CSE522, Winter 2011, Learning Theory Lecture 10 - 02/01/2011

Proof of Sauer's Lemma

Lecturer: Ofer Dekel *Scribe: Andrew Guillory*

1 Review

In previous lectures we showed that, in the case of binary classification, the Radamacher complexity of a hypothesis class is closely related to the VC dimension of the class. When discussing VC dimension, we focus on binary classification. In this setting, $\mathcal{Y} = \{+1, -1\}$, and $\forall h \in \mathcal{H}\ h : \mathcal{X} \to \{+1, -1\}$. Define 0-1 loss (also called error indicator loss)

$$l(h; (x, y)) = \begin{cases} 1 & \text{if } h(x) \neq y \\ 0 & \text{if } h(x) = y \end{cases}$$

In the case of binary classification, $h(x) \neq y \Leftrightarrow yh(x) = -1$ and $h(x) = y \Leftrightarrow yh(x) = +1$. Then, we can use the equivalent definition

$$l(h; (x, y)) = \frac{1}{2} - \frac{1}{2} yh(x)$$

This loss function is $\frac{1}{2}$-Lipschitz in $h(x)$. We can then use our result for the Radamacher complexity of $l \circ \mathcal{H}$ for Lipschitz loss functions to get

$$R_m(l \circ \mathcal{H}) \leq \frac{1}{2} R_m(\mathcal{H})$$

and similarly for empirical Radamacher complexity

$$\hat{R}_m(l \circ \mathcal{H}, S) \leq \frac{1}{2} \hat{R}_m(\mathcal{H}, S)$$

Observe that for binary classification,

$$\begin{aligned} \hat{R}_m(\mathcal{H}, S) &= \frac{2}{m} E_\sigma \max_{h \in \mathcal{H}} \sum_{i=1}^{n} \sigma_i h(x_i) \\ &= \frac{2}{m} E_\sigma \max_{a \in A_S} \sum_{i=1}^{n} \sigma_i a_i \end{aligned}$$

where $A_s = \{(h(x_1), \ldots h(x_m)) : h \in \mathcal{H}\}$. $A_S \subseteq \{+1, -1\}^m$ is a finite set ($|A_S| \leq 2^m$). Also $\forall a \in A_s$ we have $||a|| = \sqrt{m}$. Since A_s is finite and a has bounded length, we can apply Massart's finite class lemma.

Theorem 1 (Massart's Finite Class Lemma)**.** $\forall A \subseteq \mathbb{R}^m$ *with* $|A| < \infty$ *and* $\max_{a \in A} ||a|| \leq \rho$

$$E_\sigma \max_{a \in A} \sum_{i=1}^{m} \sigma_i a_i \leq \rho \sqrt{2 \log |A|}$$

In our case

$$\hat{R}_m(\mathcal{H}, S) \leq \frac{2}{\sqrt{m}} \sqrt{2 \log |A|}$$

Using this bound on empirical Radamacher complexity, we get a bound on Radamacher complexity

$$R_m(\mathcal{H}) \leq \frac{2}{\sqrt{m}} \sqrt{2 \log g_{\mathcal{H}}(m)}$$

where $g_{\mathcal{H}}(m)$ is the *growth function* of H defined

$$g_{\mathcal{H}}(m) = \max_{S \in \mathcal{X}^m} |A_S|$$

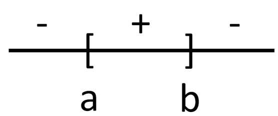
Note that $g_{\mathcal{H}}(m) \leq 2^m$.

We say that $\mathcal{H}$ *shatters* S if $\mathcal{H}$ can label S in $2^{|S|}$ different ways. The VC dimension of a hypothesis class $\mathcal{H}$ is the size of the largest set $\mathcal{H}$ can shatter.

$$\begin{aligned} VCdim(\mathcal{H}) &= \max\{|S| : \mathcal{H} \text{ shatters } S\} \\ &= \max\{m : g_{\mathcal{H}}(m) = 2^M\} \end{aligned}$$

2 Examples

In some very simple cases we are able to calculate the growth function explicitly. Consider for example the case where $\mathcal{H}$ is the class of intervals on the real line given by $h = [a, b]$ with $a \in \mathbb{R}$, $b \in \mathbb{R}$, $a \leq b$



$$h(x) = \begin{cases} +1 & \text{if } a \leq x \leq b \\ -1 & \text{otherwise} \end{cases}$$

Theorem 2. *Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H}$ be intervals on the real line.*

$$g_{\mathcal{H}}(m) = 1 + m + \binom{m}{2}$$

Proof. First count the number of labelings with at least one positive point followed by at least one negative point. There are $\binom{m}{2}$ such labelings since we must choose the first positive point and the first negative point after the positive points.

Second count the number of labelings with at least one positive point but no negative labeled points following the positive points. There are m such labelings since we must choose the first positive point.

Third count the number of labelings with no positive points. There is only 1 such labeling.

Summing these three cases gives that there are $1 + m + \binom{m}{2}$ different labelings. $\square$

We now show that for intervals on the real line $VCdim(\mathcal{H}) = 2$. Note this result follows from the growth function, but we show it more directly here to illustrate the common structure of this kind of proof.

Theorem 3. *Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{H}$ be intervals on the real line.*

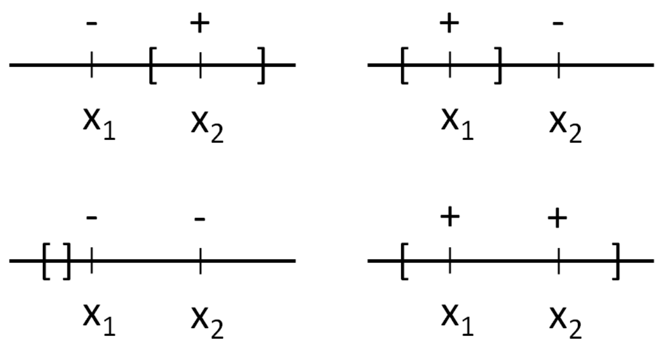
$$VCdim(\mathcal{H}) = 2$$

Proof. As is standard for proofs showing the VC dimension of a class, there are two steps

1. Lower bound

$$VCdim(\mathcal{H}) \geq 2$$

To show the lower bound, we simply need to give an example set S of 2 points such that $\mathcal{H}$ shatters S. Example:



2. Upper bound

$$VCdim(\mathcal{H}) < 3$$

   To show the upper bound, we must argue that $\forall S \in \mathcal{X}^3$, $|A_S| < 2^3$. Equivalently, we must show $\forall S \in \mathcal{X}^3\ \exists y \in \{+1,-1\}^3$ such that $\mathcal{H}$ cannot attain the labeling y.

   Let x_1, x_2, x_3 be 3 arbitrary points. Without loss of generality, assume $x_1 \le x_2 \le x_3$. Let $y = (+1,-1,+1)$. Assume some $h \in \mathcal{H}$ with $h = [a,b]$ attains these labels. $y_1 = +1$ so $a \le x_1$. $y_3 = +1$ so $x_3 \le b$. Then

$$a \le x_1 \le x_2 \le x_3 \le b$$

   $y_2 = -1$ so this is a contradiction.

□

3 Sauer's Lemma

Sauer's Lemma relates the growth function of a class to its VC dimension.

Theorem 4 (Sauer's Lemma)**.** *If $d = VCdim(\mathcal{H})$ then*

$$g_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} = \Phi_d(m)$$

As we just showed, this result is tight for intervals on the real line. A standard result shows $\Phi_d(m)$ is bounded above by a polynomial of degree d.

Lemma 5.

$$\Phi_d(m) \le \left(\frac{em}{d}\right)^d = O(m^d)$$

Sauer's Lemma then tells us that the growth function $g_{\mathcal{H}}$ grows exponentially fast for $m \le d$ but then only grows like a polynomial for $m \ge d$. We can then conclude that, so long as the VC dimension of our class is finite (i.e. $VCdim(\mathcal{H}) < \infty$), as $m \to \infty$ we have $R_m(l \circ \mathcal{H}) = \tilde{O}(\frac{1}{\sqrt{m}}) \to 0$. Here $\tilde{O}$ means ignoring log terms. Essentially, finite VC dimension implies "learnability".

Before proving Sauer's Lemma, we first note that we define $\binom{m}{k}$ as follows

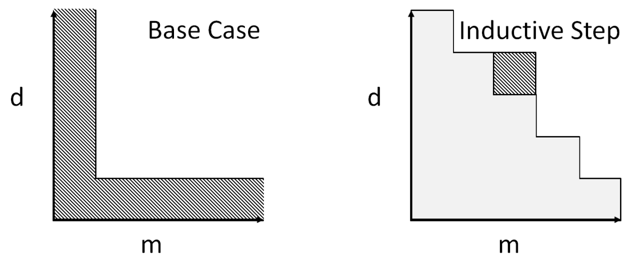
$$\binom{m}{k} = \begin{cases} \frac{m!}{(m-k)!k!} & \text{if } 0 \le k \le m \\ 0 & \text{otherwise} \end{cases}$$

Defining it in this way (i.e. defining it to be 0 for $k > m$ and $k < 0$) makes the proof cleaner. We also state the following basic fact from combinatorics

$$\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1} \tag{1}$$

To see that this holds note that to choose k items from m we can either include or excluse the first item. If we include it, we must choose $k-1$ items from the remaining $m-1$. If we exclude it, we must choose k items from the remaining $m-1$.

Proof of Sauer's Lemma. The proof uses complete induction on $m+d$. In the base case we show that the lemma holds for any d and $m = 0$ and for any m and $d = 0$. In the inductive step we show the lemma holds for any m, d with $m + d = k$ for some constant k assuming that it holds for all m, d with $m + d < k$.



- **Base Case** For any d and $m = 0$

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{0}{i} = 1$$

  and $g_{\mathcal{H}}(0) \leq 1$ since we can label 0 points at most 1 way.

  For any m and $d = 0$

$$\Phi_d(m) = \sum_{i=0}^{0} \binom{m}{i} = 1$$

  and $g_{\mathcal{H}}(m) = 1$ since $VCdim(\mathcal{H}) = 1$ implies we label everything with the same label.

- **Inductive Step** Choose $S \in \mathcal{X}^m$ such that we can choose $g_{\mathcal{H}}(m)$ different hypotheses from $\mathcal{H}$ that label S in $g_{\mathcal{H}}(m)$ different ways. Call this set of hypotheses A. Define $S' = S \setminus \{x_1\}$ ($|S'| = m-1$). Define $A' \subseteq A$ to be the smallest subset of A that labels S' in the maximal number of different ways. If 2 hypotheses in A label S' the same way then one is in A' and one is in $A \setminus A'$. $A = (A \setminus A') \cup A'$ and this partition of A into A' and $A \setminus A'$ forms the basis of our inductive argument.

  If A' shatters $T \subseteq S'$ then A also shatters T. therefore $|T| \leq d$ and $VCdim(A') \leq d$.

  If $A \backslash A'$ shatters $\tilde{T} \subseteq S'$ then A shatters $\tilde{T} \cup \{x_1\}$ This is by construction of A' since for every $h \in (A \backslash A')$ there is a corresponding $h \in A'$ that disagrees on x_1. Therefore $|\tilde{T}| \leq d-1$ and $VCdim(A \backslash A') \leq d-1$.

We can now use the inductive hypothesis on both A' and $A \setminus A'$

$$
\begin{aligned}
g_{\mathcal{H}}(m) = |A| &= |A'| + |A \setminus A'| \\
&\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\
&= \sum_{i=0}^{d} \binom{m}{i}
\end{aligned}
$$

The last step used Equation 1.

□

We can finally conclude that for binary classification with 0-1 loss

$$
\begin{aligned}
R_m(l \circ \mathcal{H}) &\leq \frac{1}{2} R_m(H) \\
&\leq \frac{1}{\sqrt{m}} \sqrt{2 \log g_{\mathcal{H}}(m)} \\
&\leq \frac{1}{\sqrt{m}} \sqrt{2d \log \frac{em}{d}} \\
&= \tilde{O}\left(\frac{\sqrt{d}}{\sqrt{m}}\right)
\end{aligned}
$$

This follows from (in order of application) (1) our bound for the Radamacher complexity for Lipschitz loss functions, (2) Massart's finite class lemma, and (3) Sauer's lemma.