ELEG/CISC 867: Advanced Machine Learning **Spring 2019**

Lecture 6: Fundamental Theorem of Statistical Learning

Lecturer: Xiugang Wu *03/12/2019, 03/14/2019*

Last time we have shown that a class of infinite VC-dimension is not learnable. The converse statement is also true, leading to the fundamental theorem of statistical learning theory.

Theorem 0.1 (Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0-1 loss. Then the following statements are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *ERM is a successful PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is PAC learnable.*
4. *The VC-dimension of $\mathcal{H}$, denoted by d, is finite.*

We have shown 1) $\to$ 2) in previous lectures. The implication 2) $\to$ 3) is trivially by the definition of PAC learnability. The implication 3) $\to$ 4) follows from No Free Lunch Theorem: if the VC-dimension of $\mathcal{H}$ is infinite, then $\mathcal{H}$ is not learnable. Here we will show that 4) $\to$ 1). The proof is based on two main claims:

- Sauer's Lemma: If VC-d$(\mathcal{H}) = d$, then even though $\mathcal{H}$ might be infinite, when restricting it to a finite set $C \subseteq \mathcal{X}$, its "effective" size, $|\mathcal{H}_C|$, is only $O(|C|^d)$.
- Uniform convergence holds whenever the hypothesis class has a "small effective" size, i.e. $|\mathcal{H}_C|$ grows polynomially with $|C|$.

1 Sauer's Lemma and the Growth Function

Definition 1.1 (Growth Function) *The growth function of a hypothesis class $\mathcal{H}$, denoted by $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as*

$$\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}: |C| = n} |\mathcal{H}_C|.$$

In words, $\tau_{\mathcal{H}}(n)$ is defined as the maximal number of different functions from a set C of size n to $\{0,1\}$ that can be obtained by restricting $\mathcal{H}$ to C.

Obviously, if VC-d$(\mathcal{H}) = d$, then for any $n \leq d$ we have $\tau_{\mathcal{H}}(n) = 2^n$. In such cases, $\mathcal{H}$ induces all possible functions from C to $\{0,1\}$. The following lemma, proposed independently by Sauer, Shelah and Perles, shows that when n becomes larger than the VC-dimension, the growth function increases polynomially rather than exponentially with n.

Lemma 1.1 (Sauer-Shelah-Perles) *Let $\mathcal{H}$ be a hypothesis class with VC-d$(\mathcal{H}) = d < \infty$. Then for all n, $\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$. In particular, if $n > d$ then $\tau_{\mathcal{H}}(n) \leq (en/d)^d$.*

2 Uniform Convergence for Classes of Small Effective Size

We now show that uniform convergence holds whenever the hypothesis class has a "small effective" size, i.e. $|\mathcal{H}_C|$ grows polynomially with $|C|$. In particular, we have the following theorem, which relates the generalization error to the growth function of $\mathcal{H}$.

Theorem 2.1 *Let $\mathcal{H}$ be a class and let $\tau_{\mathcal{H}}$ be its growth function. Then for every P and every $\delta \in (0,1)$, we have*

$$P^n\left(\sup_{h\in\mathcal{H}} |L(h,P) - L(h,P_{Z^n})| \le \frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}\right) \ge 1-\delta. \tag{1}$$

2.1 Proof of Theorem 0.1

Before we prove Theorem 2.1, we will first use it to conclude the proof of Theorem 0.1, i.e. to show 4) $\to$ 1) in Theorem 0.1. From Sauer's lemma we have that for $n > d$, $\tau_{\mathcal{H}}(2n) \le (2en/d)^d$. Combining this with Theorem 2.1, we have

$$P^n\left(\sup_{h\in\mathcal{H}} |L(h,P) - L(h,P_{Z^n})| \le \frac{4+\sqrt{d\log(2en/d)}}{\delta\sqrt{2n}}\right) \ge 1-\delta.$$

For simplicity assuming that $\sqrt{d\log(2en/d)} \ge 4$, we have

$$P^n\left(\sup_{f\in\mathcal{F}} |L(f,P) - L(f,P_{Z^n})| \le \frac{1}{\delta}\sqrt{\frac{2d\log(2en/d)}{n}}\right) \ge 1-\delta.$$

To ensure the generalization error is at most ϵ we need that

$$n \ge \frac{2d\log n}{(\delta\epsilon)^2} + \frac{2d\log(2e/d)}{(\delta\epsilon)^2}.$$

A sufficient condition for the above to hold is that

$$n \ge 4\frac{2d}{(\delta\epsilon)^2}\log(2d/(\delta\epsilon)^2) + \frac{4d\log(2e/d)}{(\delta\epsilon)^2}.$$

3 Proof of Uniform Convergence

To show Theorem 2.1 we will show that

$$\mathbb{E}_{Z^n\sim P^n}\left[\sup_{h\in\mathcal{H}} |L(h,P) - L(h,P_{Z^n})|\right] \le \frac{4+\sqrt{\log(\tau_{\mathcal{F}}(2n))}}{\sqrt{2n}}, \tag{2}$$

which will then imply (1) via Markov inequality. To show (2), we will apply a two-sample trick so that we can restrict $\mathcal{H}$ to some C, forming an small effective size hypothesis class $\mathcal{H}_C$, and then apply the union bound over $\mathcal{H}_C$.

3.1 Two-Sample Trick

To bound the L.H.S. of (2), we will use the two-sample trick. First note that for every $h \in \mathcal{H}$, we can rewrite

$$L(h, P) = \mathbb{E}_{\tilde{Z}^n \sim P^n}[L(f, P_{\tilde{Z}^n})],$$

where $\tilde{Z}^n$ is an additional i.i.d. sample. Therefore,

$$\begin{aligned}
\mathbb{E}_{Z^n \sim P^n}\left[\sup_{h \in \mathcal{H}} |L(h, P) - L(h, P_{Z^n})|\right] &= \mathbb{E}_{Z^n \sim P^n}\left[\sup_{h \in \mathcal{H}} |\mathbb{E}_{\tilde{Z}^n \sim P^n}[L(h, P_{\tilde{Z}^n}) - L(h, P_{Z^n})]|\right] \\
&\leq \mathbb{E}_{Z^n \sim P^n}\left[\sup_{h \in \mathcal{H}} \mathbb{E}_{\tilde{Z}^n \sim P^n}|L(h, P_{\tilde{Z}^n}) - L(h, P_{Z^n})|\right] \\
&\leq \mathbb{E}_{Z^n \sim P^n}\mathbb{E}_{\tilde{Z}^n \sim P^n}\left[\sup_{h \in \mathcal{H}} |L(h, P_{\tilde{Z}^n}) - L(h, P_{Z^n})|\right] \\
&= \mathbb{E}_{Z^n, \tilde{Z}^n \sim P^n}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n (\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))\right|\right]. \qquad (3)
\end{aligned}$$

Since $(Z^n, \tilde{Z}^n)$ are chosen i.i.d., nothing will change if we swap Z_i and $\tilde{Z}_i$ in (3); if we do so, instead of the term $(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))$ we will have $-(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))$ in (3). Therefore, for every $v^n \in \{\pm 1\}^n$ we have that the R.H.S. of (3) equals

$$\mathbb{E}_{Z^n, \tilde{Z}^n \sim P^n}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n v_i(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))\right|\right].$$

Since this holds for every $v^n \in \{\pm 1\}^n$, it also holds if we sample each component of v^n according to the uniform distribution U on $\{\pm 1\}$. Hence the R.H.S. of (3) also equals

$$\begin{aligned}
&\mathbb{E}_{V^n \sim U^n}\mathbb{E}_{Z^n, \tilde{Z}^n \sim P^n}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n V_i(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))\right|\right] \\
&= \mathbb{E}_{Z^n, \tilde{Z}^n \sim P^n}\mathbb{E}_{V^n \sim U^n}\left[\sup_{h \in \mathcal{H}} \frac{1}{n}\left|\sum_{i=1}^n V_i(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))\right|\right]. \qquad (4)
\end{aligned}$$

3.2 Restrict $\mathcal{H}$ to C

We now show that the inner expectation in (4) can be upper bounded irrespective of Z^n and $\tilde{Z}^n$. For any Z^n and $\tilde{Z}^n$, let $C(Z^n, \tilde{Z}^n)$ be the instances appearing in Z^n and $\tilde{Z}^n$. Then we can take the supremum in (4) only over $h \in \mathcal{H}_{C(Z^n, \tilde{Z}^n)}$, i.e.,

$$\text{inner expectation of (4)} = \mathbb{E}_{V^n \sim U^n}\left[\max_{h \in \mathcal{H}_{C(Z^n, \tilde{Z}^n)}} \frac{1}{n}\left|\sum_{i=1}^n V_i(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i))\right|\right].$$

For any h and $i \in [1:n]$, let $W_{h,i}$ be defined as

$$W_{h,i} = V_i(\ell(h(\tilde{X}_i), \tilde{Y}_i) - \ell(h(X_i), Y_i)).$$

Clearly, $\{W_{h,i}\}_{i=1}^n$ are a sequence of independent random variables, each of which takes values in $[-1, 1]$ and has mean 0. Therefore, we have by Hoeffding's inequality that for any h and any $\rho > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n W_{h,i}\right| \geq \rho\right) \leq 2e^{-2n\rho^2}.$$

Applying the union bound over $h \in \mathcal{H}_{C(Z^n,\tilde{Z}^n)}$, we have for any $\rho > 0$,

$$\mathbb{P}\left(\max_{h\in\mathcal{H}_{C(Z^n,\tilde{Z}^n)}} \frac{1}{n}\left|\sum_{i=1}^n W_{h,i}\right| \geq \rho\right) \leq 2|\mathcal{H}_{C(Z^n,\tilde{Z}^n)}|e^{-2n\rho^2},$$

which, via some technical lemma, implies that

$$\begin{aligned}\mathbb{E}\left[\max_{h\in\mathcal{H}_{C(Z^n,\tilde{Z}^n)}} \frac{1}{n}\left|\sum_{i=1}^n W_{h,i}\right|\right] &\leq \frac{4+\sqrt{\log(|\mathcal{H}_{C(Z^n,\tilde{Z}^n)}|)}}{\sqrt{2n}} \\ &\leq \frac{4+\sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\sqrt{2n}}.\end{aligned}$$

Plugging this back into (4), we have proved (2).

4 Quantitative Version of Fundamental Theorem of Learning

Finally, we provide a stronger, quantitative version of the fundamental theorem of statistical learning.

Theorem 4.1 (Fundamental Theorem of Statistical Learning – Quantitative Version) *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0-1 loss. Then there exists C_1, C_2 such that*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity satisfying*

$$C_1\frac{d+\log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}^{UC}(\epsilon,\delta) \leq C_2\frac{d+\log(1/\delta)}{\epsilon^2};$$

2. *$\mathcal{H}$ is PAC learnable with sample complexity satisfying*

$$C_1\frac{d+\log(1/\delta)}{\epsilon^2} \leq n_{\mathcal{H}}(\epsilon,\delta) \leq C_2\frac{d+\log(1/\delta)}{\epsilon^2}.$$