CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 15**

Growth function and VC-dimension

Lecturer: Peter L. Bartlett *Scribe: Richard Bourgon*

Relating $R(\hat{f})$ and $R(f^\star)$

At the end of last class, we saw that for $F \subseteq [0,1]^{\mathcal{X}}$, with probability greater than $1-\delta$ over the training data,

$$\sup_{f \in F} \left(\mathbb{E}f - \hat{\mathbb{E}}f \right) \leq 2R_n(F) + \sqrt{\frac{\log 1/\delta}{2n}} \tag{1}$$

Here, $R_n(F)$ denotes the Rademacher average:

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

Consider this inequality, with the loss class ℓ_F in place of F. Recall that

$$\ell_F \equiv \{\ell_f : f \in F\} \equiv \{(x,y) \mapsto \ell(y, f(x)) : f \in F\}$$

where ℓ is a loss function. If $\hat{f}$ is then defined to be $\operatorname{argmin}_{f \in F} \hat{\mathbb{E}}\ell_f$, then with probability greater than $1-\delta$,

$$\mathbb{E}\ell_{\hat{f}} \leq \inf_{f \in F} \mathbb{E}\ell_f + 2R_n(\ell_F) + c\sqrt{\frac{\log 1/\delta}{n}}$$

To see this, assume for convenience that $\exists f^\star$ for which the minimal risk is achieved. Then

$$\mathbb{E}\ell_{\hat{f}} - \mathbb{E}\ell_{f^\star} = (\mathbb{E}\ell_{\hat{f}} - \hat{\mathbb{E}}\ell_{\hat{f}}) + (\hat{\mathbb{E}}\ell_{\hat{f}} - \hat{\mathbb{E}}\ell_{f^\star}) + (\hat{\mathbb{E}}\ell_{f^\star} - \mathbb{E}\ell_{f^\star})$$

The first term can be controlled by (1) above, the second term is non-positive by the definition of $\hat{f}$, and the third term — because it deals with a fixed function $f^\star$ — can be controlled by the Hoeffding inequality.

Rademacher averages for the loss class

Consider the two-class classification problem i.e. $F \subseteq \{\pm 1\}^{\mathcal{X}}$, and 0/1 loss. Then

$$\begin{aligned}\ell_f(x,y) &= 1[y \neq f(x)] \\ &= \frac{1-yf(x)}{2}\end{aligned}$$

Letting D_n represent the training data $(X_1,Y_1),\ldots,(X_n,Y_n)$, we may relate the Rademacher averages for the loss class ℓ_F and for F itself as follows:

$$\begin{aligned}R_n(\ell_F) &= \mathbb{E}\,\mathbb{E}\left(\sup_{f\in F}\frac{1}{n}\sum_i \epsilon_i \frac{1-Y_i f(X_i)}{2}\;\Big|\; D_n\right) \\ &= \mathbb{E}\,\mathbb{E}\left(\frac{1}{n}\sum_i \frac{\epsilon_i}{2} + \sup_{f\in F}\frac{1}{n}\sum_i \frac{(-\epsilon_i Y_i)f(X_i)}{2}\;\Big|\; D_n\right) \\ &= 0 + \mathbb{E}\,\mathbb{E}\left(\sup_{f\in F}\frac{1}{n}\sum_i \frac{(-\epsilon_i Y_i)f(X_i)}{2}\;\Big|\; D_n\right) \\ &= \mathbb{E}\,\mathbb{E}\left(\sup_{f\in F}\frac{1}{n}\sum_i \frac{\epsilon_i f(X_i)}{2}\;\Big|\; D_n\right)\end{aligned}$$

because $Y_i \in \{\pm 1\}$ and $\epsilon_i \stackrel{d}{=} -\epsilon_i$. But this last expression is just $R_n(F)/2$.

Some examples: finite F

We saw previously that for $A \subset \mathbb{R}^n$ and $R \equiv \max_{a\in A}\|a\|_2$,

$$\mathbb{E}\sup_{a\in A}\sum_{i=1}^n \epsilon_i a_i \leq R\sqrt{2\log|A|} \qquad (2)$$

We can apply this result to any class F which can be parameterized with a finite number of bits. Consider, for example, the linear threshold functions with parameter values in some finite set S where $|S| = 2^k$:

$$F = \{x \mapsto \mathrm{sign}(w^T x - w_0) : (w_0, w_1, \ldots, w_n)^T \in S\}$$

Here, $R = 1/\sqrt{n}$, so $R_n(F) \leq \sqrt{2k/n}$. Another example in this vein is to let F be the set of decision trees with N decision nodes for $\mathcal{X} = \{\pm 1\}^d$, the d-cube. Any such decision tree can be described by enumerating the nodes of the tree in breadth-first order, and for each node listing a value from the set $\{1,\ldots,d,+,-\}$. These values have the following meaning:

(i) A value $i \in \{1,\ldots,d\}$ indicates that the node is an internal node, and i represents the component of X checked at that node. (Without loss of generality, we can always assume that the left branch corresponds to $X_i = -1$, say.)

(ii) A value in $\{+,-\}$ indicates that the node is a leaf with the corresponding decision output.

It is easy to see that any decision tree with N internal nodes has $2N+1$ nodes (internal nodes and leaves), and so can be uniquely specified by such a list of length $2N+1$, and each element of the list comes from a set of size $d+2$. Thus, $|F| \le (d+2)^{2N+1}$.

Combining equations (1) and (2), we see that with probability greater than $1-\delta$,

$$R(\hat{f}) \le \inf_{f \in F} R(f) + \sqrt{\frac{2(2N+1)\log(d+2)}{n}} + \sqrt{\frac{\log 1/\delta}{2n}}$$

Extending to infinite F

The result in equation (2) requires that $|A|$ be finite, but is still applicable to infinite F, provided that the restriction of F to the training data is finite. As before, let D_n represent the training data, and let

$$F|_{D_n} \equiv \{\big(f(X_1), \ldots, f(X_n)\big) : f \in F\}$$

When $F \subseteq \{\pm 1\}^{\mathcal{X}}$, $F|_{D_n} \subseteq \{\pm 1\}^n$, and thus $\big|F|_{D_n}\big| < \infty$. As a consequence,

$$\begin{aligned}
R_n(F) &\equiv \mathbb{E}\left(\sup_{f \in F} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right) \\
&= \mathbb{E}\,\mathbb{E}\left(\max_{a \in F|_{D_n}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i a_i \,\Big|\, D_n\right) \\
&\le \mathbb{E}\sqrt{\frac{2\log\big|F|_{D_n}\big|}{n}} \\
&\le \sqrt{\frac{2\log \mathbb{E}\big|F|_{D_n}\big|}{n}}
\end{aligned}$$

where the last line follows from two applications of Jensen's inequality. As an example of how $\big|F|_{D_n}\big|$ may be quite small even when F is uncountably infinite, recall our proof last class of the Glivenko-Cantelli Theorem. In this case, we defined

$$F \equiv \{x \mapsto 1[x \le \theta] : \theta \in \mathbb{R}\}$$

as saw that $\big|F|_{D_n}\big| \le n+1$.

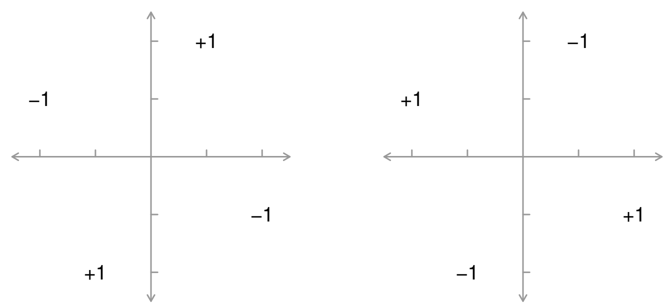
The growth function

Define the *growth function* as follows:

$$\Pi_F(n) \equiv \max\{\big|F|_{D_n}\big| : D_n \in \mathcal{X}^n\}$$

Clearly, $\Pi_F(n) \le 2^n$, and $\Pi_F(n) \le |F|$ as well. Also, for finite F, $\lim_{n\to\infty} \Pi_F(n) = |F|$.

As an example, again take F to be the set of linear threshold functions on $\mathbb{R}^2$. When $n = 2$ all four maps (+1/+1, +1/-1, -1/+1, and -1/-1) are clearly in F and thus $\Pi_F(2) = 4$. For $n = 3$, all 2^3 maps can still be obtained — provided that X_1, X_2, and X_3 are not colinear — and so $\Pi_F(3) = 8$. For $n = 4$ however, there are two maps which cannot be obtained even when the training data are optimally configured:



Since all other maps for this configuration of the training data *can* be obtained, and since no other configuration improves matters, $\Pi_F(4) = 2^4 - 2 = 14$.

The *Vapnik-Chervonenkis dimension* of the class F is related to the growth function: it is defined to be the largest value of n for which, under at least one configuration of the training data, all 2^n maps can still be obtained. More formally, F *shatters* a subset S of $\mathcal{X}$ if $\left|F|_S\right| = 2^{|S|}$, and then

$$d_{VC}(F) \equiv \max\{|S| : S \subseteq \mathcal{X} \text{ and } F \text{ shatters } S\}$$

An equivalent definition is

$$d_{VC}(F) \equiv \max\{n : \Pi_F(n) = 2^n\}$$

In the present example, $d_{VC}(F) = 3$.

Above, we examined the cases $n = 2$, 3, and 4 for $\mathcal{X} = \mathbb{R}^2$. The general result on $\Pi_F(n)$ for linear threshold functions (which we will prove later) is

Theorem 1.1. When F is the set of linear threshold functions on $\mathbb{R}^d$,

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

A corollary, which we will prove directly without use of the theorem, is

Corollary 1.2. When F is the set of linear threshold functions on $\mathbb{R}^d$, $d_{VC}(F) = d + 1$.

Proof.

Observe that $X_1, \dots, X_n$ is shattered by F iff the vectors which make up the columns of the $(d+1) \times n$ matrix

$$A \equiv \begin{pmatrix} X_1 & \cdots & X_n \\ 1 & \cdots & 1 \end{pmatrix}$$

are linearly independent in $\mathbb{R}^{d+1}$:

- Suppose the columns of A are linearly independent. Then for any $\sigma \in \{\pm 1\}^{d+1}$, the system of n equations defined by $(w^T w_0)A = \sigma^T$ has a solution. Thus, there is a linear threshold function in F which can reproduce the assignment of an arbitrary σ_i to each Y_i.
- Conversely, suppose the columns of A are not linearly independent. For F to shatter $X_1, \dots, X_n$ we must be able to chose w and w_0 such that $\left((w^T w_0)A\right)^T$ lies in each of the 2^n orthants of $\mathbb{R}^n$. If the rank of A is less than n, however, this is not possible; if it were, it would imply that A^T maps $\mathbb{R}^{d+1}$ onto $\mathbb{R}^n$, contradicting the assumption of linear dependence of the columns of A.

Now, clearly, for $n > d+1$, no choice of the X_i makes the columns of A linearly independent. For $n = d+1$, however, any choice of the X_i which make A invertible satisfy the criterion. Thus $d_{VC}(F) = d+1$.

□

Note that the linear independence argument used in the corollary can be applied to any linearly parameterized class of decision functions e.g.

$$F \equiv \{x \mapsto \operatorname{sign}\langle \Phi(x), w \rangle : w \in \mathbb{R}^n\}$$

Sauer's Lemma

One consequence of Theorem 1.1 is that for $n \leq d$, $\Pi_F(n)$ increases exponentially with n; for $n > d$, however, $\Pi_F(n)$ increases polynomially with n. This is true more generally:

Theorem 1.3. (Sauer) If $F \subseteq \{\pm 1\}^{\mathcal{X}}$ and $d_{VC}(F) \leq d$, then $\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i}$. When $n \leq d$, this is just 2^n. If $n > d$, on the other hand, $\Pi_F(n) < (en/d)^d$ i.e. $\Pi_F(n)$ grows polynomially in n.

Proof.

Fix some $(X_1, \dots, X_n) \in \mathcal{X}^n$ and represent $F|_{D_n}$ as a table (of at most 2^n rows). For example, if $n = 5$, the table might be

$f(X_1)$	$f(X_2)$	$f(X_3)$	$f(X_4)$	$f(X_5)$
-	+	-	+	+
+	-	-	+	+
+	+	+	-	+
-	+	+	-	-
-	-	-	+	-

Now produce a new table for $F^\star$ by proceeding down each column, taking the columns in numerical order, and switching "+" to "-," *provided that the switch does not duplicate a row already present in the table.* After transforming just columns 1 and 2 in the table above, for instance, we would obtain

$f^\star(X_1)$	$f(X_2)$	$f(X_3)$	$f(X_4)$	$f(X_5)$
-	+	-	+	+
-	-	-	+	+
-	+	+	-	+
-	+	+	-	-
-	-	-	+	-

$f^\star(X_1)$	$f^\star(X_2)$	$f(X_3)$	$f(X_4)$	$f(X_5)$
-	+	-	+	+
-	-	-	+	+
-	-	+	-	+
-	-	+	-	-
-	-	-	+	-

Note that the first "+" in column 2 could not be switched because it would have resulted in duplication of row 2. If we continue through all 5 columns in our example, we obtain

$f^\star(X_1)$	$f^\star(X_2)$	$f^\star(X_3)$	$f^\star(X_4)$	$f^\star(X_5)$
-	+	-	-	-
-	-	-	+	+
-	-	-	-	+
-	-	-	-	-
-	-	-	+	-

Observe that

(i) The table for $F^\star$ has the same number of distinct rows as the table for F.

(ii) If we change any "+" to a "-" anywhere in the $F^\star$ table, we create a duplicate row in that table.

(iii) Whenever $F^\star$ shatters a subset of $\{X_1, \ldots, X_n\}$, then the same must be true for F. For instance, $\{X_4, X_5\}$ is shattered by both the original and transformed tables in our example. To see why this is true, compare the table before and after transforming a single column, say column 1, and suppose that there is a shattered set in the transformed table (after shifting column 1). If the shattered set does not include column 1, then it is certainly a shattered set in the original table (before shifting that column). If the shattered set includes the column, then fix a sign assignment to the other columns in the shattered set, and notice that there are at least two rows with that sign assignment: one with a '+' in column 1 and one with a '-'. The '+' row must have been present in the original table, and for the '+' row to have remained, there must have been another row in the original table that is identical except that it had a '-' in column 1. Thus, both rows were present in the original table. It follows that the set was shattered in the original table.

The subset $\{X_1, X_2\}$, however, demonstrates that the implication does not go in the other direction.

We can easily find subsets shattered by $F^\star$, by taking any row of the $F^\star$ table and selecting the X_i which have a "+" in that row. (To see that this is true, apply observation (ii) repeatedly.) This fact, combined with observation (iii) then implies that no single row of the $F^\star$ table has more than d "+" signs, and thus that the $F^\star$ table has at most $\sum_{i=0}^{d} \binom{n}{i}$ rows. Combining this fact with (i) gives

$$\left|F|_{D_n}\right| \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Now suppose that $n > d$. Then

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

$\square$