CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 18**

Covering numbers and metric entropy

Lecturer: Peter L. Bartlett *Scribe: Xiaoyue Zhao*

1 Growth function estimates for parameterized binary classes

Recall that the growth function for a class $H \subseteq \{\pm 1\}^{\mathcal{X}}$ is

$$\Pi_H(n) = \max\left\{\left|H_{|S}\right| : S \subseteq \mathcal{X},\, |S| = n\right\},$$

and a k-combination of $\mathrm{sign}(F)$ is a class of the form

$$H = \left\{x \mapsto g(\mathrm{sign}(f_1(a,x)), \ldots, \mathrm{sign}(f_k(a,x))) : a \in \Re^d\right\}$$

for some $g : \{\pm 1\}^k \to \{\pm 1\}$ and $f_1, \ldots, f_k \in F$. The prototypical example of such a class is a set of functions that can be computed with a small number of arithmetic operations and comparisons involving real numbers.

Theorem 1.1. If H is a k-combination of sign(F), then

$$\Pi_H(n) \leq \sum_{i=0}^{d} (\begin{pmatrix} kn \\ i \end{pmatrix} \sup_{f_j, x_i} CC(\cap_{j=1}^{i}\{a \in R^d : f_j(a, x_i) = 0\}),$$

where $CC(A)$ is the number of connected components of A.

The proof uses the following lemma, due to Warren, which relates the number of cells formed in $\Re^d$ by deleting some sets of solutions to certain equations to the number of connected components in intersections of these solution sets. The regularity condition forbids the solution sets from intersecting in non-generic ways; for instance, it forbids the dimensionality of the intersection of the tangent spaces at the intersection from being too large. (For details, see Anthony and Bartlett, *Neural Network Learning: Theoretical Foundations*, CUP 1999.) This regularity condition is generically true.

Lemma 1.2. If $f_1, \cdots, f_k$ have regular zero-set intersections, then the solution sets $Z_i = \{a \in R^d : f_i(a) = 0\}$ have

$$CC(R^d - \cup_{i=1}^{k} Z_i) \leq \sum_{I \subseteq \{1, \cdots, k\}} CC(\cap_{i \in I} Z_i).$$

Proof idea of the therorem:

Fix $\{x_1, \cdots, x_n\} \subseteq \mathcal{X}$ and $f_1, \cdots, f_k \in F$. Define the zero sets

$$Z_{ij} = \{a \in R^d : f_i(a, x_j) = 0\}, i = 1, \cdots, k, j = 1, \cdots, n.$$

These sets carve R^d into a number of cells. Consider the remainder when we remove these zero sets,

$$U = R^d - \cup_{i,j} Z_{i,j}.$$

For each distinct dichotomy of $S = \{x_1, \cdots, x_n\}$, choose a representative $a_1, \cdots, a_N \in R^d$. If two distinct representatives a_1, a_2 lie in U, then they must lie in distinct connected components of U. (If one of the a_i lies in one of the zero sets, we can always perturb the zero sets. We ignore these details here, but they are straightforward.) It follows that the growth function $\Pi_H(n)$ is no more than the number of connected components of the set U. Applying Warren's lemma, we have

$$\begin{aligned}\Pi_H(n) &\leq \sum_{i=0}^{n} \sum_{\substack{I \subseteq \{(1,1),\cdots,(k,n)\} \\ |I|=i}} CC\left(\bigcap_{i,j\in I} Z_{ij}\right) \\ &\leq \sum_{i=0}^{d} \binom{kn}{i} \sup_{f_j, x_i} CC\left(\bigcap_{j=1}^{i} \{a \in R^d : f_j(a, x_i) = 0\}\right),\end{aligned}$$

where we have used the fact that the intersection of more than d generic zero sets in R^d is empty. $\square$

As we have seen, this result has the consequence that any parameterized class of functions,

$$F = \left\{x \mapsto h(a,x) : a \in \Re^d\right\}$$

that uses d real parameters and can be computed in no more than t arithmetic operations on real numbers has $d_{VC}(H) \leq 4d(t+2)$. This is a powerful and broadly applicable result.

2 Covering numbers and metric entropy

In this section, we consider another way of measuring the complexity of a class of functions. We shall see that it is closely related to the Rademacher averages.

Definition. For a metric space (S, ρ) and a subset $T \subseteq S$, we say that $\hat{T} \subseteq T$ is an ε-cover of T if

$$\forall t \in T, \exists \hat{t} \in \hat{T}, \text{with } \rho(t, \hat{t}) < \varepsilon.$$

The ε-covering number of T is

$$N(\varepsilon, T, \rho) = \min\{|\hat{T}|, \hat{T} \subseteq T \text{ is a } \varepsilon - \text{cover of T}\}.$$

The entropy function for T is $\log N(\varepsilon, T, \rho)$.

We can think of the entropy function as (proportional to) the number of bits required to identify an element of the cover $\hat{T}$, or equivalently to identify an element of T within ϵ.

(As an aside, notice that our definition is for what is sometimes called a proper cover, since we insist that the cover is a subset of T. We can define and alternative notion of covering numbers—call then N'—without this restriction, and it is closely related: $N'(\epsilon, T) \leq N(\epsilon, T) \leq N(\epsilon/2, T)$.)

Example. Consider a function class $F \subseteq [0,1]^x$, with the metric of $L_2(P_n)$, $\rho(f,g) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2}$.

Equivalently, we can consider the set of restrictions of functions in F to the support of P_n, $T \subseteq [0,1]^n$, equipped with the metric $\rho(a,b) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2}$.

Suppose that F, and hence T lies in a d-dimensional subspace. Then it is easy to see that

$$N(\varepsilon, T, \rho) \leq (\frac{c}{\varepsilon})^d, \quad \log N(\varepsilon, T, \rho) \leq d \log(\frac{c}{\varepsilon}).$$

(We just choose a basis, and make an ϵ-grid.)

Theorem 2.1. $F \subseteq [-1,1]^x, x_1, \cdots, x_n \in \mathcal{X}$

$$\hat{R}_n(F) \leq \inf_{\alpha}\left(\sqrt{\frac{2\log N(\alpha, F, L_2(P_n))}{n}} + \alpha\right).$$

Proof: Fix $\alpha > 0$, let $\hat{F}$ be a minimal α-cover of F,

$$\begin{aligned}
\hat{R}_n(F) &= E\sup_{f\in F}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right) \\
&= E\sup_{\hat{f}\in\hat{F}}\sup_{f\in F\cap B_\alpha(\hat{f})}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \hat{f}(X_i) + \frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - \hat{f}(X_i))\right) \\
&\leq E\sup_{\hat{f}\in\hat{F}}\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i \hat{f}(X_i) + \alpha\right) \\
&= \hat{R}_n(\hat{F}) + \alpha \\
&\leq \sqrt{\frac{2\max_{\hat{f}\in\hat{F}}||\hat{f}||^2_{L_2(P_n)}\log|\hat{F}|}{n}} + \alpha \\
&\leq \sqrt{\frac{2\log N(\alpha, F, L_2(P_n))}{n}} + \alpha,
\end{aligned}$$

where the first inequality follows from Cauchy-Schwartz in $L_2(P_n)$, and the second inequality is Massart's finite class lemma. □

Example. $\log N(\alpha, F, L_2(P_n)) \leq d\log\frac{1}{\alpha}$. Set $\alpha = \frac{1}{\sqrt{n}}$. $\hat{R}_n(F) \leq \sqrt{\frac{d\log n}{n}} + \frac{1}{\sqrt{n}}$.

As another example, if $F_{|S} = [-1,1]^n$, then $N(\alpha, F, L_2(P_n)) \approx \alpha^{-n}$. In this case, the log covering numbers grow linearly with n, so the upper bound on the Rademacher averages does not decrease with n. We can view the ratio $\log(N(\alpha, F, L_2(P_n)))/n$ as a measure of how much the restrictions $F_{|S}$ 'fills up' the n-cube, $[-1,1]^n$.

Example. $F \subseteq \{\pm\}^x$, $F_{|S} \subseteq \{\pm 1\}^n$.

$$\lim_{\alpha\to 0} N(\alpha, F, L_2(P_n)) = |F_{|S}| \leq \Pi_F(n).$$

Choosing $\alpha \approx 0$ shows that

$$\hat{R}_n(F) \leq c\sqrt{\frac{d_{VC}(F)\log(\frac{n}{d_{VC}(F)})}{n}}.$$

Example. F: nondecreasing functions from R to [0,1]. We can define a cover of $F_{|S}$ as follows. Discretize the range to an α-grid. Then for each of the values y in the grid, specify the smallest x_i for which $f(x_i) \geq y$. This specifies any non-decreasing function to within α. (Actually, it is uniformly within α, which implies it is that close in $L_2(P_n)$.) And the size of this cover is $n^{1/\alpha}$. Thus we have

$$N(\alpha, F, L_2(P_n)) \leq n^{\lceil\frac{1}{\alpha}\rceil+1}, \quad \log N(\alpha, F, L_2(P_n)) \leq c\frac{\log n}{\alpha}.$$

So

$$\begin{aligned}
\hat{R}_n(F) &= \inf_{\alpha}\sqrt{\frac{C\log n}{\alpha n}} + \alpha \\
&\leq C'\left(\frac{\log n}{n}\right)^{\frac{1}{3}}, \quad \text{set } \alpha = \left(\frac{\log n}{n}\right)^{\frac{1}{3}}.
\end{aligned}$$

3 Chaining and Dudley's entropy integral:

It turns out that this easy upper bound on the Rademacher averages can be loose. We shall see that, for the example of nondecreasing functions it gives the wrong rate. The problem arises because our argument involves finding a single scale α and thresholding at that value: the behavior of functions below that scale is estimated by a term α, and the richness of functions above that scale is estimated by the size of the minimal α-cover. There is a more refined approach that gives better bounds. It involves successively refining the scale at which we examine our functions. This technique, called chaining, leads to the following improvement.

Theorem 3.1. (Dudley). For $F \subseteq R^x$, $x_1, \cdots, x_n \in \mathcal{X}$,

$$\begin{aligned}\hat{R}_n(F) &= E \sup_{f \in F} \frac{1}{n} \sum \varepsilon_i f(X_i) \\ &\leq 12 \int_0^\infty \sqrt{\frac{\log(N(\alpha, F, L_2(P_n)))}{n}} d\alpha.\end{aligned}$$