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

Lecture 5: VC Dimension

Lecturer: Xiugang Wu *03/07/2019*

Last time, we decomposed the excess error $L(h_{Z^n}, P) - \min_h L(h, P)$ of the $\text{ERM}_{\mathcal{H}}$ rule into the approximation error

$$L_{\text{app}} = \min_{h \in \mathcal{H}} L(h, P) - \min_{h} L(h, P)$$

and estimation error

$$L_{\text{est}} = L(h_{Z^n}, P) - \min_{h \in \mathcal{H}} L(h, P).$$

The approximation error depends on the fit of our prior knowledge to the underlying unknown distribution P. In contrast, the definition of PAC learnability requires that the estimation error would be bounded uniformly over all distributions.

This lecture will discuss which classes $\mathcal{H}$ are PAC learnable and how to characterize the sample complexity of learning $\mathcal{H}$. We will first show that even infinite classes can be learnable, and thus, finiteness of the hypothesis class is not a necessary condition for learnability. We then introduce the notion of Vapnik-Chervonenkis dimension (VC-dimension), which plays a key role in characterizing the learnability of $\mathcal{H}$. Indeed, we will finally prove the fundamental theorem of statistical learning theory, which states the sufficient and necessary condition for $\mathcal{H}$ to be PAC learnable –i.e. its VC-dimension VC-d($\mathcal{H}$) is finite– and determines the sample complexity of learning $\mathcal{H}$ in terms of VC-d($\mathcal{H}$).

1 Infinite-Size Classes Can Be Learnable

In previous lectures, we have shown that finite classes are learnable and the sample complexity of a finite hypothesis class is upper bounded by the log of its size. But what if the class is of infinite size? The next example shows that even an infinite class can be learnable, and thus the size of the hypothesis class is not the right characterization of the its learnability and sample complexity.

Example 1.1 *Let $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{H}$ be the set of threshold functions over $\mathbb{R}$, i.e.*

$$\mathcal{H} \triangleq \{h_a : a \in \mathbb{R}\}$$

where h_a is the threshold function such that $h_a(x) = \mathbb{I}(x > a)$. Suppose the realizability assumption holds, i.e. assume there exists some $h^*(x) = \mathbb{I}(x > a^*)$ in $\mathcal{H}$ such that $L(h^*, P) = 0$. We can show that $\mathcal{H}$ is PAC learnable using the ERM rule, with sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq \log(2/\delta)/\epsilon$.

Proof: For arbitrary underlying distribution P, let $a_0 < a^* < a_1$ be such that

$$P(X \in (a_0, a^*)) = P(X \in (a^*, a_1)) = \epsilon.$$

Obviously, for any predictor h_b with $b \in [a_0, a_1]$, we have

$$L(h_b, P) \leq \epsilon,$$

and for any predictor h_b with $b \notin [a_0, a_1]$, we have

$$L(h_b, P) > \epsilon.$$

For an ERM learner to output some h_b with $b \notin [a_0, a_1]$, the training sample Z^n must be such that

$$b_0(Z^n) < a_0 \text{ or } b_1(Z^n) > a_1,$$

where $b_0(Z^n)$ and $b_1(Z^n)$ are defined as the largest and smallest x appearing in Z^n with its label being 0 and 1, respectively, i.e.,

$$\begin{aligned} b_0(Z^n) &= \max\{x : (x, 0) \in Z^n\} \\ b_1(Z^n) &= \min\{x : (x, 1) \in Z^n\}. \end{aligned}$$

Therefore, for any $\epsilon > 0$ we can bound the probability of failure of the ERM learner as follows:

$$\begin{aligned} P^n(L(h_{Z^n}, P) > \epsilon) &\leq P^n(b_0(Z^n) < a_0 \text{ or } b_1(Z^n) > a_1) \\ &\leq P^n(b_0(Z^n) < a_0) + P^n(b_1(Z^n) > a_1) \\ &= P^n(X_i \notin [a_0, a^*], \forall i \in [1:n]) + P^n(X_i \notin [a^*, a_1], \forall i \in [1:n]) \\ &= 2(1-\epsilon)^n \\ &\leq 2e^{-n\epsilon}. \end{aligned}$$

Therefore, the class $\mathcal{H}$ is PAC learnable, with sample complexity $n_{\mathcal{H}}(\epsilon, \delta) \leq \log(2/\delta)/\epsilon$. □

2 The VC Dimension

From the above example, we see that while finiteness of $\mathcal{H}$ is a sufficient condition for learnability, it is not necessary. Indeed, as we will see, the correct characterization of the learnability of a hypothesis class $\mathcal{H}$ is given by a property, called the VC dimension, of $\mathcal{H}$.

2.1 Restriction of $\mathcal{H}$ to C

Definition 2.1 (Restriction of $\mathcal{H}$ to C) *Let $\mathcal{H}$ be a hypothesis class and let $C = \{c_1, c_2, \ldots, c_m\} \subseteq \mathcal{X}$. The restriction of $\mathcal{H}$ to C is the set of functions from C to $\{0, 1\}$ that can be derived from $\mathcal{H}$. That is,*

$$\mathcal{H}_C = \{(h(c_1), \ldots, h(c_m)) : h \in \mathcal{H}\}$$

where we represent each function from C to $\{0, 1\}$ as a vector in $\{0, 1\}^{|C|}$.

2.2 Shattering

If the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$, we say $\mathcal{H}$ shatters the set C.

Definition 2.2 (Shattering) *A hypothesis class $\mathcal{H}$ shatters a finite set $C \subseteq \mathcal{X}$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0, 1\}$, i.e. $|\mathcal{H}_C| = 2^{|C|}$.*

In light of the proof of the NFL theorem, we see that whenever some set C is shattered by $\mathcal{H}$, the adversary is not restricted by $\mathcal{H}$ on this set C. The corollary below immediately follows.

Corollary 2.1 *Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{0, 1\}$. Let n be the training sample size. Assume that there exists a set $C \subseteq \mathcal{X}$ of size $2n$ that is shattered by $\mathcal{H}$. Then, for any learning algorithm A there exists a distribution P over $\mathcal{X} \times \{0, 1\}$ such that $L(h, P) = 0$ for some predictor $h \in \mathcal{H}$ but $L(A(Z^n), P) \geq 1/8$ with probability of at least 1/7.*

2.3 VC Dimension

In view of the above corollary, the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$ is an important property of $\mathcal{H}$; indeed, this is defined to be the VC dimension of $\mathcal{H}$.

Definition 2.3 (VC-dimension) *The VC-dimesion of a hypothesis class $\mathcal{H}$, denoted by VC-d($\mathcal{H}$), is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrarily large size we say that $\mathcal{H}$ has infinite VC-dimension.*

With the definition of the VC dimension, we immediately have the following theorem.

Theorem 2.1 *If a hypothesis class $\mathcal{H}$ is of infinite VC dimension, then $\mathcal{H}$ is not PAC learnable.*

In the next lecture, we will show that the converse to the above theorem is also true, i.e. a finite VC dimension guarantees learnability. Therefore, the VC dimension characterizes PAC learnability. But before we delve into more theory in the next time, let us conclude today's lecture with several examples on how to calculate the VC dimension.

2.4 Examples

Remember that to show VC-d($\mathcal{H}$) $= d$, we need to show that:

- There exists a set C of size d that is shattered by $\mathcal{H}$;
- Every set C of size $d + 1$ cannot be shattered by $\mathcal{H}$.

Threshold Functions. Let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ be the class of threshold functions over $\mathbb{R}$, where $h_a(x) = \mathbb{I}(x > a)$. For any $C = \{c_1\}$, $\mathcal{H}$ shatters C; therefore VC-d($\mathcal{H}$) ≥ 1. For any $C = \{c_1, c_2\}$, $\mathcal{H}$ does not shatter C. Therefore, VC-d($\mathcal{H}$) $= 1$.

Intervals. Let $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b\}$ be the class of intervals over $\mathbb{R}$, where $h_{a,b}(x) = \mathbb{I}(x \in (a, b))$. For any $C = \{c_1, c_2\}$ with $c_1 < c_2$, $\mathcal{H}$ shatters C; therefore VC-d($\mathcal{H}$) ≥ 2. For any $C = \{c_1, c_2, c_3\}$ with $c_1 \leq c_2 \leq c_3$, the labelling $(1, 0, 1)$ cannot be obtained by any interval $h_{a,b}$ and therefore $\mathcal{H}$ does not shatter C. We can then conclude that VC-d($\mathcal{H}$) $= 2$.

Axis Aligned Rectangles. Let $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}^2\}$ be the class of axis aligned rectangles over $\mathbb{R}^2$, where $h_{a,b}(x) = \mathbb{I}(x_1 \in [a_1, b_1], x_2 \in [a_2, b_2])$. We can find $C \subseteq \mathbb{R}^2$ with 4 points that can be shattered by $\mathcal{H}$; therefore VC-d($\mathcal{H}$) ≥ 4. For any $C \subseteq \mathbb{R}^2$ with 5 points, the labelling $(1, 1, 1, 1, 0)$, where the "0" is for the innermost point, cannot be obtained by any axis aligned rectangle $h_{a,b}$ and therefore $\mathcal{H}$ does not shatter C. We can then conclude that VC-d($\mathcal{H}$) $= 4$.

Finite Classes. Let $\mathcal{H}$ be a finite class of functions over $\mathcal{X}$. For any $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$, we have $2^{|C|} = |\mathcal{H}_C| \leq |\mathcal{H}|$ and hence VC-d($\mathcal{H}$) $\leq \log |\mathcal{H}|$. Note, however, that VC-d($\mathcal{H}$) can be significantly smaller than $\log |\mathcal{H}|$. For example, let $\mathcal{X} = [1 : k]$ for some integer k. Consider the class of threshold functions $\mathcal{H} = \{h_a : a \in [1/2 : k + 1/2]\}$ over $\mathcal{X}$, where $h_a(x) = \mathbb{I}(x > a)$. Then VC-d($\mathcal{H}$) $= 1$ but $\log |\mathcal{H}| = \log(k + 1)$.

Convex Sets. Let $\mathcal{H} = \{h_A : A \subseteq \mathbb{R}^2, A \text{ is convex}\}$ be the class of convex sets over $\mathbb{R}^2$, where $h_A(x) = \mathbb{I}(x \in A)$. For any integer k, we can find a set $C \subseteq \mathbb{R}^2$ with $|C| = k$ that can be shatter by $\mathcal{H}$ (can you see why?) and hence VC-d$(\mathcal{H}) = \infty$.