CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 15**

Rademacher averages and Vapnik-Chervonenkis dimension

Lecturer: Peter Bartlett *Scribe: Taylor Berg-Kirkpatrick*

1 Intro

Risk bounds, complexity

- Rademacher averages
- Vapnik-Chervonenkis dimension

Last time: Choose $\hat{f} \in F$ so that sample average of the loss is minimized.

$$\hat{f} = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

Interested in how the risk of f compares to the best risk in our class.

$$R(\hat{f}) - \inf_{f \in F} R(f)?$$

One approach is to find bound that holds uniformly over class.

$$\sup_{f \in F} \left| R(f) - \hat{R}(f) \right| \leq \dots$$

2 Rademacher averages

Rademacher averages are a measure of the complexity of a class.

Definition. For a class $F \subseteq \mathbb{R}^{\mathcal{X}}$, i.i.d. $X_1, \dots, X_n$, and Rademacher R.V.s $\epsilon_1, \dots, \epsilon_n$ (i.e. i.i.d. taking on ± 1 with equal probability), define the Rademacher averages of F as

$$R_n(F) = \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i)$$

Want to bound

$$\sup_{f \in F} \left(\mathbb{E}f - \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right)$$

McDiarmid $\Rightarrow$

$$\sup_{f\in F}\left(\mathbb{E}f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right)\leq\mathbb{E}\sup_{f\in F}\left(\mathbb{E}f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right)+c\sqrt{\frac{\log\frac{1}{\delta}}{n}}$$

We saw this idea in the GC theorem, dealt with awkward expectation using symmetrization.

But for $F\subseteq\mathbb{R}^{\mathcal{X}}$,

$$\begin{aligned}\mathbb{E}\sup_{f\in F}\left(\mathbb{E}f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right)&=\mathbb{E}\sup_{f\in F}\mathbb{E}\left[\frac{1}{n}\sum_{1=1}^{n}f(X_i')-\frac{1}{n}\sum_{i=1}^{n}f(X_i)|X_1,\dots,X_n\right]\\&\leq\mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{1=1}^{n}\left(f(X_i')-f(X_i)\right)\\&=\mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{1=1}^{n}\epsilon_i\left(f(X_i')-f(X_i)\right)\\&\leq\mathbb{E}\left[\sup_{f\in F}\left(\frac{1}{n}\sum_{1=1}^{n}\epsilon_i f(X_i')\right)+\sup_{f\in F}\left(-\frac{1}{n}\sum_{1=1}^{n}\epsilon_i f(X_i)\right)\right]\\&=2R_n(F)\end{aligned}\tag{1}$$

Combining yeilds:

Theorem 2.1. For $F\subseteq[-1,1]^{\mathcal{X}}$, *w.p.* $\geq 1-\delta$

$$\sup_{f\in F}\left(\mathbb{E}f-\frac{1}{n}\sum_{i=1}^{n}f(X_i)\right)\leq 2R_n(F)+c\sqrt{\frac{\log\frac{1}{\delta}}{n}}$$

Hence, for

$$\hat{f}=\arg\min_{f\in F}\hat{\mathbb{E}}\ell_f$$
$$f^*=\arg\min_{f\in F}\mathbb{E}\ell_f$$

w.p. $\geq 1-\delta$,

$$\begin{aligned}\mathbb{E}\ell_{\hat{f}}&\leq\hat{\mathbb{E}}\ell_{\hat{f}}+2R_n(\ell_F)+c\sqrt{\frac{\log\frac{1}{\delta}}{n}}\\&\leq\hat{\mathbb{E}}\ell_{f^*}+2R_n(\ell_F)+c\sqrt{\frac{\log\frac{1}{\delta}}{n}}\\&\leq\mathbb{E}\ell_{f^*}+2R_n(\ell_F)+c'\sqrt{\frac{\log\frac{1}{\delta}}{n}}\end{aligned}\tag{2}$$

where the last inequality follows from an application of Hoeffding's inequality to ℓ_{f^*}.

i.e.

$$R(\hat{f}) \leq \inf_{f \in F} R(f) + 2R_n(\ell_F) + c'\sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

Example. (Rademacher average of binary class versus Rademacher average of discrete loss class) Classification
Let $F \subseteq \{\pm 1\}^{\mathcal{X}}$, $\ell =$ 0-1 loss, $\ell_f(x,y) = \frac{1-yf(x)}{2}$, so

$$\begin{aligned}
R_n(\ell_F) &= \mathbb{E}\left[\sup_{f \in F} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i \left(\frac{1 - Y_i f(X_i)}{2}\right)\right] \\
&= \mathbb{E}\mathbb{E}\sup_{f \in F}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i}{2} - \frac{\frac{1}{n}\sum_{i=1}^{n}\epsilon_i Y_i f(X_i)}{2} | X_1, \ldots, X_n, Y_1, \ldots, Y_n\right] \\
&= \frac{1}{2}\mathbb{E}\sup_{f \in F}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right) \\
&= \frac{1}{2}R_n(F) \qquad (3)
\end{aligned}$$

Recall:

Lemma 2.2. For $A \subseteq \mathbb{R}^n$ with $\max_{a \in A} \frac{1}{n}\sum_{i=1}^{n} a_i^2 = R^2$,

$$\mathbb{E}\max_{a \in A} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i a_i \leq R\sqrt{\frac{2\log|A|}{n}}$$

Hence for finite $F \subseteq [-1,1]^{\mathcal{X}}$,

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

For example, consider a class that is parameterized by k bits,

$$F = \{x \mapsto f(x,\theta) : \theta \in \{0,1\}^k\}.$$

Then $R_n(F) \leq \sqrt{2(k/n)\log 2}$.

3 Growth function & Vapnik-Chervonenkis dimension

Pattern classification: $F \subseteq \{\pm 1\}^{\mathcal{X}}$, $\ell = $ 0-1 loss,

$$\begin{aligned} R_n(F) &= \mathbb{E}\mathbb{E}\left[\max_{a \in F\restriction_{X_1^n}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i | X_1^n\right] \\ &\le \sqrt{\frac{2}{n}}\mathbb{E}\sqrt{\log |F\restriction_{X_1^n}|} \\ &\le \sqrt{\frac{2\log \mathbb{E}|F\restriction_{X_1^n}|}{n}} \end{aligned} \tag{4}$$

Example. $F = \{x \mapsto 1[x \le \theta] : \theta \in \mathbb{R}\}$, and $x_{(1)} \le x_{(2)} \le x_{(3)} \le \ldots$, so F can only split the $x_{(i)}$'s n+1 ways.

Definition. The growth function of $F \subseteq \{\pm 1\}^{\mathcal{X}}$ is

$$\Pi_F(n) = \max\{|F\restriction_{x_1^n}| : \{x_1, \ldots, x_n\} \subseteq \mathcal{X}\}$$

For the example above, the growth function is $n + 1$.

Some observations:

$$\Pi_F(n) \le |F|$$

(and for $|F| < \infty$, $\lim_{n\to\infty} \Pi_F(n) = |F|$)

$$\Pi_F(n) \le 2^n$$

Also

$$R(\hat{f}) \le R(f^*) + 2R_n(F) + \frac{c}{\sqrt{n}}$$

where

$$\begin{aligned} 2R_n(F) &\le c\mathbb{E}\sqrt{\frac{\log|F\restriction_{X_1^n}|}{n}} \\ &\le c\sqrt{\frac{\log \Pi_F(n)}{n}} \end{aligned} \tag{5}$$

e.g. We'll see that linear threshold functions on $\mathbb{R}^d$ have growth function

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

Definition. We say F shatters $S \subseteq \mathcal{X}$ if $|F\restriction_S| = 2^{|S|}$.
The VC dimension of F is

$$\begin{aligned} d_{VC}(F) &= \max\{|S| : S \subseteq \mathcal{X}, \text{F shatters S}\} \\ &= \max\{n : \Pi_F(n) = 2^n\} \end{aligned} \tag{6}$$

e.g. For linear threshold functions

$$2\sum_{k=0}^{d}\binom{n-1}{k} = \begin{cases} 2\sum_{k=0}^{n-1}\binom{n-1}{k} - 2\sum_{k=d+1}^{n-1}\binom{n-1}{k} & \text{where } d < n-1, \\ 2^n - (\ldots) & \text{when } d \geq n-1. \end{cases}$$

$$\Pi_f(n) \leq 2^n \Leftrightarrow n > d+1, \text{ i.e. } d_{VC}(F) = d+1$$

We can calculate the VC-dimension more directly, as follows.

$$\{x_1, \ldots, x_n\} \text{ shattered by } \{x \mapsto sign(\theta' x + \theta_0)\} \Leftrightarrow \left\{\begin{pmatrix} x_1 \\ 1 \end{pmatrix}, \ldots, \begin{pmatrix} x_n \\ 1 \end{pmatrix}\right\} \text{ linearly independent.}$$

Therefore, since max basis of $\mathbb{R}^d + 1$ has size $d+1$, $d_{VC}(F) = d+1$.

Note:

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

Lemma 3.1. [Sauer's Lemma]

$$d_{VC}(F) \leq d \Rightarrow \Pi_F(n) \leq \sum_{i=0}^{d}\binom{n}{i}$$

$$\text{and for } n \geq d, \text{ this is } \leq \left(\frac{en}{d}\right)^d$$