CS281B/Stat241B (Spring 2003) **Lecture: 17**

Growth function estimates for parameterized binary classes

Lecturer: Peter Bartlett *Scribe: Yuan Yao*

1 Rademacher Complexity for Kernel Classes

Recall that in the last lecture, we mentioned the following result for kernel function classes.

Theorem 1.1. Let $\mathcal{H}$ be a reproducing kernel Hilbert space with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For the sample set, $S = \{x_1, x_2, \ldots, x_n\}$, define the kernel matrix as $K_{ij} = K(x_i, x_j)$. Consider $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq \lambda\}$. Then

$$\hat{R}_n(F) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i) | S\right] \leq \frac{\lambda \sqrt{\mathrm{trace}(K)}}{n}$$

and

$$R_n(F) := \mathbb{E}\hat{R}_n(F) \leq \lambda \sqrt{\frac{\sum_{j=1}^{\infty} \theta_j}{n}}$$

where $\theta_1 \geq \theta_2 \geq \theta_3 \geq \ldots$ the eigenvalues of the integral operator

$$T_k : f \mapsto \int_{\mathcal{X}} k(\cdot, x) f(x) dx,$$

such that $k(x_1, x_2) = \sum_{j=1}^{\infty} \theta_j \psi_j(x_1) \psi_j(x_2)$.

Proof. Since

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) &= \sup_{\|f\|_{\mathcal{H}} \leq \lambda} \left\langle \frac{1}{n} \sum_{i=1}^{n} \epsilon_i k(x_i, \cdot), f(\cdot) \right\rangle \\
&= \lambda \left\| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i K(x_i, \cdot) \right\| \\
&= \lambda \sqrt{\frac{1}{n^2} \sum_{i,j=1}^{n} \epsilon_i \epsilon_j K(x_i, x_j)} \\
&= \frac{\lambda}{n} \sqrt{\epsilon^T K \epsilon},
\end{aligned}$$

then

$$\begin{aligned}
\hat{R}_n(F) &:= \mathbb{E}_{\epsilon|K} \frac{\lambda}{n} \sqrt{\epsilon^T K \epsilon | K} \\
&\leq \frac{\lambda}{n} \sqrt{\mathbb{E}_{\epsilon|K} [\epsilon^T K \epsilon | K]} \\
&= \frac{\lambda}{n} \sqrt{\sum_{i=1}^{n} K(x_i, x_i)} = \frac{\lambda}{n} \sqrt{\mathrm{trace}(K)},
\end{aligned}$$

and

$$\begin{aligned} R_n(F) &:= \mathbb{E}\hat{R}_n(F) \\ &\leq \frac{\lambda}{n}\sqrt{\mathbb{E}\sum_{i,j=1}^{n}\epsilon_i\epsilon_j K(X_i, X_j)} \\ &= \frac{\lambda}{n}\sqrt{n\mathbb{E}K(X_i, X_i)} \\ &= \lambda\sqrt{\frac{\mathbb{E}\sum_{j=1}^{n}\infty\theta_j\psi_j^2(X_i)}{n}} \\ &= \lambda\sqrt{\frac{\sum_{j=1}^{n}\theta_j}{n}}, \qquad \text{since } \psi_j \text{ is orthonormal.} \end{aligned}$$

□

2 Growth Function for Parameterized Binary Function Classes

Recall that when $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$, we have a similar result to Massart's Finite Set Lemma,

$$R_n(F) \leq \sqrt{\frac{2\log \Pi_F(n)}{n}}$$

with the *growth function of $\mathcal{F}$ restricted on S*

$$\Pi_F(n) = \max\{|F_{|S}| : S \subseteq \mathcal{X}, |S| \leq n\}.$$

Now we are going to study the upper bounds for the growth function $\Pi_F(n)$ when F has special parameterized forms.

Proposition 2.1. For the set $\mathcal{F}$ of *linear threshold functions* on $\mathbb{R}^d$,

$$\mathcal{F} = \{x \mapsto \text{sign}(w^T x - \theta) : w \in \mathbb{R}^d, \theta \in \mathbb{R}\} = \{x \mapsto f(\alpha, x) : \alpha \in \mathbb{R}^{d+1}\}$$

we have

$$\Pi_F(n) = 2\sum_{k=0}^{d}\binom{n-1}{k}.$$

Proof. **Sketch.** Fix an arbitrary set $S = \{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{X} = \mathbb{R}^d$.

1. Assume S is in *general position*, that is, every subset of at most $d+1$ points from $\{(x_1, 1)^T, \ldots, (x_n, 1)^T\}$ is linearly independent. Then

$$\mathcal{F}_{|S} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}.$$

This will not lose generality since arbitrary small perturbation will change the points not in general position to general position.

2. Consider the parameter space $\mathbb{R}^{d+1} = \{(w, \theta) : w \in \mathbb{R}^d, \theta \in \mathbb{R}\}$. For each $x_i \in S$, define

$$P_i = \{(w, \theta) : w^T x_i - \theta = 0\},$$

which is a hyperplane through origin in $\mathbb{R}^{d+1}$.

- (w, θ) and (w', θ') give distinct labelings of S iff they are on "opposite" sides of some P_i, so

$$\Pi_{\mathcal{F}}(n) = CC(\mathbb{R}^{d+1} - \cup_{i=1}^{n} P_i) =: C(n, d+1), \quad CC = \text{number of connected components}$$

3. Suppose we have n hyperplanes through origin in $\mathbb{R}^d$, which divide $\mathbb{R}^d$ into $C(n, d)$ connected components (cells). Adding a new hyperplane through origin might intersect some cells and divide each of them into half. Since each of the cell intersecting with the new hyperplane leaves a connected component as the trace restricted on the new hyperplane (with dimension $d-1$), then the number of such cells equal $C(n, d-1)$, the number of connected components on the new hyperplane formed by the old n hyperplanes. Thus

$$C(n+1, d) = C(n, d) + C(n, d-1)$$

Notice that $C(1, d) = 2$ and $C(1, 1) = 2$, we obtain the result. (For more detail of the proof, see pp. 156-159, *Statistical Learning Theory*, by V. N. Vapnik.) □

Remark. For the more general parameterized classes, for example

$$\mathcal{F} = \{x \mapsto \text{sign}(f(a, x)) : a \in \mathbb{R}^d, f : \mathbb{R}^d \times \mathcal{X} \to \mathbb{R}\},$$

it might be natural to ask if $d_{VC}(F)$ depends on d for smoothly parameterized f. Unfortunately the following counterexample shows that a function smoothly dependent on one parameter may have infinite VC-dimension.

Example. **(Counterexample 1)** The one-parameter function class

$$\mathcal{F} = \{f : f(w, x) = \sin wx, w \in \mathbb{R}\}$$

has VC-dimension

$$d_{VC}(F) = \infty.$$

To see this, choose $x_i = 2^i$, $(i = 1, \ldots, n)$. For $(b_0, \ldots, b_n) \in \{0, 1\}^n$, set $w = \pi c$ with binary representation of c

$$c = 0b_1 b_2 \cdots b_n 1.$$

Then

$$\sin(wx_i) = \sin(\pi \times 2^i \times (0b_1 b_2 \cdots b_n 1)) \in (b_i \pi, (b_i + 1)\pi))$$

which implies $\text{sign}(\sin(wx_i)) = 1 - 2b_i$.

We have another example.

Example. **(Counterexample 2)** Consider the neural network

$$f(a, x) = \sum_{i=1}^{k} \alpha_i \sigma(v_i^T x) + \alpha_0.$$

The question is, for what σ is $d_{VC}(F) < \infty$? Actually, there exists a σ, which is analytic, increasing, convex on $x < 0$ with $\lim_{x \to -\infty} \sigma(x) = 0$, and concave on $x > 0$ with $\lim_{x \to +\infty} \sigma(x) = 1$, such that $d_{VC}(F) = \infty$ for $k = 2$. One example is

$$\sigma(x) = \frac{1}{1 + e^{-x}} + cx^3 e^{-x^2} \sin x.$$

The following will study the upper bound for the growth function of a rather general binary function class.

Definition. Let F be a class of real functions on $\mathbb{R}^d \times \mathcal{X}$. Then H is a *k-combination* of $\mathrm{sign}(F)$ if

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

for some $g : \{\pm 1\}^k \to \{\pm 1\}$.

Example. If $h : \mathbb{R}^d \times \mathbb{R}^n \to \{\pm 1\}$ can be computed by an algorithm that takes as input $(a,x) \in \mathbb{R}^d \times \mathbb{R}^n$ and returns $h(a,x)$ after at most t operations of the following kinds

- arithmetic operations on real numbers $\{+, -, \times, /\}$;
- jumps conditioned on comparisons $\{>, <, \geq, \leq, =, \neq\}$;
- output $\{\pm 1\}$.

Then h is a k-combination of $\mathrm{sign}(F)$, for $k \leq 2^t$, and $F =$ polynomials of degree at most 2^t. The following theorem gives the upper bound on the growth function by the number of connected components of algebraic varieties, i.e. the complexity aspect of Bezout Theorem.

Theorem 2.2.

$$\Pi_H(n) \leq \sum_{i=0}^{d} \binom{kn}{i} \sup_{f_j \in F, x_j \in \mathcal{X}, \epsilon_j \in \mathbb{R}} CC\left(\bigcap_{j=1}^{i} \{a : f_j(a, x_j) + \epsilon_j = 0\}\right)$$

Corollary 2.3. If F is a polynomially parameterized class of degree at most m, then

$$\Pi_H(n) \leq 2\left(\frac{2enkm}{d}\right)^d$$

and

$$d_{VC}(H) \leq 2d \log(ekm).$$

Hence, out t-step computable H has

$$d_{VC}(H) \leq 4d(t+2).$$

Remark. Adding $x \mapsto e^x$ to the allowable operation set will change $d_{VC}(H) \leq ct^2d^2$.