CSC 588: Machine learning theory **Spring 2022**

Lecture 9: Sauer's Lemma and Its Applications

Lecturer: Chicheng Zhang *Scribe: Jing Wu*

1 Sauer's lemma

Lemma 1. *$\mathcal{H}$ is hypothesis class. Let $S = \{x_1, ..., x_n\}$ be a set of unlabeled examples then*

$$|\Pi_{\mathcal{H}}(S)| \leq |\{T \subseteq S : \mathcal{H} \text{ shatters } T\}|, \tag{1}$$

where $|\cdot|$ is the number of the elements of a finite set

Proof. The proof is given by mathematical induction.

Base case. Let $n = 1$ and $S = \{x_1\}$. If $\mathcal{H}$ agrees on x_1, without loss of generality all classifiers in $\mathcal{H}$ will produce positive classification results on x_1. Then $\Pi_{\mathcal{H}}(S) = \{(+)\}$ and only $T = \varnothing \subseteq S$ is shattered by $\mathcal{H}$. Both sides of (1) are equal to 1. If $\mathcal{H}$ disagrees on x_1, $\Pi_{\mathcal{H}}(S) = \{(+), (-)\}$. Two subsets of S, $\{x_1\}$ and $\varnothing$, are shattered by $\mathcal{H}$. Both sides of (1) are equal to 2. We have completed the proof for the base case.

Inductive case. Assume that $\forall S'$ of size $n-1$, $|\Pi_{\mathcal{H}}(S')| \leq |\{T \subseteq S' : \mathcal{H} \text{ shatters } T\}|$. We construct a set of hypothesis class $\mathcal{H}_S$ by selecting a representative from $\mathcal{H}$ for every labeling $(l_1, ..., l_n)$ in $\Pi_{\mathcal{H}}(S)$. Therefore by construction $|\mathcal{H}_S| = |\Pi_{\mathcal{H}}(S)|$. Hypothesis class $\mathcal{H}_S$ can be decomposed to $\mathcal{H}_1$ and $\mathcal{H}_2$ by the following procedure:

- For every labeling in S', $(l_1, ..., l_{n-1})$ if both $(l_1, ..., l_{n-1}, +)$ and $(l_1, ..., l_{n-1}, -)$ are achievable by $\mathcal{H}_S$, i.e. $\exists h_1, h_2 \in \mathcal{H}_S$ s.t. $(h_1(x_1), ..., h_1(x_n)) = (l_1, ..., l_{n-1}, +)$ and $(h_2(x_1), ..., h_2(x_n)) = (l_1, ..., l_{n-1}, -)$, then we send one to $\mathcal{H}_1$ and the other to $\mathcal{H}_2$.
- On the other hand, if only one of $(l_1, ..., l_{n-1}, +)$ and $(l_1, ..., l_{n-1}, -)$ is achievable, we send it to $\mathcal{H}_1$.

Observations:

$$|\mathcal{H}_1| \geq |\mathcal{H}_2|, \tag{2}$$

$$|\mathcal{H}_1| = |\Pi_{\mathcal{H}_1}(S')| \quad \text{and} \quad |\mathcal{H}_2| = |\Pi_{\mathcal{H}_2}(S')|, \tag{3}$$

$$|\mathcal{H}_S| = |\mathcal{H}_1| + |\mathcal{H}_2|. \tag{4}$$

(2) is true because every time we send an element to $\mathcal{H}_2$ we send another element to $\mathcal{H}_1$. (3) comes from the fact that classifiers in $\mathcal{H}_1$ and $\mathcal{H}_2$ generates unique labeling in S. So does S'. (4) is because any classifier in $\mathcal{H}_S$ gets sent to one of $\mathcal{H}_1$ and $\mathcal{H}_2$ by construction.

Then we consider a subset T of S. We make the following two further observations:

1. If $\mathcal{H}_1$ shatters T, then $\mathcal{H}_S$ shatters T. This is because $\mathcal{H}_1 \subseteq \mathcal{H}_S$.
2. If $\mathcal{H}_2$ shatters T then $\mathcal{H}_S$ shatters $T \cup \{x_n\}$. The reason is that, if $h_2 \in \mathcal{H}_2$ that achieves some labeling $(b_1, \ldots, b_k)$ on T, then by the decomposition rule, there must exists its twin $h_1 \in \mathcal{H}_1$ so that h_1 and h_2 produce label $(b_1, ..., b_k, +)$ and $(b_1, ..., b_k, -)$ on $T \cup \{x_n\}$, i.e. conditioned on achieving labeling $(b_1, \ldots, b_k)$ on T, both $+$ and $-$ are achievable for x_n by classifiers in $\mathcal{H}_S$. Since $\mathcal{H}_2$ shatters T, then $\mathcal{H}_S$ shatters $T \cup \{x_n\}$.

For S of size n, applying inductive hypothesis on $(\mathcal{H}_1, S')$ and $(\mathcal{H}_2, S')$, we have

$$\begin{aligned}|\mathcal{H}_S| &= |\mathcal{H}_1| + |\mathcal{H}_2|, \\ &= |\Pi_{\mathcal{H}_1}(S')| + |\Pi_{\mathcal{H}_2}(S')|, \\ &\leq |\{T \subseteq S' : \mathcal{H}_1 \text{ shatters } T\}| + |\{T \subseteq S' : \mathcal{H}_2 \text{ shatters } T\}|. \end{aligned} \tag{5}$$

Observe that

$$\{T \subseteq S' : \mathcal{H}_1 \text{ shatters } T\} = \{T \subseteq S : x_n \notin T, \ \mathcal{H}_1 \text{ shatters } T\} \subseteq \{T \subseteq S : x_n \notin T, \ \mathcal{H} \text{ shatters } T\}, \tag{6}$$

$$\{T \subseteq S' : \mathcal{H}_2 \text{ shatters } T\} = \{T \subseteq S : x_n \notin T, \ \mathcal{H}_2 \text{ shatters } T\} \subseteq \{T \subseteq S : x_n \notin T, \ \mathcal{H} \text{ shatters } T \cup \{x_n\}\}, \tag{7}$$

where the subset in (6) and (7) come from the "further observations" 1 and 2. Furthermore, note that the right hand side of (7) has size

$$|\{T \subseteq S : x_n \notin T, \ \mathcal{H} \text{ shatters } T \cup \{x_n\}\}| = |\{T \subseteq S : x_n \in T, \ \mathcal{H} \text{ shatters } T\}|,$$

where the equality is from the observation that there exists a one-to-one correspondence between the elements in the two families of sets: for any set T on the set family on the LHS, $T \cup \{x_n\}$ belongs to the set family on the RHS; for any set T on the set family on the RHS, $T \setminus \{x_n\}$ belongs to the set family on the LHS.

Putting (6) and (7) into (5) we obtain that $|\Pi_{\mathcal{H}}(S)| = |\mathcal{H}_S| \leq |\{T \subseteq S : x_n \notin T, \ \mathcal{H} \text{ shatters } T\}| + |\{T \subseteq S : x_n \in T, \ \mathcal{H} \text{ shatters } T\}| = |\{T \subseteq S : \mathcal{H} \text{ shatters } T\}|$. We have shown that the lemma is true for S of size n which complete the proof by induction.

2 Application of Sauer's lemma

First we recall the following corollary of Sauer's Lemma from last time.

Corollary 2. *If VC($\mathcal{H}$) = d and $n \geq 2$, then $\mathcal{S}(\mathcal{H}, n) \leq n^{d+1}$.*

Proof. By definition of the growth function $\mathcal{S}(\mathcal{H}, n) = \max_{S:|S|=n} |\Pi_{\mathcal{H}}(S)|$.
From Sauer's lemma, $|\Pi_{\mathcal{H}}(S)| \leq |\{T \subseteq S : \mathcal{H} \text{ shatters } T\}|$ for any S. Since the VC dimension of $\mathcal{H}$ is d, if $\mathcal{H}$ shatters T then $|T| \leq d$. So we have $\{T \subseteq S : \mathcal{H} \text{ shatters } T\} \subseteq \{T \subseteq S : |T| \leq d\}$. The size of set of the right hand size is $\sum_{i=0}^{d} \binom{n}{i}$ which is bounded by n^{d+1} numerically whenever $n \geq 2$. Therefore,

$$\mathcal{S}(\mathcal{H}, n) \leq n^{d+1}. \tag{8}$$

□

Example: bounding the VC dimension of composite hypothesis classes using Sauer's Lemma. Suppose we have a base hypothesis class $\mathcal{B}$, let define $\mathcal{B}_{f,k} = \{f(h_1(x), ..., h_k(x)) : h_1, ..., h_k \in \mathcal{B}\}$ for $f : \{0,1\}^k \longrightarrow \{0,1\}$ being a fixed Boolean function. Here are some examples of f:

1. $f(y_1, ..., y_k) = y_1 \oplus ... \oplus y_k$, (9)
2. $f(y_1, ..., y_k) = y_1 \vee ... \vee y_k$, (10)
3. $f(y_1, ..., y_k) =$ majority of $(y_1 ... y_k)$. (11)

Can we upper bound the complexity (VC dimension / growth function) of $\mathcal{B}_{f,k}$? The answer is yes. Here is a claim.

Claim 3.

$$\mathcal{S}(\mathcal{B}_{f,k}, n) \leq n^{2kd}. \tag{12}$$

Proof. Fix $S = (x_1, ..., x_n)$.

For each $h \in \mathcal{B}$, the number configurations of $(h(x_1), ..., h(x_n))$ is bounded by the growth function by definition, which is further up bounded by n^{d+1} from corollary. Therefore the total number of configurations of the matrix

$$M_{h_1,...,h_k} = \begin{bmatrix} h_1(x_1) & h_1(x_2) & \dots & h_1(x_n) \\ & \dots & & \\ h_k(x_1) & h_k(x_2) & \dots & h_k(x_n) \end{bmatrix}$$

is bounded by $n^{k(d+1)} \leq n^{2kd}$. As determining $M_{h_1,...,h_k}$ fully determines

$$\left(f\left(h_1(x_1), ..., h_k(x_1)\right), \dots, f\left(h_1(x_n), ..., h_k(x_n)\right)\right),$$

by evaluating function f on all n columns, the number of possible labelings $\mathcal{B}_{f,k}$ achieves on S is at most the number of possible $M_{h_1,...,h_k}$'s, then we get (12).

Theorem 4. *Let $v = \mathrm{VC}(\mathcal{B}_{f,k})$ be the VC dimension of $\mathcal{B}_{f,k}$, then $v \leq 8kd \ln(8kd) = \tilde{\mathcal{O}}(kd)$.*

Proof. From the definition of VC dimension, $\mathcal{S}(\mathcal{B}_{f,k}, v) = 2^v$.
From the claim 3 we also have $\mathcal{S}(\mathcal{B}_{f,k}, v) \leq v^{2kd}$. Then

$$\begin{aligned} 2^v \leq v^{2kd} \Rightarrow v &\leq 2kd \log_2 v, \\ &\leq 4kd \ln v. \end{aligned} \tag{13}$$

By the following lemma, (13) leads to $v \leq 8kd \ln(8kd)$ by letting $a = 4kd$ and $b = 0$. □

Lemma 5. *If $a > 0, b \geq 0, x > 0$ and $x \leq a \ln x + b$ then $x \leq 2a \ln(2a) + 2b$.*

Proof. Since $\forall t > 0$, $\ln t \leq t$.
Substituting $t = \frac{x}{2a}$, we have

$$\ln x \leq \ln(2a) + \frac{x}{2a}.$$

Then from our assumption

$$\begin{aligned} x &\leq a \ln x + b, \\ &\leq a\left(\ln(2a) + \frac{x}{2a}\right) + b, \\ &\leq a \ln(2a) + \frac{x}{2} + b. \end{aligned} \tag{14}$$

Therefore $x \leq 2a \ln(2a) + 2b$ from (14). □

3 Uniform convergence

If $\mathcal{H}$ has a finite VC dimension then as the number of training examples m increases, the empirical error converges to its generalization error and the difference is bounded in terms of $\mathcal{H}$'s VC dimension.

Theorem 6. *Suppose hypothesis class $\mathcal{H}$ has VC dimension d. Then given set of m i.i.d. training samples $(x_1, y_1), ..., (x_n, y_n)$ from distribution D. With probability at least $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} |err(h, S) - err(h, D)| \leq c_1 \sqrt{\frac{\ln \mathcal{S}(\mathcal{H}, n) + \ln \frac{1}{\delta}}{n}} \leq c_2 \sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}} \tag{15}$$

for some absolute constants $c_1, c_2 > 0$.