CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 18**

Rademacher Averages

Lecturer: Peter Bartlett *Scribe: Rosen Diankov*

1 Introduction

Definition. For labeled data $\mathcal{X} = \{(X_1, Y_1), ...(X_n, Y_n)\}$, a function class F, and uniformly distributed random variables $\epsilon_i \in \{\pm 1\}$, the Rademacher average for F is

$$\hat{R}_n(F) = \mathbb{E}\left[\sup_{f \in F} \frac{1}{n} \sum_i \epsilon_i f(X_i) \mid \mathcal{X}\right] \tag{1}$$

$$R_n(F) = \mathbb{E}\,\hat{R}_n(F) \tag{2}$$

It measures how well functions in F can align to the random directions of ϵ_i. We can use the results from the bounded difference concentration inequality to get a bound on the error on the functions. Recall that with probability $\geq 1 - e^{-2\epsilon^2 n}$,

$$\sup_{f \in F} |\mathbb{E}\, f - \hat{\mathbb{E}}_n f| \leq \mathbb{E} \sup_{f \in F} |\mathbb{E} f - \hat{\mathbb{E}}_n f| + \epsilon \tag{3}$$

Theorem 1.1. For $F \subseteq \mathbb{R}^X$,

$$\mathbb{E} \sup_{f \in F} \left(\mathbb{E} f - \frac{1}{n} \sum_i f(X_i)\right) \leq 2R_n(F) \tag{4}$$

$$= \mathbb{E} \sup_{f \in F} \frac{2}{n} \sum_i \epsilon_i f(X_i) \tag{5}$$

where last expectation is over ϵ_i's and X_i's.

Proof. Introduce independent (of each other and everything else) X_i''s with the same distribution as the x_i's. Then, we have (with $\mathbb{E}'$ denoting expectation w.r.t. X_i''s):

$$\mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_i (\mathbb{E} f - f(X_i))$$

$$= \mathbb{E} \sup_{f \in F} \mathbb{E}' \left[\frac{1}{n} \sum_i (f(X_i') - f(X_i))\right]$$

$$\leq \mathbb{E}\mathbb{E}' \sup_{f \in F} \frac{1}{n} \sum_i (f(X_i') - f(X_i))$$

$$
\begin{aligned}
&= \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_i \left(f(X_i') - f(X_i) \right) \\
&= \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_i \epsilon_i \left(f(X_i') - f(X_i) \right) \quad , \quad \epsilon_i \in \{\pm 1\} \\
&\leq \mathbb{E} \sup_{f \in F} \frac{2}{n} \sum_i \epsilon_i f(X_i) = 2R_n(F)
\end{aligned}
$$

$\square$

By symmetry,

$$
\mathbb{E} \sup_{f \in F} \left(\frac{1}{n} \sum_i f(X_i) - \mathbb{E}f \right) \leq 2R_n(F) \ .
$$

Theorem 1.2. For $F \subseteq [0,1]^X$, with probability $\geq 1 - \delta$,

$$
\sup_{f \in F} \mathbb{E}f - \frac{1}{n} \sum_i f(X_i) \ \leq \ 2R_n(F) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \tag{6}
$$

$$
= (\text{complexity of } F) + \Theta\left(\frac{1}{\sqrt{n}} \right) \tag{7}
$$

Also notice that

$$
\sup \frac{1}{n} \sum_i \epsilon_i f(X_i)
$$

satisfies the bounded differences condition with $c_i = \frac{2}{n}$ and so is concentrated. Thus, w.p. $\geq 1 - \delta$,

$$
\sup_{f \in F} \mathbb{E}f - \hat{\mathbb{E}}f \leq \sup_{f \in F} \frac{2}{n} \sum_i \epsilon_i f(X_i) + c\sqrt{\frac{\log 1/\delta}{n}} \ .
$$

The main concern is with loss functions. Let $\ell_F = \{\ell_f : \ell_f(X,Y) = \ell(Y, f(X)), f \in F\}$. If $\hat{f} = \operatorname{argmin} \hat{\mathbb{E}}\ell_f$ and $F \subseteq [0,1]$, then with probability $\geq 1 - \delta$

$$
L(\hat{f}) = \mathbb{E}\ell_{\hat{f}} = \left(\mathbb{E}\ell_{\hat{f}} - \hat{\mathbb{E}}\ell_{\hat{f}} \right) + \underbrace{\left(\hat{\mathbb{E}}\ell_{\hat{f}} - \hat{\mathbb{E}}\ell_{f^*} \right)}_{\leq 0} + \underbrace{\left(\hat{\mathbb{E}}\ell_{f^*} - \mathbb{E}\ell_{f^*} \right)}_{\text{bound using Hoeffding's}} + \mathbb{E}\ell_{f^*} \tag{8}
$$

$$
\leq 2R_n(\ell_F) + c\sqrt{\frac{\log \frac{1}{\delta}}{n}} + \inf_{f \in F} L(f) \tag{9}
$$

Note that $R_n(\ell_F)$ is different from $R_n(F)$. However, can often find ways to relate the two. For classification where $Y_i \in \{\pm 1\}$, $F \subseteq \{\pm 1\}^X$, $\ell(Y, \hat{Y}) = 1[Y \neq \hat{Y}]$, write $\ell(Y, \hat{Y}) = \frac{1 - Y\hat{Y}}{2}$. Then,

$$R_n(\ell_F) = \mathbb{E}\left[\mathbb{E}\,[\sup_{f\in F}\frac{1}{n}\sum_i^n \epsilon_i\left(\frac{1-Y_i f(X_i)}{2}\right)]|(X_i,Y_i)...\right] \tag{10}$$

$$= \frac{1}{2}\,\mathbb{E}\left[\mathbb{E}\,\sup_{f\in F}\frac{1}{n}\sum_i^n(-\epsilon_i Y_i)f(X_i)|(X_i,Y_i)...\right] \quad \text{(can take out the 1 since } \mathbb{E}\,\epsilon_i = 0) \tag{11}$$

$$= \frac{1}{2}\,R_n(F) \tag{12}$$

Lemma 1.3. For finite $F \subseteq [-1,1]^X$,

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

Clearly,

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