Covering Number Estimates

Mathematical Analysis of Machine Learning Algorithms
(Chapter 5)

Packing Number

Definition 1 (Packing Number)

Let $(\mathcal{V}, d)$ be a pseudometric space with metric $d(\cdot, \cdot)$. A finite subset $\mathcal{G}(\epsilon) \subset \mathcal{G}$ is an ϵ-packing of $\mathcal{G}$ if $d(\phi, \phi') > \epsilon$ for all $\phi, \phi' \in \mathcal{G}(\epsilon)$. The ϵ-packing number of $\mathcal{G}$, denoted by $M(\epsilon, \mathcal{G}, d)$, is the largest cardinality of ϵ-packing of $\mathcal{G}$.

Covering Number versus Packing Number

Theorem 2 (Thm 5.2)

For all $\epsilon > 0$, we have

$$N(\epsilon, \mathcal{G}, d) \leq M(\epsilon, \mathcal{G}, d) \leq N(\epsilon/2, \mathcal{G}, d).$$

It is often convenient to use packing number because an ϵ-packing always belong to $\mathcal{G}$. This means that any assumption of $\mathcal{G}$ holds for an ϵ-packing of $\mathcal{G}$.

Proof of Theorem 2

Let $\mathcal{G}(\epsilon) = \{\phi_1, \ldots, \phi_M\} \subset \mathcal{G}$ be a maximal ϵ-packing of $\mathcal{G}$. Given any $\phi \in \mathcal{G}$, by the definition of maximality, we know that there exists $\phi_j \in \mathcal{G}(\epsilon)$ so that $d(\phi_j, \phi) \leq \epsilon$. This means that $\mathcal{G}(\epsilon)$ is also an ϵ cover of $\mathcal{G}$. Therefore $N(\epsilon, \mathcal{G}, d) \leq M$. This proves the first inequality.
On the other hand, let $\mathcal{G}'(\epsilon/2)$ be an $\epsilon/2$ cover of $\mathcal{G}$. By definition, for any $\phi_j \in \mathcal{G}(\epsilon)$, there exists $\tilde{g}(\phi_j) \in \mathcal{G}'(\epsilon/2)$ such that $d(\tilde{g}(\phi_j), \phi_j) \leq \epsilon/2$. For $j \neq i$, we know that $d(\phi_i, \phi_j) > \epsilon$, and thus triangle inequality implies that

$$d(\tilde{g}(\phi_j), \phi_i) \geq d(\phi_i, \phi_j) - d(\tilde{g}(\phi_j), \phi_j) > \epsilon/2 \geq d(\tilde{g}(\phi_i), \phi_i).$$

Therefore $\tilde{g}(\phi_i) \neq \tilde{g}(\phi_j)$. This implies the map $\phi_j \in \mathcal{G}(\epsilon) \to \tilde{g}(\phi_j) \in \mathcal{G}'(\epsilon/2)$ is one to one. Therefore $|\mathcal{G}(\epsilon)| \leq |\mathcal{G}'(\epsilon/2)|$. This proves the second inequality.

Finite Dimensional Space

Theorem 3 (Thm 5.3)

Let $\|\cdot\|$ *be a seminorm on* $\mathbb{R}^k$. *Let* $B(r) = \{z \in \mathbb{R}^k : \|z\| \leq r\}$ *be the* $\|\cdot\|$*-ball with radius* r. *Then*

$$M(\epsilon, B(r), \|\cdot\|) \leq (1 + 2r/\epsilon)^k.$$

Moreover,

$$N(\epsilon, B(r), \|\cdot\|) \geq (r/\epsilon)^k.$$

Proof of Theorem 3 (I/II)

Let $\{z_1, \ldots, z_M\} \subset B(r)$ be a maximal ϵ packing of $B(r)$. Let $B_j = \{z \in \mathbb{R}^k : \|z - z_j\| \leq \epsilon/2\}$, then $B_j \cap B_k = \emptyset$ for $j \neq k$ and $B_j \subset B(r + \epsilon/2)$ for all j. It follows that

$$\sum_{j=1}^{M} \mathrm{volume}(B_j) = \mathrm{volume}(\cup_{j=1}^{M} B_j) \leq \mathrm{volume}(B(r + \epsilon/2)).$$

Let $v = \mathrm{volume}(B(1))$. Since $\mathrm{volume}(B_j) = (\epsilon/2)^k v$ and $\mathrm{volume}(B(r + \epsilon/2)) = (r + \epsilon/2)^k v$, we have

$$M(\epsilon/2)^k v \leq (r + \epsilon/2)^k v.$$

This implies the first bound.

Proof of Theorem 3 (II/II)

Let $\{z_1, \ldots, z_N\} \subset \mathbb{R}^k$ be a cover of $B(r)$. If we define $B_j = \{z \in \mathbb{R}^k : \|z - z_j\| \leq \epsilon\}$, then $B(r) \subset \cup_j B_j$. Therefore

$$\mathrm{volume}(B(r)) \leq \mathrm{volume}(\cup_{j=1}^N B_j) \leq \sum_{j=1}^N \mathrm{volume}(B_j).$$

Let $v = \mathrm{volume}(B(1))$. Since $\mathrm{volume}(B_j) = (\epsilon)^k v$ and $\mathrm{volume}(B(r)) = r^k v$, we have

$$r^k v \leq N \epsilon^k v.$$

This implies the second bound.

Lipschitz Function Class

Theorem 4 (Thm 5.4)

Consider

$$\{\phi(w, Z) : w \in \Omega\}, \tag{1}$$

where $\Omega \subset \mathbb{R}^k$ *is a compact set.*
Assume that $\Omega \subset \mathbb{R}^k$ *is a compact set so that* $\Omega \in B(r)$ *with respect to a norm* $\|\cdot\|$. *Assume for all* z, $\phi(w, z)$ *is* $\gamma(z)$ *Lipschitz with respect to* w*:*

$$|\phi(w, z) - \phi(w', z)| \leq \gamma(z)\|w - w'\|.$$

Given $p \geq 1$, *let* $\gamma_p = (\mathbb{E}_{Z\sim D}|\gamma(Z)|^p)^{1/p}$. *Then*

$$N_{[]}(2\epsilon, \mathcal{G}, L_p(\mathcal{D})) \leq (1 + 2\gamma_p r/\epsilon)^k.$$

Proof of Theorem 4

Let $\{w_1, \dots, w_M\}$ be an ϵ/γ_p packing of Ω. Then it is also an ϵ/γ_p cover of Ω.
Let

$$\phi_j^L(z) = \phi(w_j, z) - \gamma(z)\epsilon/\gamma_p$$

and

$$\phi_j^U(z) = \phi(w_j, z) + \gamma(z)\epsilon/\gamma_p.$$

Then $\{[\phi_j^L, \phi_j^U] : j = 1, \dots, M\}$ is an 2ϵ $L_p(\mathcal{D})$-bracketing cover.

We can now apply Theorem 3 to obtain the desired result.

Empirical L_1 Covering of VC-class

Theorem 5 (Thm 5.5)

If a binary valued function class $\mathcal{G} = \{\phi(w, Z) : w \in \Omega\}$ *is a VC class, then for* $\epsilon \le 1$*:*

$$\ln M(\epsilon, \mathcal{G}, L_1(\mathcal{S}_n)) \le 3d + d \ln(\ln(4/\epsilon)/\epsilon).$$

Proof of Theorem 5 (I/II)

Given $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$. Let $Q = \{\phi_1, \ldots, \phi_m\}$ be a maximal ϵ $L_1(\mathcal{S}_n)$ packing of $\mathcal{G}$. Q is also an L_1 ϵ-cover of $\mathcal{G}$. Consider the empirical distribution, denoted by $\mathcal{S}_n$, which puts a probability of $1/n$ on each Z_i. We have for $j \neq k$:

$$\Pr_{Z \sim \mathcal{S}_n}[\phi_j(Z) = \phi_k(Z)] = 1 - \mathbb{E}_{Z \sim \mathcal{S}_n}|\phi_j(Z) - \phi_k(Z)| < 1 - \epsilon.$$

Now consider random sample with replacement from $\mathcal{S}_n$ for T times to obtain samples $\{Z_{i_1}, \ldots, Z_{i_T}\}$. We have

$$\Pr(\{\forall \ell : \phi_j(Z_{i_\ell}) = \phi_k(Z_{i_\ell})\}) < (1 - \epsilon)^T \leq e^{-T\epsilon}.$$

That is, with probability larger than $1 - e^{-T\epsilon}$,

$$\exists \ell : \phi_j(Z_{i_\ell}) \neq \phi_k(Z_{i_\ell}).$$

Taking the union bound for all $j \neq k$, we have with probability larger than $1 - \binom{m}{2} \cdot e^{-T\epsilon}$, for all $j \neq k$:

$$\exists \ell : \phi_j(Z_{i_\ell}) \neq \phi_k(Z_{i_\ell}).$$

Proof of Theorem 5 (II/II)

If we take $T = \lceil \ln(m^2)/\epsilon \rceil$, then $e^{-T\epsilon}\binom{m}{2} \leq 1$. Then there exists T samples $\{Z_{i_\ell} : \ell = 1, \ldots, T\}$ such that $\phi_j \neq \phi_k$ for all $j \neq k$ when restricted to these samples. Since $\mathrm{VC}(G) = d$, we obtain from Sauer's lemma:

$$m \leq \max[2, eT/d]^d \leq \max[2, e(1 + \ln(m^2)/\epsilon)/d]^d.$$

The theorem holds automatically when $m \leq 2^d$. Otherwise,

$$\ln m \leq d \ln(1/\epsilon) + d \ln((e\epsilon/d) + (2e/d)\ln(m)).$$

Let $u = d^{-1} \ln m - \ln(1/\epsilon) - \ln\ln(4/\epsilon)$ and let $\epsilon \leq 1$, we can obtain the following bound by using the upper bound of $\ln m$:

$$\begin{aligned} u &\leq -\ln\ln(4/\epsilon) + \ln((e\epsilon/d) + 2e(u + \ln(1/\epsilon) + \ln\ln(4/\epsilon))) \\ &\leq \ln \frac{2e(u + 0.5 + \ln(1/\epsilon) + \ln\ln(4/\epsilon))}{\ln(4/\epsilon)} \leq \ln(4u + 7), \end{aligned}$$

where the last inequality is obtained by taking $\sup$ over $\epsilon \in (0, 1]$. By solving this inequality we obtain a bound $u \leq 3$. This implies the desired result.

A More Refined Result

Theorem 6 ([Haussler, 1995], Thm 5.6)

Let $\mathcal{G}$ be a binary valued function class with $\textsc{vc}(G) = d$*. Then*

$$\ln M(\epsilon, \mathcal{G}, L_1(\mathcal{S}_n)) \leq 1 + \ln(d+1) + d\ln(2e/\epsilon).$$

D. Haussler (1995). "Sphere packing numbers for subsets of the Boolean *n*-cube with bounded Vapnik-Chervonenkis dimension". In: *Journal of Combinatorial Theory, Series A* 69.2, pp. 217–232 .

Corollary 7 (Cor 5.7)

If $\textsc{vc}(\mathcal{G}) = d$*, then for all distributions* $\mathcal{D}$ *over* Z*, we have*

$$\ln N(\epsilon, \mathcal{G}, L_p(\mathcal{D})) \leq 1 + \ln(d+1) + d\ln(2e/\epsilon^p)$$

for $\epsilon \in (0, 1]$ *and* $p \in [1, \infty)$*.*

VC-Subgraph Class

One may extend the concept of VC dimension to real valued functions by introducing the definition of VC subgraph class.

Definition 8

A real valued function class of $z \in \mathcal{Z}$

$$\mathcal{G} = \{\phi(w, Z) : w \in \Omega\}$$

is a VC-subgraph class, if the binary function class

$$\mathcal{G}_{\text{subgraph}} = \{\mathbb{1}(t < \phi(w, z)) : w \in \Omega\}$$

defined on $(z, t) \in \mathcal{Z} \times \mathbb{R}$ is a VC class. The VC dimension (some times also called pseudo-dimension) of $\mathcal{G}$ is

$$\mathrm{vc}(\mathcal{G}) = \mathrm{vc}(\mathcal{G}_{\text{sub}-\text{graph}}).$$

Example: Linear Functions

Example 9

The d dimensional linear functions of the form

$$f_w(x) = w^\top x$$

is VC subgraph class of VC dimension $d + 1$.
This is because $w^\top x - t$ is linear function in $d + 1$ dimension, and we have shown that it has VC dimension $d + 1$.

Example: Composition with Monotone Function

Example 10

If $\mathcal{F} = \{f(w, x) : w \in \Omega\}$ is a VC subgraph class and h is monotone function, then

$$h \circ \mathcal{F} = \{h(f(w, x)) : w \in \Omega\}$$

is a VC subgraph class with

$$\mathrm{vc}(h \circ \mathcal{F}) \leq \mathrm{vc}(\mathcal{F}).$$

In particular, if $f(w, x) = w^\top x$ is a d-dimensional linear function, then $h(f(w, x))$ has VC dimension $d + 1$.

Covering Number of VC-Subgraph Class

Theorem 11 (Thm 5.11)

Assume that $\mathcal{G}$ is a VC subgraph class, with VC dimension d, and all $\phi \in \mathcal{G}$ are bounded: $\phi(Z) \in [0, 1]$. Then for any distribution $\mathcal{D}$ over Z, $\epsilon \in (0, 1]$ and $p \in [1, \infty)$, we have

$$\ln N(\epsilon, \mathcal{G}, L_p(\mathcal{D})) \leq 1 + \ln(d + 1) + d \ln(2e/\epsilon^p).$$

Moreover,

$$\ln N_\infty(\epsilon, \mathcal{G}, n) \leq d \ln \max[2, en/(d\epsilon)].$$

Proof of Theorem 11 (I/II)

Let U be a random variable distributed uniformly over $[0, 1]$. Then for all $a \in (0, 1)$: $\mathbb{E}_U \mathbb{1}(U \le a) = a$. Thus for all $\phi, \phi' \in \mathcal{G}$:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}} |\phi(Z) - \phi'(Z)|^p \\
=&\mathbb{E}_{\mathcal{D}} |\mathbb{E}_U [\mathbb{1}(U \le \phi(Z)) - \mathbb{1}(U \le \phi'(Z))]|^p \\
\le&\mathbb{E}_{\mathcal{D}} \mathbb{E}_U |\mathbb{1}(U \le \phi(Z)) - \mathbb{1}(U \le \phi'(Z))|^p.
\end{aligned}$$

The last inequality used the Jensen's inequality. Therefore

$$\ln N(\epsilon, \mathcal{G}, L_p(\mathcal{D})) \le \ln N(\epsilon, G_{\text{subgraph}}, L_p(\mathcal{D} \times U(0, 1))).$$

This leads to the first desired bound.

Proof of Theorem 11 (II/II)

The second bound can be proved by discretizing U into intervals with thresholds $\min(1, \epsilon(2k+1))$ for $k = 0, 1, \ldots$ with no more than $\lceil (2\epsilon)^{-1} \rceil \leq 1/\epsilon$ thresholds. This gives an ϵ-cover of U in Euclidean distance.

We can then approximate $\mathbb{E}_U$ by average over the thresholds to get ϵ L_∞ cover with the discretization. Let the set of thresholds be U'.

If $\mathcal{D}$ contain n data points, then $\mathcal{D} \times U'$ contains at most $n|U'| \leq n/\epsilon$ points, and one may apply Sauer's lemma to obtain a cover on these points. This implies the second bound.

Regularized Linear Function Class

$$\mathcal{F} = \{f(w, x) = w^\top \psi(x) : w \in \Omega, x \in \mathcal{X}\} \tag{2}$$

where $\psi(x)$ is a known feature vector.

Theorem 12 (Thm 5.18)

Let $w = [w_1, w_2, \ldots] \in \mathbb{R}^\infty$ *and* $\psi(x) = [\psi_1(x), \psi_2(x), \ldots] \in \mathbb{R}^\infty$*. Let* $\Omega = \{w : \|w\|_2 \leq A\}$*. Given a distribution* $\mathcal{D}$ *on* $\mathcal{X}$*. Assume there exists* $B_1 \geq B_2 \geq \cdots$ *such that*

$$\mathbb{E}_{x \sim \mathcal{D}} \sum_{i \geq j} \psi_i(x)^2 \leq B_j^2.$$

Define $\tilde{d}(\epsilon) = \min\{j \geq 0 : AB_{j+1} \leq \epsilon\}$*. Then the function class* $\mathcal{F}$ *of* (2) *satisfies:*

$$\ln N(\epsilon, \mathcal{F}, L_2(\mathcal{D})) \leq \tilde{d}(\epsilon/2) \ln \left(1 + \frac{4AB_1}{\epsilon}\right).$$

Proof of Theorem 12

Given $\epsilon > 0$. Consider $j = \tilde{d}(\epsilon/2)$ such that $AB_{j+1} \leq \epsilon/2$. Let

$$\mathcal{F}_1 = \left\{ \sum_{i=1}^{j} w_i \psi_i(x) : w \in \Omega \right\}$$

$$\mathcal{F}_2 = \left\{ \sum_{i>j} w_i \psi_i(x) : w \in \Omega \right\}.$$

Since $\|f\|_{L_2(\mathcal{D})} \leq \epsilon/2$ for all $f \in \mathcal{F}_2$, we have $N(\epsilon/2, \mathcal{F}, L_2(\mathcal{D})) = 1$. Moreover, Theorem 3 implies that

$$\ln N(\epsilon/2, \mathcal{F}_1, L_2(\mathcal{D})) \leq \tilde{d}(\epsilon/2) \ln \left(1 + \frac{4AB_0}{\epsilon} \right).$$

Note that $\mathcal{F} \subset \mathcal{F}_1 + \mathcal{F}_2$, we have

$$\ln N(\epsilon, \mathcal{F}, L_2(\mathcal{D})) \leq \ln N(\epsilon/2, \mathcal{F}_1, L_2(\mathcal{D})) + \ln N(\epsilon/2, \mathcal{F}_2, L_2(\mathcal{D})).$$

This implies the result.

Example

Example 13

Assume that $B_j = j^{-q}$, then

$$\ln N(\epsilon, \mathcal{F}, L_2(\mathcal{D})) = O\left(\epsilon^{-q} \ln(1/\epsilon)\right).$$

If $B_j = O(c^j)$ for some $c \in (0, 1)$, then

$$\ln N(\epsilon, \mathcal{F}, L_2(\mathcal{D})) = O\left((\ln(1/\epsilon))^2\right).$$

L_2 Regularized Empirical L_∞ Covering Number

Theorem 14 (Thm 5.20)

Assume that $\Omega = \{w : \|w\|_2 \leq A\}$ *and* $\|\psi(x)\|_2 \leq B$*, then the function class* (2) *has the following covering number bound:*

$$\ln N(\mathcal{F}, \epsilon, L_\infty(\mathcal{S}_n)) \leq \frac{36A^2B^2}{\epsilon^2} \ln[2\lceil(4AB/\epsilon) + 2\rceil n + 1].$$

Empirical L_∞ covering number bounds can be used in margin analysis and can be used to derive better bounds in multiclass classification.

Summary (Chapter 5)

- Packing number and relationship with covering number
- Finite dimensional function classes
- L_p covering for VC class.
- VC-subgraph class.
- Regularized linear function class.