Lecture 4: The Vapnik-Chervonenkis Dimension

Akshay Krishnamurthy
akshay@cs.umass.edu

September 21, 2017

1 Recap

So far we've seen two main sample complexity bounds for different learning settings. In the original (realizable) PAC model we saw an $O(\log(|\mathcal{H}|/\delta)/\epsilon)$ sample complexity bound, and for the agnostic PAC model we saw an $O(\log(|\mathcal{H}/\delta)/\epsilon^2)$ bound. Both bounds apply for finite $\mathcal{H}$. For the latter, we used Hoeffding's inequality which states that for iid random variables X_i with mean μ and range $X_i \in [a,b]$

$$\mathbb{P}(|\bar{X} - \mu| > \epsilon) \leq 2\exp(-2n\epsilon^2/(b-a)^2).$$

The key limitation of the two results above is that they require finite $\mathcal{H}$. While we can use the discretization trick from last lecture, it is somehow unsatisfying because it depends on the implementation. Today we'll see how to move beyond finite $\mathcal{H}$.

2 Shattering and VC Dimension

Let's start more pessimistically and think about when we should not be able to learn. In the proof of the No Free Lunch Theorem, we ran into some trouble because the set of T labeling functions $f_1, \ldots, f_T$ could realize all possible labelings on a set of examples C. In the proof, we showed that because this happened, if we didn't see a large fraction of the examples from C we were forced to incur significant error. This concept of a hypothesis class being able to realize all labelings on a set of examples will be important going forward, and is known as **shattering**.

Definition 1 (Restriction and Shattering)**.** *The **restriction** of a class $\mathcal{H}$ to a set of examples $C = \{c_1, \ldots, c_n\} \in \mathcal{X}$ is a subset of $\{0,1\}^{|C|}$ given by $\mathcal{H}_C = \{(h(c_1), \ldots, h(c_n)) \mid h \in \mathcal{H}\}$. We say that $\mathcal{H}$ **shatters** C if $|\mathcal{H}_C| = 2^{|C|}$.*

$\mathcal{H}$ shatters C precisely when all the labelings can be realized by the functions in $\mathcal{H}$. Going back to the No Free Lunch Theorem, this is precisely how we showed that learning was not possible.

Corollary 2 (No Free Lunch)**.** *Let $\mathcal{H}$ be a hypothesis class and assume there exists a set $C \subset \mathcal{X}$ of size $2n$ such that $\mathcal{H}$ shatters C. Then there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ and a predictor $h \in \mathcal{H}$ such that $R(h) = 0$ but for any learning algorithm A, $\mathbb{P}_{S\sim\mathcal{D}^n}[R(A(S))] \geq 1/8] \geq 1/7$.*

While we haven't proven anything yet, a form of converse seems plausible. If C is such that $|\mathcal{H}_C| \ll 2^{|C|}$, then we might expect that we can actually learn for distributions supported just on C. By seeing some of the labeled data, you should be able to infer the labels elsewhere. In fact, for distributions supported just on C, the real hypothesis space in consideration is just $\mathcal{H}_C$, which is not only finite, but also quite small! This brings us to the definition of Vapnik-Chervonenkis (VC) dimension.

Definition 3 (VC Dimension)**.** *The VC-dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$) is the size of the largest set $C \subset \mathcal{X}$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitrary size, then VCdim($\mathcal{H}$) $= \infty$.*

The main result in this lecture is that hypothesis classes with finite VC dimension have the uniform convergence property, and hence they are agnostic PAC learnable. This reveals the importance of shattering: We saw from the no free lunch theorem that if large sets can be shattered then learning is not possible, and the VC theorem will say that if large sets cannot be shattered, then learning is possible.

First let us turn to some examples:

Example 1 (VC Dimension of threshold functions)**.** *Let $\mathcal{X} = [0,1]$ and let $\mathcal{H} = \{h_b(x) = \mathbf{1}\{x \geq b\}, b \in [0,1]\}$ be the set of threshold functions. It is easy to see that VCdim($\mathcal{H}$) $= 1$, since clearly the set $C = \{1/2\}$ can be labeled both positive and negative, but for a set $C = \{a, b\}$ with $a \leq b$ we cannot simultaneously label a positively and b negatively. If the thresholds can also be revereсed, e.g., $\mathbf{1}\{x < b\}$ then the VC-dimension is 2. (Exercise: generalize to linear thresholds in d dimensions, where the VC-dimension is $d + 1$.)*

Example 2 (VC dimension of axis-aligned rectangles)**.** *Let $\mathcal{X} = \mathbb{R}^2$ and consider $\mathcal{H} = \{h_{a_1,a_2,b_1,b_2}(x_1, x_2) = \mathbf{1}\{a_1 \leq x_1 \leq a_2 \wedge b_1 \leq x_2 \leq b_2\}\}$ which are the axis-aligned rectangles. It is easy to see that the set $C = \{(-1, 0), (1, 0), (0, -1), (0, 1)\}$ can be shattered by $\mathcal{H}$. Now consider any set $C = \{c_1, \ldots, c_5\}$ of 5 points. Take the four points "on the perimeter" (the one with smallest x_1, largest x_1, smallest x_2, and largest x_4, break ties arbitrarily) and consider the labeling that sets these four to 1 and the final point to 0. This labeling cannot be realized by $\mathcal{H}$, so the set cannot be shattered. Thus VCdim($\mathcal{H}$) $= 4$.*

Example 3 (VC dimension of finite classes)**.** *For any finite hypothesis class $\mathcal{H}$, we must have VCdim($\mathcal{H}$) $\leq \log_2(|\mathcal{H}|)$, since for any set C we must get $\mathcal{H}_C \leq |\mathcal{H}|$ (If $2^{|C|} > |\mathcal{H}|$ then we cannot shatter C). Thus PAC learnability for finite classes follows from learnability for VC classes.*

As a preview of what to expect and to build some intuition, if VCdim($\mathcal{H}$) $= d$ then there exists a set of d examples that can be shattered. If the $\mathcal{X}$-marginal distribution was uniform over these d examples, then, since they can be shattered, we are back to the finite hypothesis space situation with 2^d hypotheses. Since in the finite-hypothesis sample complexity bounds we have logarithmic dependence on the number of hypotheses, this suggests that we should expect a linear dependence on the VC-dimension. This is exactly what we'll prove.

3 VC Theorem

Theorem 4 (Finite VC classes have uniform convergence property)**.** *Let $\mathcal{H}$ be a hypothesis class with VCdim($\mathcal{H}$) $= d < \infty$. Then there is an absolute constant $c > 0$ such that $\mathcal{H}$ has the uniform convergence property with*

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

1. This immediately implies that $\mathcal{H}$ is agnostically-PAC learnable with $O(\frac{d\log(d/\epsilon)+\log(1/\delta)}{\epsilon^2})$ sample complexity.

2. It turns out that $\mathcal{H}$ is also (realizable) PAC-learnable with sample complexity $O(\frac{d\log(d/\epsilon)+\log(1/\delta)}{\epsilon})$, although it is not implied directly by Theorem 4.

3. In fact slightly better bounds are possible for both agnostic and realizable cases, but they will require more advanced tools. We might get to them in the next couple of lectures.

4. The result as stated applies only for binary classification with 0/1 loss. We'll see how to move beyond binary classification and to other loss functions in upcoming lectures.

Proof. The key insight in the proof is that if $\mathcal{H}$ has finite VC-dimension, then it actually looks like a finite-sized class once n is large enough. Specifically, in the first step, we'll show that for any finite set $C \subset \mathcal{X}$, the effective size of $\mathcal{H}$, which is $|\mathcal{H}_C|$, is actually $O(|C|^d)$ which grows only polynomially in $|C|$ rather than exponentially. Then in the second step we'll show that small effective size is adequate for us to apply the union bound, which is similar to what we did in the finite-$|\mathcal{H}|$ case.

Step 1: Polynomial growth of $\mathcal{H}_C$. We'll define the growth function to quantify the size of $\mathcal{H}_C$.

Definition 5 (Growth function)**.** *The growth function of $\mathcal{H}$ is $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ is defined as*

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

One key point to the proof is to understand how $\tau_{\mathcal{H}}(n)$ depends on n. If VCdim($\mathcal{H}$) $= d$ then clearly $\tau_{\mathcal{H}}(n) = 2^n$ for all $n \leq d$. The next lemma gives an upper bound on $\tau_{\mathcal{H}}$ for $n \geq d$ (technically for all n).

Lemma 6 (Sauer-Shelah Lemma)**.**

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

In particular, for $n > d+1$, this implies $\tau_{\mathcal{H}}(n) \leq (en/d)^d$.

Proof. We instead show a stronger inequality: For any $C = \{c_1, \dots, c_n\}$ and any $\mathcal{H}$

$$|\mathcal{H}_C| \leq |\{B \subset C : \mathcal{H} \text{ shatters } B\}|. \tag{1}$$

This statement is sufficient here since if $\text{VCdim}(\mathcal{H}) = d$ then $\mathcal{H}$ cannot shatter any subset B of size $|B| > d$. Since there are $\binom{n}{i}$ subsets of size i, this implies the bound on the growth function.

To prove (1), we proceed by induction. For $n = 1$ either both sides equal 1 or both sides equal 2 since either there is only one available labeling, in which case you shatter the empty set, or both labelings. Now inductively assume that (1) holds for all sets of size $k < n$. Fix $\mathcal{H}$ and $C = \{c_1, \dots, c_n\}$ and let $C' = \{c_2, \dots, c_n\}$.

We define two sets $Y_0 = \{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_C \vee (1, y_2, \dots, y_n) \in \mathcal{H}_C\}$ and $Y_1 = \{(y_2, \dots, y_n) : (0, y_2, \dots, y_n) \in \mathcal{H}_C \wedge (1, y_2, \dots, y_n) \in \mathcal{H}_C\}$. These definitions imply that $|\mathcal{H}_C| = |Y_0| + |Y_1|$ (While in Y_1 we only count each sequence once, the other time is included in Y_0).

Now by induction since $Y_0 = \mathcal{H}_{C'}$ we get

$$|Y_0| \leq |\{B \subset C' : \mathcal{H} \text{ shatters } B\}| = |\{B \subset C : c_1 \notin B \wedge \mathcal{H} \text{ shatters } B\}|$$

For Y_1, define another hypothesis class:

$$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H} \text{ s.t. } (1 - h'(c_1), h'(c_2), \dots, h'(c_n)) = (h(c_1), \dots, h(c_n))\}$$

which contains all the hypotheses for which the hypothesis that agrees everywhere on C except for c_1 is also in $\mathcal{H}$. It is then clear that if $\mathcal{H}'$ shatters some set $B \subset C'$ then it also shatters $B \cup \{c_1\}$. Moreover working through the definitions $Y_1 = \mathcal{H}'_{C'}$ and hence

$$|Y_1| = |\mathcal{H}'_{C'}| \leq \{B \subset C' : \mathcal{H}' \text{ shatters } B\}| = |\{B \subset C' : \mathcal{H}' \text{ shatters } B \cup \{c_1\}\}| \quad \leq |\{B \subset C : \mathcal{H} \text{ shatters } B \cup \{c_1\}\}|$$

Adding up the bounds for Y_0 and Y_1 concludes the proof. The approximation $(en/d)^d$ is somewhat technical, but using the fact that $\binom{n}{d} \leq (en/d)^d$ the asymptotics shouldn't be too surprising. □

Step 2. Symmetrization The next step is to show that if the growth function is well behaved, we obtain uniform convergence.

Lemma 7. *For a class $\mathcal{H}$ with growth function $\tau_{\mathcal{H}}$,*

$$\mathbb{E}_{S \sim \mathcal{D}^n} \sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)| \leq \sqrt{\frac{2 \log(2\tau_{\mathcal{H}}(2n))}{n}}.$$

Armed with the lemma, we can apply Markov's inequality to get a high-probability uniform convergence statement, but we can significantly improve the dependence on δ using McDiarmid's inequality, which we saw previously. We'll return to this after the proof, which contains the important concept of symmetrization.

Proof.

$$\begin{aligned}
\mathbb{E}_S \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| &= \mathbb{E}_S \sup_{h \in \mathcal{H}} |\mathbb{E}_{S'} \hat{R}_{S'}(h) - \hat{R}_S(h)| \leq \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} |\hat{R}_{S'}(h) - \hat{R}_S(h)| \\
&= \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{h(x_i') \neq y_i'\} - \mathbf{1}\{h(x_i) \neq y_i\} \right| \\
&= \mathbb{E}_{\epsilon_{1:n}} \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (\mathbf{1}\{h(x_i') \neq y_i'\} - \mathbf{1}\{h(x_i) \neq y_i\}) \right|
\end{aligned}$$

The first inequality here follows from Jensen's inequality, since $|\cdot|$ is convex and so is the maximum of convex functions (or the maximum of expectation is smaller than expected maximum). In the third line we simply write out the definition of empirical risk, and in the fourth line we introduce the **rademacher** random variables $\epsilon_1, \ldots, \epsilon_n \sim \mathrm{Unif}(\{-1,1\})$, using the fact that (x_i, y_i) and (x_i', y_i') are iid.

Now fix S, S' and let C be the examples appearing in S, S', we know that $|C| \leq 2n$ and we can replace the supremum by the supremum over $\mathcal{H}_C$

$$\mathbb{E}_{\epsilon_{1:n}} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (\mathbf{1}\{h(x_i' \neq y_i'\} - \mathbf{1}\{h(x_i) \neq y_i\}) \right| = \mathbb{E}_{\epsilon_{1:n}} \max_{h \in \mathcal{H}_C} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i (\mathbf{1}\{h(x_i' \neq y_i'\} - \mathbf{1}\{h(x_i) \neq y_i\}) \right|$$

Now observe that we have just $\tau_{\mathcal{H}}(2n)$ terms in the maximum, but actually each term is itself centered and bounded, so we at a high-level we can do the same thing we did in the proof of Hoeffding's inequality. The only difference is that we are working on the expectation instead of in high probability. Nevertheless, if we call $\theta_h = \frac{1}{n}\sum_{i=1}^n \epsilon_i(\ell(h, z_i') - \ell(h, z_i))$ as the term in the sum, we get

$$\begin{aligned}
\mathbb{E}_\epsilon \max_{h \in \mathcal{H}_C} |\theta_h| &= \frac{1}{\lambda} \log \exp \lambda \mathbb{E}_\epsilon \max_{h \in \mathcal{H}_C} \max\{\theta_h, -\theta_h\} \leq \frac{1}{\lambda} \log \mathbb{E}_\epsilon \exp\left(\lambda \max_{h \in \mathcal{H}_c} \max\{\theta_h, -\theta_h\}\right) \\
&\leq \frac{1}{\lambda} \log \sum_{h \in \mathcal{H}_C} 2\mathbb{E}_\epsilon \exp(\lambda \theta_h) = \frac{1}{\lambda} \log \sum_{h \in \mathcal{H}_C} 2(\mathbb{E}_\epsilon \exp(\lambda \epsilon / n(\ell(h,z) - \ell(h,z'))))^n \\
&\leq \frac{1}{\lambda} \log \sum_{h \in \mathcal{H}_C} 2\exp\left(\lambda^2/(2n)\right).
\end{aligned}$$

This derivation applies for any λ, and if we set $\lambda = \sqrt{2n \log(2|\mathcal{H}_C|)}$ we get

$$\mathbb{E}[\max_{h \in \mathcal{H}_C} |\theta_h|] \leq \sqrt{\frac{2\log(2|\mathcal{H}_C|)}{n}}.$$

Using the definition of the growth function, this proves that

$$\mathbb{E}_{S \sim \mathcal{D}^n} \sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)| \leq \sqrt{\frac{2\log(2\tau_{\mathcal{H}}(2n))}{n}}$$

Note that the constants here are different than from those in SSBD. □

Step 3. Applying McDiarmid's inequality. The last step in the proof is to take the bound on the expectation and apply McDiarmid's inequality to obtain exponential concentration. Observe that the functional

$$f(S) = \sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)|$$

satisfies the bounded differences property with constant $1/n$, since each term can only move by at most $1/n$ when we swap out a sample, so the maximum also can only move by at most $1/n$, thus McDiarmid's inequality states

$$\mathbb{P}(|f(S) - \mathbb{E}f(S)| > t) \leq 2\exp(-2nt^2).$$

Setting the RHS to be at most δ, using the upper bound on the expected maximum and re-arranging shows that with probability at least $1 - \delta$

$$\sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\frac{2\log(2\tau_{\mathcal{H}}(2n))}{n}}.$$

Step 4: Putting things together. Finally, using the bound on $\tau_{\mathcal{H}}$ from Sauer-Shelah, we get

$$\sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}} + \sqrt{\frac{2d\log(2en/d)}{n}}.$$

for $n \geq O(\frac{d\log(d/\epsilon) + \log(1/\delta)}{\epsilon^2})$, this is at most ϵ. Remember that Sauer-Shelah required $n \geq d+1$ to use the approximation, but if $n \leq d+1$ then this bound is trivial since the LHS is always at most 1. □