Machine Learning **Spring 2021**

Lecture 3: VC Theory: Generalization Theory for ERM

Lecturer: Liwei Wang *Scribe: Yixin Yang, Siqi Yang, Fan Chen, Lexing Zhang*

3.1 From finite number to infinite number

There are two measurements of the performance of a classifier:

- Training error: $P_S(y \neq f(x)) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{y_i \neq f(x_i)\}$
- Generalization error[1]: $P_{\mathcal{D}}(y \neq f(x)) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{y \neq f(x)\}\right]$

We are interested in the **generalization gap** $P_{\mathcal{D}}(y \neq f(x)) - P_S(y \neq f(x))$. Intuitively, it is closely related to the complexity of our model.

Even though it's unlikely that the number of classifiers can be finite in practice, it's important for us to first consider the finite case:

Theorem 3.1 *Suppose we have a finite model class $\mathcal{F}, |\mathcal{F}| < \infty$. For any classifier f in $\mathcal{F}$, we have the bound*

$$\mathbb{P}(P_{\mathcal{D}}(y \neq f(x)) - P_S(y \neq f(x)) \geq \epsilon) \leq |\mathcal{F}|e^{-2n\epsilon^2}$$

Proof:

$$\begin{aligned}
& \mathbb{P}(P_{\mathcal{D}}(y \neq f(x)) - P_S(y \neq f(x)) \geq \epsilon) \\
\leq\ & \mathbb{P}(\exists f \in \mathcal{F}, P_D(y \neq f(x)) - P_S(y \neq f(x)) \geq \epsilon) \\
\leq\ & \sum_{f\in\mathcal{F}} \mathbb{P}\left(P_{\mathcal{D}}(y \neq f(x)) - P_S(y \neq f(x)) \geq \epsilon\right) \qquad \text{(Union bound)} \\
\leq\ & |F|e^{-2n\epsilon^2} \qquad \text{(Chernoff Bound)}
\end{aligned}$$

■

Thus in conclusion, when $|\mathcal{F}|$ is finite, the generalization gap can be estimated by applying the union bound. That is, for a finite collection of events $A_1, A_2, \ldots, A_n$ we have

$$\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_i \mathbb{P}\left(A_i\right)$$

What if the model class $\mathcal{F}$ is infinite?

[1] We use $\mathcal{D}$ to denote the distribution of $z = (x, y)$.

3.2 VC-Theory

For the sake of simplicity, we first introduce some notations:

- for $f \in \mathcal{F}$ and $z = (x, y)$ sampled from the distribution $\mathcal{D}$, define $\phi_f(z) = \mathbb{1}\{f(x) \neq y\}$
- define $\Phi = \{\phi_f : f \in \mathcal{F}\}$

Our ultimate goal in this section is to bound the following quantity:

$$\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \mathbb{E}_z\left[\phi(z)\right]\right| \geq \epsilon\right)$$

3.2.1 Step1: Double Sample Trick

Lemma 3.2 *Let $X, X_1, X_2, \ldots, X_{2n}$ be i.i.d. Bernoulli random variables and $p = \mathbb{E}[X]$, $V_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, V_2 = \frac{1}{n}\sum_{i=n+1}^{2n} X_i$. If $n \geq \frac{\ln 2}{\epsilon^2}$, then*

$$\frac{1}{2}\mathbb{P}\left(|V_1 - p| \geq 2\epsilon\right) \leq \mathbb{P}\left(|V_1 - V_2| \geq \epsilon\right) \leq 2\mathbb{P}\left(|V_1 - p| \geq \frac{\epsilon}{2}\right)$$

Proof: For the inequality on the right side,

$$|V_1 - V_2| \geq \epsilon \Rightarrow \left(|V_1 - p| \geq \frac{\epsilon}{2}\right) \vee \left(|V_2 - p| \geq \frac{\epsilon}{2}\right)$$

Using the union bound, we have

$$\mathbb{P}\left(|V_1 - V_2| \geq \epsilon\right) \leq \mathbb{P}\left(|V_1 - p| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|V_2 - p| \geq \frac{\epsilon}{2}\right) = 2\mathbb{P}\left(|V_1 - p| \geq \frac{\epsilon}{2}\right)$$

For the inequality on the left side, since $|V_1 - p| \geq 2\epsilon$ and $|V_2 - p| < \epsilon$ are independent events and

$$|V_1 - p| \geq 2\epsilon \wedge |V_2 - p| < \epsilon \Rightarrow |V_1 - V_2| \geq \epsilon$$

we have

$$\mathbb{P}\left(|V_1 - p| \geq 2\epsilon\right)\mathbb{P}\left(|V_2 - p| < \epsilon\right) \leq \mathbb{P}\left(|V_1 - V_2| \geq \epsilon\right)$$

Using Chernoff bound we can infer that $\mathbb{P}\left(|V_2 - p| < \epsilon\right) \leq \frac{1}{2}$ when $n \geq \frac{\ln 2}{\epsilon^2}$, thus

$$\frac{1}{2}\mathbb{P}\left(|V_1 - p| \geq 2\epsilon\right) \leq \mathbb{P}\left(|V_1 - V_2| \geq \epsilon\right)$$

■

Lemma 3.3 *If $n \geq \frac{\ln 2}{\epsilon^2}$, then*[2]

$$\begin{aligned}\frac{1}{2}\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\mathbb{E}\left[\phi(Z)\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(z_i)\right| \geq 2\epsilon\right) &\leq \mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right| \geq \epsilon\right) \\ &\leq 2\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\mathbb{E}\left[\phi(Z)\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(z_i)\right| \geq \frac{\epsilon}{2}\right)\end{aligned} \tag{3.1}$$

Proof: Left as homework. ■

[2] A slightly tricky point here is that: we are taking "sup" over a possibly uncountable family of random variables, so the event $\{\sup_{\phi\in\Phi}\left|\mathbb{E}\left[\phi(Z)\right] - \frac{1}{n}\sum_{i=1}^{n}\phi(z_i)\right| \geq \epsilon\}$ may not be $\mathbb{P}$-measurable. But this is just a purely technical issue in Probability Theory, and there is nothing to worry about.

3.2.2 Step2: Symmetrization

For sampling $z_i = (x_i, y_i)$ from the probability distribution $\mathcal{D}$, consider two equivalent sampling methods below:

1. Draw $z_1, \ldots, z_{2n}$ from $\mathcal{D}$ independently.
2. Draw set $\{z_1, \ldots, z_{2n}\}$ from $\mathcal{D}$ and randomly draw a permutation σ.

Consider the second sampling methods, we have

$$\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right|\geq\epsilon\right)=\mathop{\mathbb{E}}_{\{z_1,\ldots,z_{2n}\}}\left[\mathbb{P}_\sigma\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geq\epsilon\right)\right]$$

Let $N^\Phi(z_1, \ldots, z_n)$ denotes $|\{(\phi(z_1), \ldots, \phi(z_n)) : \phi \in \Phi\}|$.

Using union bound, we have [3]

$$\mathbb{P}_\sigma\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geq\epsilon\right)\leq N^\Phi(z_1,\ldots,z_{2n})\mathbb{P}_\sigma\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geq\epsilon\right)$$

Using "draw without replacement" Chernoff bound, we have

$$\begin{aligned}
&\mathbb{P}_\sigma\left(\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\right|\geq\epsilon\right)\\
=\;&2\mathbb{P}_\sigma\left(\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_{\sigma(i)})\geq\epsilon\right)\\
=\;&2\mathbb{P}_\sigma\left(\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-\frac{1}{2n}\sum_{i=1}^{2n}\phi(z_{\sigma(i)})\geq\frac{\epsilon}{2}\right)\\
=\;&2\mathbb{P}_\sigma\left(\frac{1}{n}\sum_{i=1}^{n}\phi(z_{\sigma(i)})-p\geq\frac{\epsilon}{2}\right)\leq 2e^{-2n(\frac{\epsilon}{2})^2}=2e^{-\frac{n\epsilon^2}{2}}
\end{aligned}$$

Therefore, we got

$$\begin{aligned}
\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i)-\frac{1}{n}\sum_{i=n+1}^{2n}\phi(z_i)\right|\geq\epsilon\right)&\leq\mathop{\mathbb{E}}_{\{z_1,\ldots,z_{2n}\}}\left[N^\Phi(z_1,\ldots,z_{2n})\cdot 2e^{-\frac{n\epsilon^2}{2}}\right]\\
&=\mathop{\mathbb{E}}_{\{z_1,\ldots,z_{2n}\}}\left[N^\Phi(z_1,\ldots,z_{2n})\right]\cdot 2e^{-\frac{n\epsilon^2}{2}} \qquad (3.2)
\end{aligned}$$

[3] In fact, the RHS of the inequality below should be a summation over the set $\{(\phi(z_1), \ldots, \phi(z_{2n})) : \phi \in \Phi\}$ (i.e. over all dichotomies of $\{z_1, \cdots, z_{2n}\}$), which has $N^\Phi(z_1, \ldots, z_{2n})$ elements. But for the sake of simplicity and clarity, we follow the notation used in class, so from here on, when we write $\phi(z_1), \cdots, \phi(z_{2n})$, we refer to a specific element in $\{(\phi(z_1), \ldots, \phi(z_{2n})) : \phi \in \Phi\}$ (i.e. a dichotomy).

3.2.3 Step3: Estimate Growth Function

Now we have to deal with the term $\mathop{\mathbb{E}}_{\{z_1,\dots,z_{2n}\}}\left[N^\Phi(z_1,\dots,z_{2n})\right]$ in section 3.2. In the following part, we use $N^\Phi(n)$, the growth function of Φ, to represent the "worst case":

Definition 3.4 $N^\Phi(n) := \max_{z_1,\cdots,z_n\sim\mathcal{D}} N^\Phi(z_1,\cdots,z_n)$.

Concerning the behaviour of $N^\Phi(n)$ (as function of n), there are essentially two cases:

1. For all n, $N^\Phi(n) = 2^n$
2. At some n_0, $N^\Phi(n_0) < 2^{n_0}$

In case of the first one, all work we have done before will turn out to be useless, if not trivial[4]. And in the second case, by exactly the same argument, we shall expect $N^\Phi(n)$ to behaving well. So the main problem is: we know that $\forall n \geq n_0$, $N^\Phi(n)$ is strictly smaller than 2^n, but to what extent?

The methodology is to try with some easy examples. The insight is that there is a special case with enough generality:

Example 3.5 *Set $d = n_0 - 1$ and work with fixed sample $z_1,\cdots,z_n$. We know that for every $z_{i_1},\cdots,z_{i_{d+1}} (i_1 < \cdots < i_{d+1})$, there some strings [5] of length $d+1$ which is not "attainable" [6] at $z_{i_1},\cdots,z_{i_{d+1}}$. We **assume** that these "forbidden" strings can always be chosen to be $(0,\cdots,0)$.*

Then, any attainable string of length n can only contain at most d many 0's. Thus the the number of attainable strings, i.e. $N^\Phi(z_1,\cdots,z_n)$, is no more than $\sum_{k=0}^{d}\binom{n}{k}$.

In general setting, the forbidden patterns need not be the same for different sub-strings, but (intuitively) different patterns will "overlap" to create more forbidden patterns, and therefore we may expect there are less many possibility in the general case.

Lemma 3.6 *Assume that $N^\Phi(d+1) < 2^{d+1}$ and fix a sample $z_1,\cdots,z_n$ with $n \geq d+1$. Then we have $N^\Phi(z_1,\cdots,z_n) \leq \sum_{k=0}^{d}\binom{n}{k}$.* [7]

Proof: To bound the $N^\Phi(z_1,\cdots,z_n)$, we have to give a lower bound of the number of strings not attainable. The idea here is to look at the forbidden pattern of each sub-string, and then "extend" them. That is, we may naturally **extend** a forbidden pattern w (at $\{i_1 < \cdots < i_{d+1}\}$) to a set $E(w)$ of not attainable n-strings, e.g. we can extend the pattern $(0,1,1)$ at $2,3,5$ to $\{(*,0,1,*,1)\}$

Now, for each $\mathcal{I} = \{i_1 < \cdots < i_{d+1}\} \subset \{1,2,\cdots,n\}$, we choose a forbidden pattern $w_\mathcal{I}$, and extend it to a set of n-strings $E(w_\mathcal{I})$,[8] and we have to (lower) bound the $|\bigcup_\mathcal{I} E(w_\mathcal{I})|$:

[4] In fact, in such case, our hypothesis class is in some sense not learnable. For reference, see [FML12], Chapter 3, Section 5.

[5] From now on, when we talk about **(k-)string**, we refer to a vector (of length k), with each component being 0 or 1.

[6] We say a k-string is **not attainable** or it is a **forbidden pattern** at $\mathcal{I} = \{i_1 < \cdots < i_{d+1}\}$, if it is not contain in $\left\{\left(\phi(z_{i_1}),\cdots,\phi(z_{i_{d+1}})\right) : \phi \in \Phi\right\}$.

[7] When $n \leq d$, $\sum_{k=0}^{d}\binom{n}{k} = 2^n$, so the bound in fact holds for all n.

[8] In other word, we choose a pattern that cannot be produced by $z_{i_1},\cdots,z_{i_{d+1}}$, for each $d+1$ subset of $z_1,\cdots,z_n$.

First, consider the what happen at 1-st component. There are three possibility, e.g.

$$\{(*,1,0,*,\cdots)\}$$
$$\{(*,*,1,*,\cdots)\}$$
$$\{(1,0,*,0,\cdots)\}$$
$$\{(0,0,1,*,\cdots)\}$$
$$\{(0,*,*,1,\cdots)\}$$
$$\cdots$$

and we thus consider:

$$S_0 := \bigcup_{w_{\mathcal{I}} \text{ is 0 at 1-st}} E(w_{\mathcal{I}})$$
$$S_1 := \bigcup_{w_{\mathcal{I}} \text{ is 1 at 1-st}} E(w_{\mathcal{I}})$$
$$S_2 := \bigcup_{\text{no restriction at 1-st}} E(w_{\mathcal{I}})$$

Now for each $w_{\mathcal{I}}$ having 1 at 1-st component, we change its 1-st 1 to 0 to obtain $w'_{\mathcal{I}}$, and get $S'_1 = \bigcup E(w'_{\mathcal{I}})$. From definition, we have

$$|S'_1| = |S_1|, \quad S_0 \cap S_1 = \emptyset, \quad |S_2 \cap S_1| = |S_2 \cap S'_1|$$

and therefore we have $|S_0 \cup S_1 \cup S_2| \geq |S_0 \cup S'_1 \cup S_2|$.

Apply this procedure to each place, we can finally have each pattern $w_{\mathcal{I}}^{new}$ being the $d+1$-string $(0,\cdots,0)$, and thus

$$\begin{aligned} |\{\text{not attainable } n\text{-string}\}| \geq \left|\bigcup_{\mathcal{I}} E(w_{\mathcal{I}})\right| = |S_0 \cup S_1 \cup S_2| &\geq |S_0 \cup S'_1 \cup S_2| \\ &\geq \cdots \\ &\geq \left|\bigcup_{\mathcal{I}} E(w_{\mathcal{I}}^{new})\right| \\ &= \sum_{k=d+1}^{n} \binom{n}{k} \end{aligned}$$

Therefore $N^{\Phi}(z_1,\cdots,z_n) \leq 2^n - \sum_{k=d+1}^{n} \binom{n}{k} = \sum_{k=0}^{d} \binom{n}{k}$. ■

Corollary 3.7 *Combining inequalities 3.1, 3.2 and* $\sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d = \mathcal{O}(n^d)$*, we finally have*

$$\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(z_i) - \mathbb{E}_z\left[\phi(z)\right]\right| \geq \epsilon\right) \leq 2\left(\frac{2en}{d}\right)^d e^{-\frac{n\epsilon^2}{8}}$$

Definition 3.8 *Define the VC dimension of $\mathcal{F}$ to be the maximal d such that $N^{\Phi}(d) = 2^d$. If there is no such d, the VC dimension of $\mathcal{F}$ is defined to be $+\infty$.*

In other word, if we assume $\mathcal{F}$ has VC-dimension d, then there exist $z_1,\cdots,z_d$ satisfying $N^{\Phi}(z_1,\cdots,z_d) = 2^d$, and for all $z_1,\cdots,z_{d+1}$ drawn from $\mathcal{D}$, $N^{\Phi}(z_1,\cdots,z_{d+1}) < 2^{d+1}$.

References

[FML12] Mohri. M, Rostamizadeh. A and Talwalkar. A, Foundations of Machine Learning, *MIT Press* (2012)

[1] https://en.wikipedia.org/wiki/Boole's_inequality

[2] https://en.wikipedia.org/wiki/Vapnik-Chervonenkis_dimension