CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 19**

Growth Function and Vapnik-Chervonenkis Dimension

Lecturer: Peter Bartlett *Scribe: Pierre Garrigues*

Background. Recall from the previous lecture that with probability $\geq 1-\delta$, for $\hat{f}$ the minimizer in a class F of the empirical risk $\hat{L}$, we have

$$L(\hat{f}) \leq \inf_{f\in F} L(f) + 2R_n(l_F) + c\sqrt{\frac{\log 1/\delta}{n}}$$

where the term $R_n(l_F)$ is the Rademacher average and involves the complexity of the loss class l_F.

The outline of the lecture is as follows. We'll see first how $R_n(l_F)$ and $R_n(F)$ are related, and will investigate bounds on this quantity for a finite class F. The generalization to an infinite class will lead us to introduce the growth function, and we'll derive an upper bound on this quantity involving the VC dimension of the class.

Link between $R_n(l_F)$ and $R_n(F)$. For classification ($F \subseteq \{\pm 1\}^{\mathcal{X}}$), and with the 0-1 loss, we have $R_n(l_F) = \frac{1}{2}R_n(F)$. To see that, remember that

$$l_f(x,y) = 1[y \neq f(x)] = \frac{1-yf(x)}{2}.$$

Thus the Rademacher average for l_F is

$$\begin{aligned}
R_n(l_F) &= \mathbb{E}\left(\sup_{f\in F}\frac{1}{n}\sum_i \epsilon_i \frac{1-Y_if(X_i)}{2} \mid (X_1^n, Y_1^n)\right)\\
&= \mathbb{E}\left(\frac{1}{n}\sum_i \frac{\epsilon_i}{2}\right) + \mathbb{E}\left(\sup_{f\in F}\frac{1}{n}\sum_i \frac{-(\epsilon_iY_i)f(X_i)}{2} \mid (X_1^n, Y_1^n)\right)\\
&= \frac{1}{2}\mathbb{E}\left(\sup_{f\in F}\epsilon_i' f(X_i) \mid X_1^n\right)\\
&= \frac{1}{2}R_n(F)
\end{aligned}$$

where $\epsilon_i' = \epsilon_i Y_i$ is also a Rademacher random variable.

A bound on $R_n(F)$ for a finite class F. We saw in the previous lecture that for a finite class $F \subseteq [-1,1]^{\mathcal{X}}$, then $R_n(F) \leq \sqrt{\frac{2\log|F|}{n}}$. Since l_f is such that $|l_F| \leq |F|$, we have $R_n(l_F) \leq \sqrt{\frac{2\log|F|}{n}}$. The intuition is that if the class F is big, we can always find an $f \in F$ such that it aligns itself with the ϵ_i's, and thus the Rademacher average will be big too. We now see two examples of finite classes and bound their Rademacher averages.

Example 1. F is a parameterized with a k-bit vector:

$$F = \{x \mapsto \mathrm{sign}(\theta^T x - \theta_0) : \ \theta_0, \theta_1, \dots \theta_d \in \mathcal{S}\}$$

where $|\mathcal{S}| = 2^k$. In this case, $\log |F| = c(d+1)k$, and so w.p. $\geq 1 - \delta$,

$$L(\hat{f}) \leq \inf_{f \in F} L(f) + c\sqrt{\frac{dk}{n}} + c'\sqrt{\frac{\log 1/\delta}{n}}$$

Example 2. $F =$ {decision trees with N decision nodes on $\mathcal{X} = \{\pm 1\}^d$}. In such a binary tree, a node is either an internal node, i.e. checks the value of some X_i, or a decision node giving the classification output. A node can thus take $d + 2$ values for a 2-class problem. There are N internal nodes, and $N + 1$ leaves or decision nodes. So in this case, $|F| \leq (d + 2)^{2N+1}$, and thus $R_n(F) \leq c\sqrt{\frac{N \log d}{n}}$.

Extension to F such that $|F| = \infty$. Requiring the class to be finite is very restrictive, and we would like to obtain bounds on the Rademacher averages of infinite classes. To do so, recall the lemma of Massart.

Lemma 1. For $A \subseteq \mathbb{R}^n$ with $\max_{a \in A} \frac{1}{n} \sum_i a_i^2 = R^2$, we have

$$\mathbb{E}\left(\max_{a \in A} \frac{1}{n} \sum_i \epsilon_i a_i\right) \leq R\sqrt{\frac{2 \log |A|}{n}}$$

We don't need the class to be finite, but rather the restrictions (to the sample) of the class to be finite to bound the Rademacher averages. Indeed, one can write

$$R_n(F) = \mathbb{E}\left(\max_{a \in A} \frac{1}{n} \sum_i \epsilon_i a_i \mid X_1^n\right)$$

where $A = \{(f(X_1), \dots, f(X_n))^T : f \in F\}$.

We introduce the notation $F_{|X_1^n} = \{(f(X_1), \dots, f(X_n))^T : f \in F\} \subseteq \{\pm 1\}^n$, and by extensive use of Jensen's inequality we obtain

$$\begin{aligned} R_n(F) = \mathbb{E}\left(\max_{a \in F_{|X_1^n}} \frac{1}{n} \sum_i \epsilon_i a_i \mid X_1^n\right) &\leq \sqrt{\frac{2}{n}} \mathbb{E}\sqrt{\log |F_{|X_1^n}|} \\ &\leq \sqrt{\frac{2\mathbb{E} \log |F_{|X_1^n}|}{n}} \\ &\leq \sqrt{\frac{2 \log \mathbb{E}|F_{|X_1^n}|}{n}} \end{aligned}$$

Example 3. Consider the infinite class $F = \{x \mapsto 2.1[x \leq \theta] - 1 : \theta \in \mathbb{R}\}$. We have $|F_{|X_1^n}| \leq n + 1$. To see this, consider the permutation of the x_i's such that $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. The possible evaluations of a function of F on $(x_{(1)}, x_{(2)}, \dots, x_{(n)})$ are $(-1, -1, \dots - 1), (1, -1, \dots - 1), \dots (1, 1, \dots 1)$, and there are $n + 1$ of them.

Definition. The growth function of $F \subseteq \{\pm 1\}^n$ is $\Pi_F(n) = \max\{|F_{|x_1^n}| : (x_1, \dots, x_n) \in \mathcal{X}^n\}$.

Notice that

- $\Pi_F(n) \leq |F|$
- $\Pi_F(n) \leq 2^n$
- $|F| < \infty \Rightarrow \lim_{n\to\infty} \Pi_F(n) = |F|$

In the case of linear threshold function on $\mathbb{R}^d$ ($\{x \mapsto \text{sign}(\theta^T x - \theta_0) : \theta \in \mathbb{R}^d, \theta_0 \in \mathbb{R}\}$), we will see that $\Pi_F(n) = 2\sum_{i=0}^{d}\binom{n-1}{i}$.

The VC dimension is the point at which $\Pi_F(n)$ drops below 2^n. We will see that when n is greater than the VC dimension, the growth is polynomial, which is quite a surprising and remarkable result. The bound on the Rademacher average will then be $O(\frac{\log n}{n})$, and thus goes to zero as n tends to ∞.

Definition. We say that F shatters $\mathcal{S} \subseteq \mathcal{X}$ if $|F_{|\mathcal{S}}| = 2^{|\mathcal{S}|}$. This means that we get all the dichotomies of that set. The Vapnik-Chervonenkis dimension of F is

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

For the linear threshold functions on $\mathbb{R}^d$ that we just considered, we have

$$\Pi_F(n) = 2\sum_{i=0}^{d}\binom{n-1}{i} = \begin{cases} 2^n & \text{for } d > n-1 \\ 2^n - 2\sum_{i=d+1}^{n-1}\binom{n-1}{i} & \text{otherwise} \end{cases}$$

So $\Pi_F(n) < 2^n$ iff $n > d+1$, and thus $d_{VC}(F) = d+1$.

We now prove the fact that $\Pi_F(n)$ has a polynomial growth for $n \geq d$ in the general case.

Theorem 2. (Sauer's lemma). If $F \subseteq \{\pm 1\}^{\mathcal{X}}$ and $d_{VC}(F) = d$, then $\Pi_F(n) \leq \sum_{i=0}^{d}\binom{n}{i}$, and for $n \geq d$, $\Pi_F(n) \leq \left(\frac{en}{d}\right)^d$.

Proof. We fix a set $(x_1, \ldots, x_n) \in \mathcal{X}^n$, and we consider the restriction of F to x_1^n: $F_{|x_1^n}$. We represent $F_{|x_1^n}$ via the following table

$f(x_1)$	$f(x_2)$	$\ldots$	$f(x_n)$
+	+	$\ldots$	-
-	+	$\ldots$	-
+	-	$\ldots$	-

where the rows are the possible distinct evaluations of the functions in F of the set $(x_1, \ldots, x_n)$. Thus, $|\{(f(x_1), \ldots, f(x_n)) : \ f \in F\}|$ is the number of rows in the table. The functions in F corresponding to a same evaluation or row form equivalence classes. We now consider a transformation on the table T that works as follows. For each column from the first to the last and in this order, change a "$+$" to a "$-$" in the column unless it duplicates a row already present. As an example, consider the following table:

$f(x_1)$	$f(x_2)$	$f(x_3)$
+	+	-
-	+	+
+	+	+

- transformation of the first column: $\begin{array}{ccc} - & + & - \\ - & + & + \\ + & - & + \end{array}$

- transformation of the second column: $\begin{array}{ccc} - & - & - \\ - & - & + \\ + & - & + \end{array}$

- transformation of the third column: $\begin{array}{ccc} - & - & - \\ - & - & + \\ + & - & - \end{array}$

By construction, we notice that all the rows of the transformed table T^* are distinct, and so the size of the table remains unchanged. In the sequel, we will say that T shatters a subset of columns or x_i's (we identify x_i with the ith column) if we can find all possible dichotomies of this subset in the table. Thus T shatters a subset of columns iff F shatters the x_i's corresponding to these columns.

We now prove that if T^* shatters a subset of columns, then so did T. Suppose that T^* shatters (x_i, x_j), and $i < j$. Then, we have 4 rows in T^* where the values for the ith and jth columns are $(+,+)$, $(+,-)$, $(-,+)$ and $(-,-)$. Consider the table before and after transforming the ith row. There were two rows with $(+,+)$ and $(+,-)$ in the ith and jth column before the transformation since we can only go from " $+$ " to " $-$ ". For the first " $+$ " in $(+,+)$ not to be mapped to a " $-$ " while transforming the ith column means that there was also an instance of $(-,+)$ in some other row of the table, because if there was not one would have changed the " $+$ " into a " $-$ " without duplicating an already existing row. Similarly, the first " $+$ " in $(+,-)$ implies the existence of a row with $(-,-)$. We have just proved that (x_i, x_j) was shattered before transforming the ith row. Similar arguments show that (x_i, x_j) was shattered before the transformation of the jth row. It is not difficult to extend the proof in the case where the size of the set is greater than 2. Note that the converse of this result is not true, i.e. a set can be shattered in T but not in T^*.

The table T^* exhibits the "closed-below" property, i.e. for each row, T^* shatters the set of the x_i's such that the element of the row in the ith column is a " $+$ ". To see that, note that changing any " $+$ " into a " $-$ " in T^* leads to an already existing row. We have thus all the possible dichotomies of this set in the table. Since the VC dimension of F is d, we cannot have more than d " $+$ "'s in a row, otherwise this would mean that T and hence F shatters a set of cardinality greater than d. Thus, the number of rows in T^* is no more than $\sum_{i=0}^{d} \binom{n}{i}$. Furthermore, the number of rows in T and T^* are the same. Hence

$$|\{(f(x_1), \ldots, f(x_n)) : \ f \in F\}| \leq \sum_{i=0}^{d} \binom{n}{i}$$

Taking the maximum over all possible $(x_1, \ldots, x_n) \in \mathcal{X}^n$ we finally get $\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i}$.

Now suppose $n > d$. We have

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

which completes the proof. □