CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 16**

Sauer's Lemma

Lecturer: Peter Bartlett *Scribe: Zeyu Li*

1 Recap

Consider the pattern classification setting: $F \subseteq \{\pm 1\}^{\mathcal{X}}$ and $l : x \mapsto \{0, 1\}$. For the minimizer of the empirical risk $\hat{\mathbb{E}} l_f$,

$$\hat{f} = \underset{f \in F}{\operatorname{argmin}}\ \hat{\mathbb{E}} l_f$$

with probability at least $1 - \delta$, we have:

$$\begin{aligned} \mathbb{E} l_{\hat{f}} &\leq \inf_{f \in F} \mathbb{E} l_f + 2R_n(l_F) + c\sqrt{\frac{\log \frac{1}{\delta}}{n}} \\ &= \inf_{f \in F} \mathbb{E} l_f + R_n(F) + c\sqrt{\frac{\log \frac{1}{\delta}}{n}} \end{aligned} \tag{1}$$

where l_F is defined as $\{l_f : f \in F\}$ and $R_n(F)$ is the Rademacher average:

$$R_n(F) = \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \tag{2}$$

The Rademacher averages can be bounded, for instance, as follows:

$$R_n(F) \leq \begin{cases} \sqrt{\frac{2 \log |F|}{n}}, & \text{if } |F| < \infty; \\ \sqrt{\frac{2 \log \Pi_F(n)}{n}}, & \text{if we restrict the growth function.} \end{cases} \tag{3}$$

where,

$$F_{|x_1^n} = \{(f(x_1), \ldots, f(x_n)) : f \in F\} \subseteq \{\pm 1\}^n$$

and ϵ_i are uniformly distributed random variables $\epsilon \in \{\pm 1\}$.

In this lecture, two topics are discussed:

* Sauer's Lemma;
* Rademacher averages: applications.

2 Sauer's Lemma

Definition. Growth Function:

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

Definition. VC dimension

$$d_{\mathrm{VC}}(F) = \max\{|s| : s \subseteq \mathcal{X}, f \text{ shatters } s\}$$

Here, we say that a family of binary functions F shatters a set $\mathcal{S} \in \mathcal{X}$ if $F_{|\mathcal{S}} = 2^{|\mathcal{S}|}$.

Theorem 2.1. Sauer's Lemma: If $F \subseteq \{\pm 1\}^{\mathcal{X}}$ and $d_{VC} = d$, then $\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i}$. And for $n \geq d$, $\Pi_F(n) \leq \left(\frac{en}{d}\right)^d$

That means: if $d_{VC}(F)$ is ∞, we always get exponential growth function; however, if $d_{VC}(F) = d$ is finite, the growth function increases exponentially up to d and polynomially for $n > d$.

Proof. Fix $(x_1, \ldots, x_n) \in \mathcal{X}$, and consider a table containing the values of functions in the class $F_{|x_1^n}$ restricted to the sample. For instance, consider the following example:

	x_1	x_2	x_3	x_4	x_5
f_1	-	+	-	+	+
f_2	+	-	-	+	+
f_3	+	+	+	-	+
f_4	-	+	+	-	-
f_5	-	-	-	+	-

Each row is one possible evaluation of the functions in F on the fixed sample, and the cardinality of $F_{|x_1^n}$ equals to the number of rows. We transform the table by "shifting" columns.

Definition. shifting column i: for each row, replace a "+" in column i with a "-" unless it would produce a row that is already in the table.

After applying the shifting operation in order from x_1 to x_5, we get the table($F^*_{|x_1^n}$):

	x_1	x_2	x_3	x_4	x_5
f_1	-	+	-	-	-
f_2	-	-	-	+	+
f_3	-	-	-	-	+
f_4	-	-	-	-	-
f_5	-	-	-	+	-

Observations:

(1) Size of the table unchanged, because the rows in $F^*_{|x_1^n}$ are still distinct;

(2) The table $F^*_{|x_1^n}$ exhibits "closed below" property, i.e., for each row containing a "+", replacing that "+" with a "-" produces another row in the table.

(3) $d_{VC}(F^*_{|x_1^n}) \leq d_{VC}(F_{|x_1^n})$. To see this, consider the application of the shifting operation to a single column, and notice that if F^* (after shifting) shatters a subset of columns, then so does F (before shifting).

Therefore,

(3) and (2) $\Rightarrow$ F^* can not have more than d "+"'s in a row. Hence, #row of $F^* \leq \sum_{i=0}^{d} \binom{n}{i}$;

$(1) \Rightarrow |F_{|x_1^n}| \leq \sum_{i=1}^{d} \binom{n}{i}$

Also, if $n \geq d$, we have:

$$\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$$

Because:

$$\begin{aligned}\sum_{i=0}^{d} \binom{n}{i} &\leq (\frac{n}{d})^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\ &= (\frac{n}{d})^d \left(1 + \frac{d}{n}\right)^n \\ &\leq \left(\frac{en}{d}\right)^d\end{aligned}$$

□

In summary, we have:

$$\Pi_F(n) = \begin{cases} 2^n, & n \leq d; \\ \leq \left(\frac{en}{d}\right)^d, & n > d. \end{cases} \tag{4}$$

Plug Eqn.(4) and Eqn.(3) into Eqn.(1), we have: for $d_{VC}(F) \leq d$, with probability at least $1 - \delta$,

$$\mathbb{E}l_{\hat{f}} \leq \inf_{f \in F} \mathbb{E}l_f + \sqrt{\frac{2d \cdot \log(en/d)}{n}} + c\sqrt{\frac{\log \frac{1}{\delta}}{n}}$$

Now let's look at a lower bound on the expected loss of a function class. We have the following converse of Theorem 2.1.

Theorem 2.2. Converse Theorem: For a function class F with $d_{VC}(F) \geq d$, $\delta < 1/200$ (a small constant), and any $f_n : (\mathcal{X} \times \{\pm 1\})^n \times \mathcal{X} \mapsto \{\pm 1\}$, $\exists$ probability distribution P on $\mathcal{X} \times \mathcal{Y}$, such that w.p. $\geq \delta$:

$$\mathbb{E}\left[l(Y, f_n(X_1, Y_1, \ldots, X_n, Y_n\,;\, X)) | x_1, y_1, \ldots, x_n, y_n\right] - \inf_{f \in F} \mathbb{E}l(f(X), Y) \geq c \cdot \min\left(\sqrt{\frac{d}{n}}, 1\right)$$

Proof. Proof Idea: Suppose we have a shattered set $\{x_1, \ldots, x_n\}$ by class F, and we choose our Y based on the following probability distribution P:

$$P(Y = 1 | X = x_i) = \frac{1}{2} \pm \epsilon$$

Because $\epsilon > 0$ is a very small number, it is very hard to distinguish its probability (either be 1 or 0) for each x_i. Therefore, it requires a large number of examples from each position of the set in order to get the correct estimation. This will make the learning problem harder.

Intuitively, this Theorem tells us that: there remains a probability distribution such that the expected loss drops very slowly.

3 Rademacher averages: applications

In this section, we will learn how to estimate the Rademacher averages for function classes that are built from simpler classes. The following lists some properties of Rademacher averages (recall its definition in Eqn.(2)).

(1) $F \subseteq \mathcal{G} \Rightarrow R_n(F) \leq R_n(\mathcal{G})$; [based on the definition]

(2) $R_n(c \cdot F) = |c| R_n(F)$, where $c \cdot F = \{x \mapsto c \cdot f(x) \,:\, f \in F\}$. Based on the definition:

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

where ϵ_i is uniformly distributed r.v., $\epsilon \in \pm 1$. $|c|\epsilon_i$ has the same distribution as $c\epsilon_i$, which leads to,

$$R_n(cF) = |c| \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) = |c| \cdot R_n(F)$$

(3) $R_n(F + g) = R_n(F)$, where $F + g$ is defined as $\{x \mapsto f(x) + g(x) \,:\, f \in F\}$. This is because:

$$\begin{aligned} R_n(F+g) &= \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i) + g(x_i)) \\ &= \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) + \mathbb{E} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i) \\ &= \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) = R_n(F) \end{aligned}$$

The last equality is because $g(x_i)$ is constant, therfore, its expectation is zero.

(4) For a class of functions F, let $co(F)$ represents its convex hull,

$$co(F) := \left\{ \sum_{i=1}^{k} \alpha_i f_i : k \geq 1,\, \alpha_i \geq 0,\, \|\alpha\|_1 = 1,\, f_i \in F \right\}.$$

Then we have: $R_n(F) = R_n(co(F))$. Based on the definition:

$$\begin{aligned} R_n(co(F)) &= \mathbb{E} \sup_{\substack{f_1, \ldots, f_m \in F \\ ||\alpha||_1 = 1}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \sum_{j=1}^{m} \alpha_j f_j(x_i) \\ &= \mathbb{E} \sup_{f_j \in F} \sup_{||\alpha||_1 = 1} \sum_{j=1}^{m} \alpha_j \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_j(x_i) \right) \\ &= \mathbb{E} \sup_{f_j \in F} \max_{j} \frac{1}{n} \sum \epsilon_i f_j(x_i) \\ &= R_n(F) \end{aligned}$$

where the third equality follows from the fact that the maximum of a linear function over the simplex is always achieved at one of the vertices.

(5) Ledoux-Talagrand contraction inequality: If $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfies $|\phi_i(a) - \phi_i(b)| \leq L|a - b|$), then

$$\mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_i(f(x_i)) \leq L \cdot \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)$$