CSC 588: Machine learning theory **Spring 2021**

Lecture 7: Sauer's Lemma - Proof and Applications

Lecturer: Chicheng Zhang *Scribe: Marium Yousuf*

1 Sauer's Lemma

The Sauer's Lemma tells how many labelings can a hypothesis class, $\mathcal{H}$, generate for a sequence of unlabeled examples, S.

Lemma 1. *Suppose we have a non-empty hypothesis class $\mathcal{H}$ and a sequence of unlabeled examples $S = (x_1, ...x_n)$, such that:*

$$|\Pi_{\mathcal{H}}(S)| \leq |\{T \subseteq S : \mathcal{H} \textit{ shatters } T\}|.$$

Consequently, if $\mathrm{VC}(\mathcal{H}) = d$, *then*

$$|\{T \subseteq S : \mathcal{H} \textit{ shatters } T\}| \leq \{T \subseteq S : |T| \leq d\} \leq \sum_{i=0}^{d} \binom{n}{i} = \binom{n}{\leq d},$$

such that, $\binom{n}{\leq d}$ is a polynomial in n with exponent d.

Remark: The converse of the Sauer's Lemma is true up to log factors. That is, if we can show

$$S(\mathcal{H}, n) \leq \sum_{i=0}^{d} \binom{n}{i},$$

for any n, then

$$VC(\mathcal{H}) \leq \mathcal{O}(d \ln d).$$

The proof of the converse is left as an exercise.

1.1 Proof

Sauer's Lemma can be proved by induction on n, the size of S.

Proof. Using $n = 1$ as our base case, we have two situations to consider:

1. $\mathcal{H}$ agrees on x_1:
   In this case all classifiers $h \in \mathcal{H}$ will predict *one* label unanimously, $h : x_1 \to +1$. Consequently, X_1 would not be shattered by $\mathcal{H}$, which then would only shatter $\emptyset$. Therefore,

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

2. $\mathcal{H}$ disagrees on x_1:
   In this case, we can find two classifiers $h \in \mathcal{H}$ that disagree on x_1. That is, suppose we have $h_1, h_2 \in \mathcal{H}$ such that $h_1 : x_1 \to +1$ and $h_2 : x_1 \to -1$. Therefore,

$$|\Pi_{\mathcal{H}}(S)| = 2$$

   And, since $\mathcal{H}$ shatters both X_1 and $\emptyset$, we have:

$$|\{T \subseteq S : \mathcal{H} \text{ shatters } T\}| = 2.$$

   And so the base case holds since $|\Pi_{\mathcal{H}}(S)| \leq |\{T \subseteq S : \mathcal{H} \text{ shatters } T\}|$ for both situations.

Now to apply induction on n, we assume that $\forall S'$ of size $n-1$,

$$|\Pi_{\mathcal{H}}(S')| \leq |\{T \subseteq S' : \mathcal{H} \text{ shatters } T\}|$$

as our inductive hypothesis. Now to prove for Sauer's lemma holds for $n \geq 2$, we need an upper-bound for $|\Pi_{\mathcal{H}}(S)|$, where $|S| = n$. Also, set $S' = \{x_1, ..., x_{n-1}\}$ and note that in this notation, $S = S' \cup \{x_n\}$.

$\mathcal{H}_S$: for every label $(l_1, ..., l_n)$ in $|\Pi_{\mathcal{H}}(S)|$, we select one representative in $\mathcal{H}$ and add it to $\mathcal{H}_S$ such that $|\mathcal{H}_S| = |\Pi_{\mathcal{H}}(S)|$.

Consider the example where $n = 3$, then $\mathcal{H}_s$ generates all possible labels except $x_1 \to +1$, $x_2 \to -1$, and $x_3 \to +1$. Now, to achieve the goal of upper-bounding the cardinality of $\mathcal{H}_s$, we need to decompose $\mathcal{H}_s$ into two parts. Still considering $n = 3$, we know $S' = \{x_1, x_2\}$ and $S = S' \cup \{x_3\}$, and so for every label on $S' : (l_1 = -1, l_2 = -1)$, any two $h_i \in \mathcal{H}_s$, where $i = \{1, ..., 7\}$, will achieve the mentioned labels. Now, to decompose $\mathcal{H}_s$ for $n = 3$,

1. if both $(l_1, l_2, +1)$ and $(l_1, l_2, -1)$ are achievable by $\mathcal{H}_s$, send one classifier to $\mathcal{H}_1$ and send the other to $\mathcal{H}_2$.

2. if only one of $(l_1, l_2, +1)$ and $(l_1, l_2, -1)$ is achievable by $\mathcal{H}_s$, then we send the classifier to $\mathcal{H}_1$, giving $\mathcal{H}_1$ the priority.

This can be further generalized to any n. Note that $|\mathcal{H}_1| \geq |\mathcal{H}_2|$, since $\mathcal{H}_1$ is prioritized.

Observations:

1. $|\mathcal{H}_1| = |\Pi_{\mathcal{H}_1}(S')|$ and $|\mathcal{H}_2| = |\Pi_{\mathcal{H}_2}(S')|$. This is because all classifiers $h \in \mathcal{H}_1$ and $\mathcal{H}_2$ generate unique labelings in S'.

2. If $T \subseteq S'$ and $\mathcal{H}_1$ shatters T, then $\mathcal{H}_s$ shatters T. Consider an example where $\mathcal{H}_1$ shatters $\{x_1, x_2\}$, then $\mathcal{H}_s$ also shatters $\{x_1, x_2\}$ because $\mathcal{H}_1 \subset \mathcal{H}_s$.

3. If $T \subseteq S'$ and $\mathcal{H}_2$ shatters T, then $\mathcal{H}_s$ shatters $T \cup \{x_n\}$. Consider an example where $\mathcal{H}_2$ shatters $\{x_2\}$, then $\mathcal{H}_s$ shatters $\{x_2, x_3\}$. In general, if $\mathcal{H}_2$ achieves $b_1, ..., b_{|T|}$ on T, then $\mathcal{H}$ will achieve both $(b_1, ..., b_{|T|}, +1)$ and $(b_1, ..., b_{|T|}, -1)$ on $T \cup \{x_n\}$.

Using these observations, we can conclude that

$$\begin{aligned}
|\mathcal{H}_s| &= |\mathcal{H}_1| + |\mathcal{H}_2| \\
&= |\Pi_{\mathcal{H}_1}(S')| + |\Pi_{\mathcal{H}_2}(S')| \text{ (by observation 1)} \\
&\leq |\{T \subseteq S' : T \text{ shattered by } \mathcal{H}_1| + |\{T \subseteq S' : T \text{ shattered by } \mathcal{H}_2\}| \\
&\leq |\{T \subseteq S' : T \text{ shattered by } \mathcal{H}_s| \text{ (by observation 2)} \\
&\qquad\qquad + |\{T \subseteq S' : T \cup \{x_n\} \text{ shattered by } \mathcal{H}_s\}| \text{ (by observation 3).}
\end{aligned}$$

Note that $|\{T \subseteq S' : T \cup \{x_n\}.$ shattered by $\mathcal{H}_s| = |\{T \subseteq S : X_n \in T, T$ shattered by $\mathcal{H}_s|$, that is both L.H.S and R.H.S have the same cardinality as there is a one-to-one correspondence between their elements. Using this, we have:

$$\begin{aligned}
&\implies |\{T \subseteq S' : T \text{ shattered by } \mathcal{H}_s| + |\{T \subseteq S' : T \cup \{x_n\} \text{ shattered by } \mathcal{H}_s\}| \\
&= \ |\{T \subseteq S : T \text{ shattered by } \mathcal{H}_s\}|.
\end{aligned}$$

□

Note: $|\mathcal{H}_1| \geq |\mathcal{H}_2|$; this can also be seen by noticing that $|\{T \subseteq S' : T$ shattered by $\mathcal{H}_s| \geq |\{T \subseteq S' : T \cup \{X_n\}$ shattered by $\mathcal{H}_s\}|$.

1.2 Applications for bounding VC dimensions of hypothesis classes

Consider an example with a base hypothesis class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = d$. Let k by an odd number such that, $\mathcal{H}_{\mathrm{maj},k} = \{\mathrm{maj}(h_1(X), ..., h_k(X) : h_1, ..., h_k \in \mathcal{H})\}$, where $\mathrm{maj}(y_1, ..., y_n) = \begin{cases} +1 & |\{i, y_i = +1\}| > k/2 \\ -1 & |\{i, y_i = +1\}| \leq k/2 \end{cases}$. Can we bound $\mathrm{VC}(\mathcal{H}_{(\mathrm{maj},k)})$?

Claim: $S(\mathcal{H}_{(\mathrm{maj},k)}, n) \leq n^{k(d+1)}$. Since $\mathrm{VC}(\mathcal{H}_{(\mathrm{maj},k)}) = \max\{n : 2^n = S(\mathcal{H}_{(\mathrm{maj},k)})\}$, the n that satisfies $2^n = S(\mathcal{H}_{(\mathrm{maj},k)})$ is upper-bounded. That is, the growth function $S(\mathcal{H}_{(\mathrm{maj},k)}$ is polynomial and so equating it with the exponential function 2^n, there must exist a size point, n^*, such that VC dimension is upper bounded by that n^* (Figure 1).

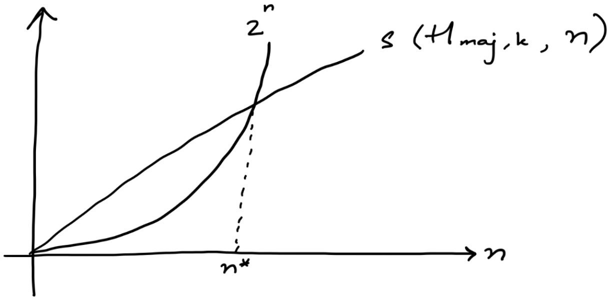

Figure 1: Comparison of a polynomial $S(\mathcal{H}_{\mathrm{maj},k}, n)$ meeting a exponential function 2^n at n^*.

Suppose we are given a dataset $x_1, ..., x_n$, how many labels can be generated using the majority class? By Sauer's lemma, we know that for each $h_i \in \mathcal{H} : i = 1, ..., k$, the number of labels is upper-bounded by n^{d+1}. The key observation here is that after fixing the labels of base classifiers of n examples, the behavior of the majority class is determined. Therefore, the number of possible configurations for $\mathrm{maj}(h_1, ..., h_k) \leq (n^{d+1})^k$, which is the upper-bound of the growth function $S(\mathcal{H}_{\mathrm{maj},k}, n)$.

Now, using this $(n^{d+1})^k$ and intuition of the plots from Figure 1, we need to find an upper-bound of the VC-dimension. Suppose $n : 2^n \leq (n^{d+1})^k$. Therefore, $n \leq 2(d+1)k \ln n$. Using Lemma 2 mentioned below, letting $a = 2(d+1)k$ and $b = 0$, we have $n \leq 4(d+1)k \ln(4(d+1)k) = \tilde{\mathcal{O}}(d \cdot k)$.

This proves that $\mathrm{VC}(\mathcal{H}_{\mathrm{maj},k}) = \tilde{\mathcal{O}}(d \cdot k)$. Notice here that if the base class is expressive, the composite class will also be expressive. Similarly, if k is higher, that is, if more base case classifiers are aggregated together, then the resulting classifier will be more complicated.

Note that if $\mathcal{H}_{\mathrm{maj},k}$ function is replaced by any other function $\mathcal{H}_{f,k}$, the VC-dimension of the resulting class will still be upper-bounded by $(n^{d+1})^k$, with the $\tilde{\mathcal{O}}(d \cdot k)$. In this case, the bound can be loose for some 'simple' f.

Lemma 2. *Given $a \geq 1/2$ and $b > 0$, if $x \leq a \ln x + b$, then $x \leq 2a \ln(2a) + 2b$.*

Proof.

$$\begin{aligned}
\text{Given:}\quad & \ln\frac{x}{2a} \le \frac{x}{2a} \\
\implies & \ln x \le \frac{x}{2a} + \ln(2a) \\
= \ & x \le a\ln x + b \\
\le \ & \frac{x}{2} + a\ln(2a) + b \\
\implies & \frac{x}{2} \le +a\ln(2a) + b \\
\implies & x \le 2a\ln(2a) + 2b
\end{aligned}$$

□

2 Uniform Convergence

Recall that if $\mathcal{H}$ has a finite VC-dimension then with high probability all classifier's empirical error will concentrate around its generalization error.

Theorem 3. *Given a hypothesis class $\mathcal{H}$ with $VC(\mathcal{H}) = d$, a set of n i.i.d. training examples $(X_1, Y_1), ..., (X_n, Y_n)$ from $\mathcal{D}$, then with probability $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} |\text{err}(h, S) - \text{err}(h, \mathcal{D})| \le c_1 \sqrt{\frac{d \ln \frac{d}{n} + \ln 1/\delta}{n}},$$

for some constant c_1.

Note that if a class has a larger VC-dimension then the uniform control of empirical error to the generalization error will be looser. When sample size increases we expect the empirical error to concentrate around the generalization error.

Consequently, ERM on $\mathcal{H}$ has an agnostic PAC sample complexity of

$$f(\varepsilon, \delta) = \mathcal{O}\left(\frac{1}{\varepsilon^2}\left(d \ln \frac{1}{\varepsilon} + \ln \frac{1}{\delta}\right)\right).$$

This implication holds because $\text{err}(\hat{h}, \mathcal{D}) - \min_{h' \in \mathcal{H}} \text{err}(h', \mathcal{D}) \le \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) \le \varepsilon$, when $n \ge \tilde{\mathcal{O}}(\frac{d}{\varepsilon^2})$. We can achieve the agnostic PAC learning goal by showing that $\tilde{\mathcal{O}}\left(\sqrt{\frac{d}{n}}\right) \le \varepsilon$.