CS861: Theoretical Foundations of Machine Learning Lecture 6 - 09/18/2023
University of Wisconsin–Madison, Fall 2023

Lecture 06: PAC bound in a finite VC class, Proof of Sauer's lemma

Lecturer: Kirthevasan Kandasamy *Scribed by: Justin Kiefel, Joseph Salzer*

In this lecture, we demonstrate how to derive a bound for the growth function of a hypothesis class via its VC-dimension. This bound is called Sauer's Lemma, which will be proved in the second section. Once we have this bound we can show that, in the case of a finite VC-dimension hypothesis class, that class is (agnostic) PAC-learnable.

1 PAC Bound in a Finite VC Class

Recall the definition of a restriction $\mathcal{L}(S,\mathcal{H})$ and the growth function $g(n,\mathcal{H})$ (see Lecture 4, definition 2 and 3 respectively) of a hypothesis class $\mathcal{H}$. In our previous lecture, we proved the following generalization bound for the estimation error using the growth function: with probability greater than $1-2e^{-2n\epsilon^2}$ we have

$$R(\hat{h}) \leq \inf_{h\in\mathcal{H}} R(h) + c_1\sqrt{\frac{2\log(g(n,\mathcal{H}))}{n}} + 2\epsilon \tag{1}$$

Furthermore, we had introduced the concept of shattering and of VC dimension, i.e, the maximal size of a set that can be shattered by $\mathcal{H}$. The following lemma provides an upper bound for the growth function based on the VC-dimension.

Lemma 1 (Sauer's Lemma)**.** *Define $\Phi_d(n) := \sum_{i=0}^{d}\binom{n}{i}$. If the VC-dimension of a hypothesis class $\mathcal{H}$ is d, then*

$$g(n,\mathcal{H}) \leq \Phi_d(n)$$

We will prove this lemma in the next section of this lecture. For now, we will demonstrate a few properties of the function $\Phi_d(n)$ and use them to derive the PAC bound similar to Equation 1 but in terms of the VC-dimension instead of the growth function.

If $n \leq d$, then $\Phi_d(n) = 2^n$. But if $n > d$,

$$\begin{aligned}
\Phi_d(n) &= (\tfrac{n}{d})^d \textstyle\sum_{i=0}^{d}\binom{n}{i}(\tfrac{d}{n})^d && \text{\textit{$i \leq d$ and $n > d$}}\\
&\leq (\tfrac{n}{d})^d \textstyle\sum_{i=0}^{n}\binom{n}{i}(\tfrac{d}{n})^i && \text{\textit{binomial expansion of} } (1+\tfrac{d}{n})^n\\
&= (\tfrac{n}{d})^d(1+\tfrac{d}{n})^n && (1+\tfrac{x}{n})^n \leq e^x\\
&\leq (\tfrac{en}{d})^d
\end{aligned}$$

Thus, when $n > d$, the growth function grows polynomially in n. We can combine this result with Equation 1 to obtain the following theorem:

Theorem 1 (PAC Bound for Finite VC-dim)**.** *Let $\mathcal{H}$ be a hypothesis class with finite VC dimension d. Let $\hat{h}$ be obtained via ERM using n i.i.d samples where $n \geq d$. Further, let $\epsilon > 0$. Then with probability of at least $1-2e^{-2n\epsilon^2}$,*

$$R(\hat{h}) \leq \inf_{h \in \mathcal{H}} R(h) + O\left(\sqrt{\frac{\log(n/d)}{(n/d)}}\right) + 2\epsilon$$

2 Proof of Sauer's Lemma

We will now provide a proof for Sauer's lemma via a modified induction argument.

For $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ and $S^X = \{x_1, \cdots, x_n\} \in \mathcal{X}^n$, define $\mathcal{H}(S^X) := \{[h(x_1), \cdots, h(x_n)] : h \in \mathcal{H}\}$. The following claim will be useful in constructing our proof:

Claim 1. $g(n, \mathcal{H}) = \max_{|S^X|=n} |\mathcal{H}(S^X)|$

Proof Recall that $\mathcal{L}(S, \mathcal{H}) = \{\ [(\ell(h(x_1), y_1), \cdots, (\ell(h(x_n), y_n)] : h \in \mathcal{H}\ \}$. There exists a bijection between $\mathcal{L}(S, \mathcal{H})$ and $\mathcal{H}(S^X)$ so that $|\mathcal{L}(S, \mathcal{H})| = |\mathcal{H}(S^X)|$. Thus,

$$g(n, \mathcal{H}) = \max_{|S|=n} |\mathcal{L}(S, \mathcal{H})| = \max_{|S|=n} |\mathcal{H}(S^X)| = \max_{|S^X|=n} |\mathcal{H}(S^X)|$$

□

The following example illustrates the bijection.

Example 2. Let $S = \{(x_1 = -1, y_1 = 0), (x_2 = 1, y_2 = 1)\}$ and $\mathcal{H}_{\text{one-sided}} = \{h_a(x) = \mathbb{1}_{\{x \geq a\}} | \forall a \in \mathbb{R}\}$. Under the zero-one loss we have

$$\mathcal{L}(S, \mathcal{H}_{\text{one-sided}}) = \{[0, 1], [0, 0], [1, 0]\}$$

and

$$\mathcal{H}_{\text{one-sided}}(S^X) = \{[0, 0], [0, 1], [1, 1]\}$$

Clearly, there is a one-to-one correspondence/bijection between these two sets.

The setup for our proof of Sauer's lemma will be via induction on $k = n + d$; where n is the number of i.i.d samples and d is the VC-dimension of our hypothesis class.

1. Base case: Show that Sauer's lemma holds...
   (a) $\forall d$ and $n = 0$
   (b) $\forall n$ and $d = 0$
2. Inductive case: Let k be some constant. Assume Sauer's lemma holds $\forall n, d$ such that $n + d < k$. Show that Sauer's lemma holds $\forall n, d$ such that $n + d = k$. See Figure 1 for a visual demo of the induction strategy.

We will begin by proving the two base cases. For the first case, let $n = 0$. The VC dimension may be any non-negative integer.

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

Notice that S^X must be empty when $|S^X| = 0$. There is only one possible labeling of zero data points. Therefore, $\mathcal{H}(S^X) = \{[]\}$, and $|\mathcal{H}(S^X)| = 1$. Applying Claim 1 we see that $g(n, \mathcal{H}) = 1$. Thus, $g(n, \mathcal{H}) =$

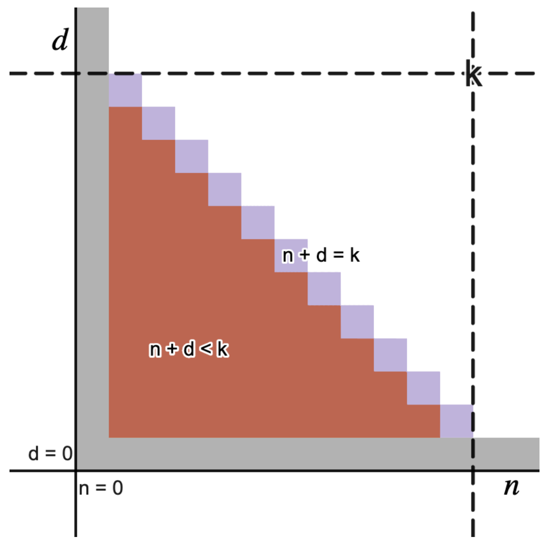

Figure 1: Visual demo of the proof by induction. The axes are n and d. The gray region represents the base case for n and d (where $n = 0$ and $d = 0$). The brownish region represents the induction hypothesis (where $n + d < k$). The purple region represents the inductive step (where $n + d = k$).

$\Phi_d(n)$, and Sauer's lemma is satisfied for $n = 0$. Now we will consider the case where $d = 0$ and n is any non-negative integer.

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

The VC dimension of $\mathcal{H}$ is 0, so the hypothesis class cannot shatter a set of size 1. Therefore, for any $x \in \mathcal{X}$, all classifiers in $\mathcal{H}$ must generate the same label. It follows that for any $S^X = \{x_1, ..., x_n\} \in X^n$, $|\{[h(x_1), ..., h(x_n)] : h \in \mathcal{H}\}| = |\mathcal{H}(S^X)| = 1$. Hence, $g(n, \mathcal{H}) = \Phi_d(n)$, and the lemma is satisfied.

We will now prove the inductive case. Assume that Sauer's lemma holds $\forall d, n$ where $d + n \leq k - 1$. Let d, n be such that $d + n = k$.

Let $S^X = \{x_1, \ldots, x_n\}$ be given. To begin with, we will construct a new hypothesis class, $\mathcal{G}$, defined only on $\{x_1, \ldots, x_n\}$ as follows. For each $[y_1, ..., y_n] \in \mathcal{H}(S^X), \exists h \in \mathcal{H}$ such that $[y_1, ..., y_n] = [h(x_1), ..., h(x_n)]$. Add one such h, restricted only to points in S^X, to $\mathcal{G}$; that is, we will add $g_h : S \to \{0, 1\}$, where $g_h(x) = h(x)$ for all $x \in S^X$, but undefined elsewhere. Therefore, $\mathcal{G}$ will have exactly one function that generates each labeling in $\mathcal{H}(S^X)$. It follows that $|\mathcal{G}(S^X)| = |\mathcal{H}(S^X)| = |\mathcal{G}|$. Next, we will partition $\mathcal{G}$ into the sets $\mathcal{G}_1$ and $\mathcal{G}_2$ using the following construction:

1. $\underline{\mathcal{G}_1}$: For every possible labeling of $\{x_1, ..., x_{n-1}\}$, add one element from $\mathcal{G}$ to $\mathcal{G}_1$.
2. $\underline{\mathcal{G}_2}$: Let $\mathcal{G}_2 = \mathcal{G} \setminus \mathcal{G}_1$.

The intuition behind this partition is to generate two hypothesis classes with VC dimension less than d. We will then apply the inductive hypothesis to each hypothesis class and bound the growth function. To demonstrate how $\mathcal{G}$ is constructed and partitioned, we present an example with a simple hypothesis class.

Example 3. Let $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$, and let $\mathcal{H}_{\text{one-sided}} = \{h_a(x) = \mathbb{1}_{\{x \geq a\}} | \forall a \in \mathbb{R}\}$. We can see that $\mathcal{H}_{one-sided}(S^X) = \{[0, 0], [0, 1], [1, 1]\}$. Due to the one-sided nature of $\mathcal{H}_{one-sided}$, it is not possible to generate the labeling $[1, 0]$. Let g_1, g_2, and g_3 be classifiers generating the predictions $[0, 0], [0, 1]$, and $[1, 1]$ respectively. Define $\mathcal{G} = \{g_1, g_2, g_3\}$. An example of $\mathcal{G}$ includes the following classifiers:

1. $g_1 = g_{h_{x_2+1}}$ which is the function h_{x_2+1} restricted to $\{x_1, x_2\}$. This function generates the label 0 for both x_1 and x_2. The function is undefined for other values.

2. $g_2 = g_{h_{(x_1+x_2)/2}}$ which is the function $h_{(x_1+x_2)/2}$ restricted to $\{x_1, x_2\}$. This function generates the labels 0 and 1 for x_1 and x_2 respectively. The function is undefined for other values.

3. $g_3 = g_{h_{x_1-1}}$ which is the function h_{x_1-1} restricted to $\{x_1, x_2\}$. This function generates the label 1 for both x_1 and x_2. The function is undefined for other values.

To construct $\mathcal{G}_1$ we will select one classifier for each labeling of $\{x_1\}$. The remaining classifiers will define $\mathcal{G}_2$. One possible partition is $\mathcal{G}_1 = \{g_1, g_3\}$ and $\mathcal{G}_2 = \{g_2\}$.

Claim 2. $|\mathcal{G}_1(S^X)| = |\mathcal{G}_1(\{x_1, ..., x_{n-1}\})|$

Proof For every labeling $\{g(x_1), ..., g(x_{n-1})\} \in \mathcal{G}_1(\{x_1, ..., x_{n-1}\})$, we have exactly one of $[g(x_1), ..., g(x_{n-1}), 0]$ or $[g(x_1), ..., g(x_{n-1}), 1]$ in $\mathcal{G}_1(S^X)$. □

Claim 3. $|\mathcal{G}_2(S^X)| = |\mathcal{G}_2(\{x_1, ..., x_{n-1}\})|$

Proof For every labeling $\{g(x_1), ..., g(x_{n-1})\} \in \mathcal{G}_1(\{x_1, ..., x_{n-1}\})$, we have exactly one of $[g(x_1), ..., g(x_{n-1}), 0]$ or $[g(x_1), ..., g(x_{n-1}), 1]$ in $\mathcal{G}_1(S^X)$. Therefore, $\mathcal{G}_2$ will have at most one of these labelings. □

We can apply the equality in Claim 2 to create a bound on $|\mathcal{G}_1(S^X)|$.

$$\begin{aligned}
|\mathcal{G}_1(S^X)| &= |\mathcal{G}_1(\{x_1, ..., x_{n-1}\})| && \text{Definition of growth function} \\
&\leq g(n-1, \mathcal{G}_1) && \text{Inductive Hypothesis} \\
&\leq \Phi_{d_{\mathcal{G}_1}}(n-1) && \Phi_d \text{ increases with } d \\
&\leq \Phi_d(n-1)
\end{aligned}$$

To show why the inductive hypothesis applies to $g(n-1, \mathcal{G}_1)$ and why $d_{\mathcal{G}_1} \leq d$, consider $\mathcal{G}_1$. This hypothesis class is a subset of $\mathcal{G}$, so any set shattered by $\mathcal{G}_1$ will also be shattered by $\mathcal{G}$. As a result, $d_{\mathcal{G}_1} \leq d_{\mathcal{G}}$. Similarly, any set shattered by $\mathcal{G}$ is shattered by $\mathcal{H}$, so $d_{\mathcal{G}_1} \leq d_{\mathcal{G}} \leq d$. Furthermore, the sum of the VC dimension and the number of samples in the second line is $d_{\mathcal{G}_1} + n - 1 \leq d + n - 1 = k - 1$.

Now consider $\mathcal{G}_2$. For every $g_2 \in \mathcal{G}_2, \exists g_1 \in \mathcal{G}_1$ which disagrees only on x_n. Therefore, if $T^X \subseteq \{x_1, ..., x_{n-1}\}$ is shattered by $\mathcal{G}_2$, $T^X \cup \{x_n\}$ must be shattered by $\mathcal{G}$. Because no set larger than d can be shattered by $\mathcal{G}$, $|T^X| \leq d - 1$. Hence, $d_{\mathcal{G}_2} \leq d - 1$. We will now apply this result with Claim 3 to create a bound on $|\mathcal{G}_2(S^X)|$.

$$\begin{aligned}
|\mathcal{G}_2(S^X)| &= |\mathcal{G}_2(\{x_1, ..., x_{n-1}\})| && \text{Definition of growth function} \\
&\leq g(n-1, \mathcal{G}_2) && \text{Inductive Hypothesis} \\
&\leq \Phi_{d_{\mathcal{G}_2}}(n-1) && \Phi_d \text{ increases with } d \\
&\leq \Phi_{d-1}(n-1)
\end{aligned}$$

With this result, we can prove the bound in Sauer's lemma.

$$\begin{aligned}
|\mathcal{H}(S^X)| &= |\mathcal{G}(S^X)| \\
&= |\mathcal{G}_1(S^X) \cup \mathcal{G}_2(S^X)|
\end{aligned}$$

$\{\mathcal{G}_1, \mathcal{G}_2\}$ is a partition of $\mathcal{G}$.

$$
\begin{aligned}
&= |\mathcal{G}_1(S^X)| + |\mathcal{G}_2(S^X)| \\
&\leq \Phi_d(n-1) + \Phi_{d-1}(n-1) \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \binom{n-1}{0} + \sum_{i=1}^{d} \binom{n-1}{i} + \sum_{i=1}^{d} \binom{n-1}{i-1} \\
&= \binom{n}{0} + \sum_{i=1}^{d} \left(\binom{n-1}{i} + \binom{n-1}{i-1} \right) \\
&= \binom{n}{0} + \sum_{i=1}^{d} \binom{n}{i} = \sum_{i=0}^{d} \binom{n}{i} = \Phi_d(n)
\end{aligned}
$$

$S^X \subseteq \mathcal{X}^n$ is arbitrary, so $g(n, \mathcal{H}) = \max_{|S^X|=n} |\mathcal{H}(S^X)| \leq \Phi_d(n)$. Therefore, Sauer's lemma holds in the inductive case.