

LECTURE 9: COVERING NUMBER AND REAL-VALUED FUNCTION CLASS

Recall that if $S = \{x_1, \dots, x_n\} \subset \mathcal{X}^n$ and \mathcal{H} is a set of binary-valued functions, then \mathcal{H}_S denotes the restriction of \mathcal{H} to S , and the growth function (shatter coefficient) is defined by

$$G_{\mathcal{H}}(n) = \sup_S |\mathcal{H}_S|.$$

Since \mathcal{H} maps \mathcal{X} into $\{0, 1\}$, \mathcal{H}_S is finite for any finite S ($|\mathcal{H}_S| \leq 2^n$).

Such a definition works well for binary-valued functions, but if \mathcal{H} is a set of real-valued functions then \mathcal{H}_S will be a set with infinite cardinality even for finite n . Essentially we do not need to divide \mathcal{H} into unique elements based on \mathcal{H}_S , but only need to partition the function class \mathcal{H} into small groups which are “local” in nature.

Definition. Let (\mathcal{W}, d) be a metric space and $\mathcal{F} \subset \mathcal{W}$. For every $\epsilon > 0$, denote by $N(\epsilon, \mathcal{F}, d)$ the minimal number of open balls (with respect to metric d) needed to cover \mathcal{F} . That is, $N(\epsilon, \mathcal{F}, d)$ is the minimal cardinality of the set $\{f_1, \dots, f_m\} \subset \mathcal{W}$ with the property that for every $f \in \mathcal{F}$ there is some f_i such that $d(f, f_i) < \epsilon$. The set $\{f_1, \dots, f_m\}$ is called an ϵ -cover of \mathcal{F} . The logarithm of the covering number is called the *entropy* of the set.

We will be interested in metrics induced by samples. For every sample $\{x_1, \dots, x_n\}$ let μ_n be the empirical measure of the sample. For $1 \leq p \leq \infty$ and a function f , put

$$\|f\|_{L_p(\mu_n)} = \left(\frac{1}{n} \sum_{i=1}^n |f(x_i)|^p \right)^{1/p}$$

and in particular, we have $\|f\|_{L_\infty(\mu_n)} = \max_{1 \leq i \leq n} |f(x_i)|$. Let $N(\epsilon, \mathcal{F}, L_p(\mu_n))$ be the covering number of \mathcal{F} at scale ϵ with respect to the norm $L_p(\mu_n)$.

Theorem 9-1. For any class \mathcal{F} of real-valued functions, any sample $S = \{x_1, \dots, x_n\}$ and $\epsilon > 0$,

$$N(\epsilon, \mathcal{F}, L_1(\mu_n)) \leq N(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq N(\epsilon, \mathcal{F}, L_\infty(\mu_n)).$$

Definition. For $\epsilon > 0$, define the *uniform covering number*

$$N_p(\epsilon, \mathcal{F}, n) = \sup_{\mu_n} N(\epsilon, \mathcal{F}, L_p(\mu_n)).$$

It is easy to see that the uniform covering number is a generalization of the growth function. Suppose that \mathcal{F} contains functions which map \mathcal{X} into $\{0, 1\}$. Then for any $S = \{x_1, \dots, x_n\}$, $\epsilon < 1$ and $p = \infty$, we have $N(\epsilon, \mathcal{F}, L_\infty(\mu_n)) = |\mathcal{H}_S|$, so $N_\infty(\epsilon, \mathcal{F}, n) = |G_{\mathcal{H}}|$.

Based on the covering number we are able to obtain uniform convergence result for real-valued function class.

Theorem 9-2. Let \mathcal{F} be a class of functions which map \mathcal{X} into $[-1, 1]$ and let μ be a probability measure on \mathcal{X} . Assume X_1, \dots, X_n are independent random variables distributed according to μ . For every $\epsilon > 0$ and $n \geq 8/\epsilon^2$,

$$P \left(\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum f(X_i) - \mathbb{E}[f(X)] \right| > \epsilon \right) \leq 8 \mathbb{E}[N(\epsilon, \mathcal{F}, L_1(\mu_n))] \exp \left(-\frac{n\epsilon^2}{128} \right),$$

where μ_n is the empirical measure on X_1, \dots, X_n .

PROOF. Consider the event $A = \{\sup_{f \in \mathcal{F}} |\sum_{i=1}^n \sigma_i f(X_i)| > n\epsilon/4\}$. We have

$$\begin{aligned} P(A) &= \mathbb{E}_\mu[\mathbb{E}_\sigma[I(A)|X_1, \dots, X_n]] \\ &= \mathbb{E}_\mu \left[\mathbb{E}_\sigma \left[I \left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| > \frac{n\epsilon}{4} \right) |X_1, \dots, X_n \right] \right]. \end{aligned}$$

For any realization of X_1, \dots, X_n , its empirical measure is μ_n . Let \mathcal{G} be an $\epsilon/8$ cover of \mathcal{F} with respect to the $L_1(\mu_n)$ norm, and we can assume that any function $g \in \mathcal{G}$ is bounded by 1. First, observe that

$$P \left(\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i f(X_i) \right| > \frac{n\epsilon}{4} \right) \leq P \left(\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \sigma_i g(X_i) \right| > \frac{n\epsilon}{8} \right).$$

This is because if there is some element $f^* \in \mathcal{F}$ which makes the LHS event true, then we can find some $g^* \in \mathcal{G}$ such that

$$\frac{1}{n} \sum_{i=1}^n |\sigma_i f^*(X_i) - \sigma_i g^*(X_i)| = \frac{1}{n} \sum_{i=1}^n |f^*(X_i) - g^*(X_i)| \leq n\epsilon.$$

So we have $\sup_{g \in \mathcal{G}} |\sum_{i=1}^n \sigma_i g(X_i)| > n\epsilon/8$. Applying the union bound, Hoeffding's inequality and utilizing the fact that $\forall g \in \mathcal{G}, \sum_{i=1}^n g(x_i)^2 \leq n$, we have

$$\begin{aligned} P \left(\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n \sigma_i g(X_i) \right| > \frac{n\epsilon}{8} \right) &\leq |\mathcal{G}| \cdot \sup_{g \in \mathcal{G}} P \left(\left| \sum_{i=1}^n \sigma_i g(X_i) \right| > \frac{n\epsilon}{8} \right) \\ &\leq 2N(\epsilon/8, \mathcal{F}, L_1(\mu_n)) \exp \left(-\frac{n\epsilon^2}{128} \right). \end{aligned}$$

The claim follows by combing this result with symmetrization (with ghost sample and Rademacher random variables).

□

Lemma. For $A \subset \mathbb{R}^n$ with $r = \max_{a \in A} \|a\|$, and $\sigma_1, \dots, \sigma_n$ being Rademacher random variables, we have

$$\mathbb{E} \sup_{a \in A} \left(\sum_{i=1}^n \sigma_i a_i \right) \leq r \sqrt{2 \log |A|}.$$

PROOF.

For any $s > 0$ we have

$$\begin{aligned} \exp \left(s \mathbb{E} \sup_{a \in A} \left(\sum_{i=1}^n \sigma_i a_i \right) \right) &\leq \mathbb{E} \exp \left(s \sup_{a \in A} \left(\sum_{i=1}^n \sigma_i a_i \right) \right) \\ &= \mathbb{E} \sup_{a \in A} \left(\exp \left(s \sum_{i=1}^n \sigma_i a_i \right) \right) \\ &\leq \sum_{a \in A} \mathbb{E} \exp \left(s \sum_{i=1}^n \sigma_i a_i \right) \\ &\leq \sum_{a \in A} \exp \left(\frac{s^2}{2} \sum_{i=1}^n a_i^2 \right) \\ &\leq |A| \exp (s^2 r^2 / 2). \end{aligned}$$

So we have

$$\mathbb{E} \sup_{a \in A} \left(\sum_{i=1}^n \sigma_i a_i \right) \leq \inf_{s > 0} \left(\frac{\log |A|}{s} + \frac{s r^2}{2} \right) = r \sqrt{2 \log |A|}.$$

□

By the lemma we have obvious that

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log |\mathcal{F}|}{n}}$$

if \mathcal{F} is finite with output values within $[-1, 1]$.

Theorem 9-3. For $\mathcal{F} \subset [-1, 1]^{\mathcal{X}}$, we have

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \inf_{\epsilon > 0} \left(\sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, L_2(\mu_n))}{n}} + \epsilon \right).$$

PROOF.

For an $\epsilon > 0$, let \mathcal{G} be an ϵ -cover of \mathcal{F} . Then we have

$$\begin{aligned} \hat{\mathcal{R}}_n(\mathcal{F}) &= \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(x_i) \right] \\ &= \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \sup_{f \in \mathcal{F} \cap B_\epsilon(g)} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) + \frac{1}{n} \sum_{i=1}^n \sigma_i (f(x_i) - g(x_i)) \right) \right] \\ &\leq \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \sigma_i g(x_i) + \epsilon \right] \\ &\leq \sqrt{\frac{2 \log N(\epsilon, \mathcal{F}, L_2(\mu_n))}{n}} + \epsilon, \end{aligned}$$

where the first equality utilizes the fact that $\mathcal{F} = \bigcup_{g \in \mathcal{G}} (\mathcal{F} \cap B_\epsilon(g))$, and the first inequality comes from the fact that $\|f - g\|_{L_2(\mu_n)} \leq \epsilon$ and C-S inequality.

□

Definition. Let (\mathcal{W}, d) be a metric space and $\mathcal{F} \subset \mathcal{W}$. For $\epsilon > 0$, a subset A is said to be an ϵ -packing of \mathcal{F} , if for all distinct $f_1, f_2 \in A$, we have $d(f_1, f_2) > \epsilon$. The ϵ -packing number $P(\epsilon, \mathcal{F}, d)$ is defined as the maximum cardinality of an ϵ -packing subset.

Both the covering number and packing number can be used to measure the size of the sets, and they are obviously related. The following simple result shows that as long as one of them can be computed, we can easily obtain a bound for the other one.

Theorem 9-4. Given a metric space (\mathcal{W}, d) . Then for all $\epsilon > 0$ and for every $\mathcal{F} \subset \mathcal{W}$, the covering number and packing number satisfy

$$P(2\epsilon, \mathcal{F}, d) \leq N(\epsilon, \mathcal{F}, d) \leq P(\epsilon, \mathcal{F}, d).$$