27 Covering Numbers

In this chapter we describe another way to measure the complexity of sets, which is called covering numbers.

27.1 Covering

DEFINITION 27.1 (Covering) Let $A \subset \mathbb{R}^m$ be a set of vectors. We say that A is r-covered by a set A', with respect to the Euclidean metric, if for all $\mathbf{a} \in A$ there exists $\mathbf{a}' \in A'$ with $\|\mathbf{a} - \mathbf{a}'\| \leq r$. We define by $N(r, A)$ the cardinality of the smallest A' that r-covers A.

Example 27.1 (Subspace) Suppose that $A \subset \mathbb{R}^m$, let $c = \max_{\mathbf{a} \in A} \|\mathbf{a}\|$, and assume that A lies in a d-dimensional subspace of $\mathbb{R}^m$. Then, $N(r, A) \leq (2c\sqrt{d}/r)^d$. To see this, let $\mathbf{v}_1, \ldots, \mathbf{v}_d$ be an orthonormal basis of the subspace. Then, any $\mathbf{a} \in A$ can be written as $\mathbf{a} = \sum_{i=1}^d \alpha_i \mathbf{v}_i$ with $\|\boldsymbol{\alpha}\|_\infty \leq \|\boldsymbol{\alpha}\|_2 = \|\mathbf{a}\|_2 \leq c$. Let $\epsilon \in \mathbb{R}$ and consider the set

$$A' = \left\{ \sum_{i=1}^d \alpha_i' \mathbf{v}_i : \forall i, \alpha_i' \in \{-c, -c+\epsilon, -c+2\epsilon, \ldots, c\} \right\}.$$

Given $\mathbf{a} \in A$ s.t. $\mathbf{a} = \sum_{i=1}^d \alpha_i \mathbf{v}_i$ with $\|\boldsymbol{\alpha}\|_\infty \leq c$, there exists $\mathbf{a}' \in A'$ such that

$$\|\mathbf{a} - \mathbf{a}'\|^2 = \|\sum_i (\alpha_i' - \alpha_i)\mathbf{v}_i\|^2 \leq \epsilon^2 \sum_i \|\mathbf{v}_i\|^2 \leq \epsilon^2 \, d.$$

Choose $\epsilon = r/\sqrt{d}$; then $\|\mathbf{a} - \mathbf{a}'\| \leq r$ and therefore A' is an r-cover of A. Hence,

$$N(r, A) \leq |A'| = \left(\frac{2c}{\epsilon}\right)^d = \left(\frac{2c\sqrt{d}}{r}\right)^d.$$

27.1.1 Properties

The following lemma is immediate from the definition.

LEMMA 27.2 *For any $A \subset \mathbb{R}^m$, scalar $c > 0$, and vector $\mathbf{a}_0 \in \mathbb{R}^m$, we have*

$$\forall r > 0, \quad N(r, \{c\,\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in A\}) \leq N(cr, A).$$

Next, we derive a contraction principle.

LEMMA 27.3 *For each $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a ρ-Lipschitz function; namely, for all $\alpha, \beta \in \mathbb{R}$ we have $|\phi_i(\alpha) - \phi_i(\beta)| \le \rho\,|\alpha - \beta|$. For $\mathbf{a} \in \mathbb{R}^m$ let $\boldsymbol{\phi}(\mathbf{a})$ denote the vector $(\phi_1(a_1), \ldots, \phi_m(a_m))$. Let $\boldsymbol{\phi} \circ A = \{\boldsymbol{\phi}(\mathbf{a}) : a \in A\}$. Then,*

$$N(\rho\, r, \boldsymbol{\phi} \circ A) \le N(r, A).$$

Proof Define $B = \boldsymbol{\phi} \circ A$. Let A' be an r-cover of A and define $B' = \boldsymbol{\phi} \circ A'$. Then, for all $\mathbf{a} \in A$ there exists $\mathbf{a}' \in A'$ with $\|\mathbf{a} - \mathbf{a}'\| \le r$. So,

$$\|\boldsymbol{\phi}(\mathbf{a}) - \boldsymbol{\phi}(\mathbf{a}')\|^2 = \sum_i (\phi_i(a_i) - \phi_i(a_i'))^2 \le \rho^2 \sum_i (a_i - a_i')^2 \le (\rho r)^2.$$

Hence, B' is an $(\rho\, r)$-cover of B. □

27.2 From Covering to Rademacher Complexity via Chaining

The following lemma bounds the Rademacher complexity of A based on the covering numbers $N(r, A)$. This technique is called *Chaining* and is attributed to Dudley.

LEMMA 27.4 *Let $c = \min_{\bar{\mathbf{a}}} \max_{\mathbf{a} \in A} \|\mathbf{a} - \bar{\mathbf{a}}\|$. Then, for any integer $M > 0$,*

$$R(A) \le \frac{c\, 2^{-M}}{\sqrt{m}} + \frac{6\, c}{m} \sum_{k=1}^{M} 2^{-k} \sqrt{\log(N(c\, 2^{-k}, A))}.$$

Proof Let $\bar{\mathbf{a}}$ be a minimizer of the objective function given in the definition of c. On the basis of Lemma 26.6, we can analyze the Rademacher complexity assuming that $\bar{\mathbf{a}} = \mathbf{0}$.

Consider the set $B_0 = \{\mathbf{0}\}$ and note that it is a c-cover of A. Let $B_1, \ldots, B_M$ be sets such that each B_k corresponds to a minimal $(c\, 2^{-k})$-cover of A. Let $\mathbf{a}^* = \operatorname{argmax}_{\mathbf{a} \in A} \langle \boldsymbol{\sigma}, \mathbf{a} \rangle$ (where if there is more than one maximizer, choose one in an arbitrary way, and if a maximizer does not exist, choose $\mathbf{a}^*$ such that $\langle \boldsymbol{\sigma}, \mathbf{a}^* \rangle$ is close enough to the supremum). Note that $\mathbf{a}^*$ is a function of $\boldsymbol{\sigma}$. For each k, let $\mathbf{b}^{(k)}$ be the nearest neighbor of $\mathbf{a}^*$ in B_k (hence $\mathbf{b}^{(k)}$ is also a function of $\boldsymbol{\sigma}$). Using the triangle inequality,

$$\|\mathbf{b}^{(k)} - \mathbf{b}^{(k-1)}\| \le \|\mathbf{b}^{(k)} - \mathbf{a}^*\| + \|\mathbf{a}^* - \mathbf{b}^{(k-1)}\| \le c\,(2^{-k} + 2^{-(k-1)}) = 3\, c\, 2^{-k}.$$

For each k define the set

$$\hat{B}_k = \{(\mathbf{a} - \mathbf{a}') : \mathbf{a} \in B_k, \mathbf{a}' \in B_{k-1}, \|\mathbf{a} - \mathbf{a}'\| \le 3\, c\, 2^{-k}\}.$$

We can now write

$$\begin{aligned} R(A) &= \frac{1}{m}\,\mathbb{E}\langle \boldsymbol{\sigma}, \mathbf{a}^* \rangle \\ &= \frac{1}{m}\,\mathbb{E}\left[\langle \boldsymbol{\sigma}, \mathbf{a}^* - \mathbf{b}^{(M)} \rangle + \sum_{k=1}^{M} \langle \boldsymbol{\sigma}, \mathbf{b}^{(k)} - \mathbf{b}^{(k-1)} \rangle\right] \\ &\leq \frac{1}{m}\,\mathbb{E}\left[\|\sigma\|\,\|\mathbf{a}^* - \mathbf{b}^{(M)}\|\right] + \sum_{k=1}^{M} \frac{1}{m}\,\mathbb{E}\left[\sup_{\mathbf{a}\in \hat{B}_k} \langle \boldsymbol{\sigma}, \mathbf{a} \rangle\right]. \end{aligned}$$

Since $\|\boldsymbol{\sigma}\| = \sqrt{m}$ and $\|\mathbf{a}^* - \mathbf{b}^{(M)}\| \leq c\,2^{-M}$, the first summand is at most $\frac{c}{\sqrt{m}}\,2^{-M}$. Additionally, by Massart lemma,

$$\frac{1}{m}\,\mathbb{E}\sup_{\mathbf{a}\in \hat{B}_k} \langle \boldsymbol{\sigma}, \mathbf{a} \rangle \leq 3\,c\,2^{-k}\,\frac{\sqrt{2\log(N(c\,2^{-k}, A)^2)}}{m} = 6\,c\,2^{-k}\,\frac{\sqrt{\log(N(c\,2^{-k}, A))}}{m}.$$

Therefore,

$$R(A) \leq \frac{c\,2^{-M}}{\sqrt{m}} + \frac{6c}{m}\sum_{k=1}^{M} 2^{-k}\sqrt{\log(N(c2^{-k}, A))}.$$

□

As a corollary we obtain the following:

LEMMA 27.5 *Assume that there are $\alpha, \beta > 0$ such that for any $k \geq 1$ we have*

$$\sqrt{\log(N(c2^{-k}, A))} \leq \alpha + \beta k.$$

Then,

$$R(A) \leq \frac{6c}{m}\,(\alpha + 2\beta)\,.$$

Proof The bound follows from Lemma 27.4 by taking $M \to \infty$ and noting that $\sum_{k=1}^{\infty} 2^{-k} = 1$ and $\sum_{k=1}^{\infty} k2^{-k} = 2$. □

Example 27.2 Consider a set A which lies in a d dimensional subspace of $\mathbb{R}^m$ and such that $c = \max_{\mathbf{a}\in A} \|\mathbf{a}\|$. We have shown that $N(r, A) \leq \left(\frac{2c\sqrt{d}}{r}\right)^d$. Therefore, for any k,

$$\begin{aligned} \sqrt{\log(N(c2^{-k}, A))} &\leq \sqrt{d\log\left(2^{k+1}\sqrt{d}\right)} \\ &\leq \sqrt{d\log(2\sqrt{d})} + \sqrt{k\,d} \\ &\leq \sqrt{d\log(2\sqrt{d})} + \sqrt{d}\,k. \end{aligned}$$

Hence Lemma 27.5 yields

$$R(A) \;\leq\; \frac{6c}{m}\left(\sqrt{d\log(2\sqrt{d})} + 2\sqrt{d}\right) \;=\; O\left(\frac{c\sqrt{d\log(d)}}{m}\right).$$

27.3 Bibliographic Remarks

The chaining technique is due to Dudley (1987). For an extensive study of covering numbers as well as other complexity measures that can be used to bound the rate of uniform convergence we refer the reader to (Anthony & Bartlet 1999).