CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 18**

$\Pi_F(n)$ for parameterized F, covering numbers, $R_n(F)$

Lecturer: Peter Bartlett *Scribe: Eilyan Bitar*

1 $\Pi_F(n)$ for parameterized F

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

For a family of classifiers F , linear in a, we have $d_{vc}(F) = d$, where $d =$ the number of parameters.

Example. $f(w,x) = \sin(wx) \Rightarrow d_{vc}(F) = \infty$, even though f is smooth.

Set $w = \pi c$, where c has a binary representation $0.b_1 b_2 \ldots.b_n 1$.

Set $x_i = 2^i$ for $i = 1, \ldots, n$. Then

$$\begin{aligned}\sin(wx_i) &= \sin(2^i \times \pi \times 0.b_1 \ldots.b_n 1)\\ &= \sin(\pi \times b_1 \ldots.b_i.b_{i+1} \ldots.b_n 1)\\ &= \sin(\pi \times b_i.b_{1+1} \ldots.b_n 1)\end{aligned}$$

which implies that $\text{sign}(\sin(wx_i)) = b_i$. Hence, we can always find a set of size n $\forall n$.

Example (Neural Nets)**.**

$$f(\theta, x) = \sum_{i=1}^{k} \alpha_i \underbrace{\sigma(\beta_i^T x)}_{\substack{\text{squashing}\\ \text{function}}} + \alpha_o$$

For what $\sigma : \mathbb{R} \to [0,1]$ is $d_{vc}(F) < \infty$?

For instance, if $\sigma(\alpha) = \underbrace{\frac{1}{1+e^{-\alpha}} + c\alpha^3 e^{-\alpha^2} \sin(\alpha)}_{\substack{\text{Looks like a sigmoid but}\\ \text{has a sinusoid hidden in it}}}$, we have $d_{vc}(F) = \infty$. Take note that σ is convex left of zero and concave right of zero.

Consider the function $h : \underbrace{\mathbb{R}^d}_{a} \times \underbrace{\mathbb{R}^m}_{x} \to \{+-1\}$ that can be computed by an algorithm that takes as input, $(a, x) \in \mathbb{R}^d \times \mathbb{R}^m$, and returns as $h(a, x)$ after $\leq t$ operations:

- arithmetic, $(+, -, \times, \div)$
- conditionals $(<, >, \leq, \geq)$
- outputs ± 1

Definition. For a class, F, of real valued functions on $\overbrace{\mathbb{R}^d}^{\textbf{cont. in a}} \times \mathcal{X}$, we say h is a <u>$k-$*combination*</u> of $\mathrm{sign}(F)$ if:

$\mathcal{H} = \{x \to g(\mathrm{sign}(f_1(a, x)), ..., \mathrm{sign}(f_k(a, x))) \mid a \in \mathbb{R}^d\}$ for fixed $g : \{\pm 1\}^k \to \{\pm 1\}$ and $f_1, ..., f_k \in F$.

E.g. For a $t-$*step* computable h, we have a 2^t-combination of $\mathrm{sign}(F)$ for $F =$ polynomials of degree $\leq 2^t$.

Theorem 1.1. For H a $k-$*combination* of $\mathrm{sign}(F)$,

$$\Pi_H(n) \;\leq\; \sum_{i=0}^{d} \binom{kn}{i} \max_{\{f_j\} \in F, \{x_j\} \in \mathcal{X}} \underbrace{CC\left(\bigcap_{j=1}^{i} \{a \mid f_j(a, x_j) = 0\}\right)}_{\text{number of connected components in the solution set}}$$

Example. Linear threshold function ($1-$*combination* of $\mathrm{sign}(F)$)

- f_j is linear in a.
- $\underbrace{CC\left(\bigcap_{j=1}^{i} \{a \mid f_j(a, x_j) = 0\}\right)}_{\text{defines a subspace}} = 0 \text{ or } 1$

Corollary 1.2. For F, polynomialy parameterized, with degree $\leq m$, we have

$$\begin{aligned} \Pi_H(n) &\leq 2\left(\frac{2enkm}{d}\right)^d \\ d_{vc}(H) &\leq 2d \log(2ekm) \end{aligned}$$

• Hence, $t - step$ computable, H has $d_{vc}(H) \leq 4d(t+2)$
(using $\Pi_H(n) < 2^n \Rightarrow d_{vc}(H) < n$).

Note: With the addition of exponentials in the model of computation, we have $d_{vc}(H) = O(t^2 d^2)$.

Proof. Proof idea of previous theorem.

• $\Pi_H(n) = \max\{|H_{|S}| \ : \ S \subseteq \mathcal{X}, |S| = n\}$
• $Z_{ij} = \{a \mid f_i(a, x_j) = 0\}$, assume regular intersections between these subspaces.

Lemma 1.3 (Warren 1960)**.**

$$CC\left(\mathbb{R}^d - \bigcup_{i,j} Z_{ij}\right) \leq \sum_{I \subseteq \{(i,j)\}} CC\left(\bigcap_{i \in I} Z_i\right)$$

□

Summarize: $d_{vc}(H) = O(dt)$ for $t - step$ computeable h: $2^t - combination$ of $sign(F)$ for $F =$ polynomial with degree $\leq 2^t$.

2 Covering Numbers

Definition. For a metric space, (S, ρ), and a subspace, $T \subseteq S$, we say that $\hat{T}$ is an $\underline{\varepsilon - cover}$ of T if $\forall\, t \in T,\ \exists\, \hat{t} \in \hat{T}$ such that $\rho(t, \hat{t}) < \varepsilon$.

Definition. The $\underline{\varepsilon - \text{covering number}}$ of (T, ρ):

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

Note: *Entropy* $:= \log N(\varepsilon, T, \rho)$

Example. $T \subseteq [0,1]^n$ is a d-dimensional subspace. A bound on the covering number for this subspace can be found in terms of a uniform grid of ε-balls over the subspace, i.e.

$N(\varepsilon, T, L_2(P_n)) \leq \left(\frac{1}{\varepsilon}\right)^d$

Consider,

- $F \subseteq [-1,1]^{\mathcal{X}}$
- $S = \{x_1, ..., x_n\} \subseteq \mathcal{X}$.
- $F_{|s} = \{(f(x_1), ..., f(x_n) \mid f \in F\} \subseteq [-1,1]^n$.
- $L_2(\hat{P}), \quad \rho(u,v) = \left(\frac{1}{n}\sum_i (u_i - v_i)^2\right)^{1/2}$.

Theorem 2.1.

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

Proof. Fix α, α-cover $\hat{F} of F$.

$$\begin{aligned}
\hat{R}_n(F) &= \mathbb{E}_\varepsilon \sup_{f\in F} \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i) \\
&= \mathbb{E} \sup_{\hat{f}\in\hat{F}} \sup_{f\in F\cap B_\alpha(\hat{f})} \left(\frac{1}{n}\sum \varepsilon_i \hat{f}(x_i) + \underbrace{\frac{1}{n}\sum \varepsilon_i (f(x_i) - \hat{f}(x_i))}_{\langle \underbrace{\varepsilon}_{||\varepsilon||=1}, \underbrace{f-\hat{f}}_{\|\cdot\|\leq\alpha}\rangle_{L_2(\hat{P})}}\right) \\
&\leq \mathbb{E}\left[\sup_{\hat{f}\in\hat{F}}\left(\frac{1}{n}\sum \varepsilon_i \hat{f}(x_i)\right) + \alpha\right]
\end{aligned}$$

Note:

- $F = \bigcup_{\hat{f}\in\hat{F}} (F \cap B_\alpha(\hat{f}))$
- $|\hat{F}| = N(\alpha, F, L_2(\hat{P}))$

□

$\Rightarrow \log N(\alpha, F) = d\log(1/\alpha)$ for the linear case.

Set $\alpha = \frac{1}{\sqrt{n}}, \Rightarrow R_n(F) \leq \sqrt{\frac{2d\log(n)}{n}} + \frac{1}{n}$