chaining and Dudley's entropy integral

Example: For the class F of non-decreasing functions from $\mathbb{R}$ to $[0,1]$, we calculated

$$\mathbb{E}\|R_n\|_F \leq \inf_{\alpha>0}\left(c\sqrt{\frac{\log n}{\alpha n}}+\alpha\right) = O\left(\left(\frac{\log n}{n}\right)^{1/3}\right).$$

But chaining gives

$$\mathbb{E}\|R_n\|_F \leq c\int_0^1 \sqrt{\frac{\log n}{\epsilon n}}\,d\epsilon = c'\left(\frac{\log n}{n}\right)^{1/2}.$$

Applications of chaining and Dudley's entropy integral

Example: For $F \subseteq \{\pm 1\}^{\mathcal{X}}$ with $d_{VC}(F) \leq d$, we have seen that Sauer's Lemma plus the finite class lemma implies

$$\mathbb{E}\|R_n\|_F \leq c' \sqrt{\frac{d \log n}{n}}.$$

However, Haussler showed that

$$\mathcal{N}(\alpha, F, d_n) \leq \left(\frac{c}{\alpha}\right)^{2d}.$$

So Dudley's entropy integral evaluates to

$$\mathbb{E}\|R_n\|_F \leq c' \sqrt{\frac{d}{n}}.$$

Overview

- Covering numbers
  - Approximating real-valued functions
  - Chaining and Dudley's entropy integral
  - Sudakov's lower bound

Sudakov's Theorem

Theorem:

$$\mathbb{E}\|R_n\|_F \geq \frac{c}{\log n} \sup_{\alpha} \left(\alpha \mathbb{E} \sqrt{\frac{\log(\mathcal{N}(\alpha, F, d_n))}{n}} \right).$$

Ignoring the $\log n$, this lower bound is the largest rectangle that we can fit under the graph of $\sqrt{\log(\mathcal{N}(\alpha, F, d_n))/n}$.

Covering numbers

- There is a gap between the upper and lower bounds on $\mathbb{E}\|R_n\|_F$ in terms of covering numbers. This gap is essential.
- We have seen that $\mathbb{E}\|R_n\|_F$ gives tight bounds on $\|P - P_n\|_F$. Covering numbers do not.
- Covering numbers are convenient: it is often easy to bound them by piecing together approximations.

Overview

- Covering numbers
  - Approximating real-valued functions
  - Chaining and Dudley's entropy integral
  - Sudakov's lower bound