Lecture #5: VC Dimension and Learning

Instructor: Aditya Bhaskara *Scribe: Asmaa Al-Juhani*

CS 5966/6966: Theory of Machine Learning

January 25th, 2017

Abstract

In this lecture, we review the notion of the growth function of a hypothesis class, and introduce *VC dimension* – a fundamental notion in learning theory. We then prove a powerful lemma that connects the VC dimension to the growth function.

1 Review and Introduction

Let us recall some of the notation we have used over the last few lectures. We started to look at infinite hypothesis classes $\mathcal{H}$ and we are trying to understand the question of when a hypothesis class is PAC learnable. In the last lecture, we introduced the notion of the *growth function* (denoted $\tau_{\mathcal{H}}(m)$ of a class $\mathcal{H}$. As a consequence of the no free lunch theorem, we saw that if the growth function is exponential, then the class $\mathcal{H}$ is not PAC learnable, and vice versa. Thus to understand the learnability of a class, we simply need to know how $\tau_{\mathcal{H}}(m)$ grows.

Recall: the growth function is the max number of sign patterns that $\mathcal{H}$ induces on a set of size m.

2 Analysing the Growth function

While the characterization is elegant, reasoning about the growth function is often not easy unless the hypothesis class is very simple.

Consider an extremely simple class, e.g., half planes in $\mathbb{R}^2$. To compute the growth function, we need to argue that only lines that pass through two of the points "matter" (a fact that takes some reasoning), and then conclude that $\tau_{\mathcal{H}}(m)$ grows as $\Theta(m^2)$.

insert image - You have planes like shown and everything on one side is labeled (plus) and the other half is labeled (minus).

If we consider a $\mathcal{H}$ that is moderately more involved, e.g., the interiors of rectangles (not necessarily axis aligned), it becomes intricate to reason about the max number of sign patterns the class can induce on a set of m points (as we need to obtain a bound that works for *all* sets of m points). Is there a notion that makes it easier to argue about learnability?

3 Shattering, VC Dimension

Let us start with the simple observation, that the *largest possible* number of sign patterns that $\mathcal{H}$ can induce on a set S of size m is at most 2^m. I.e., $\sigma_{\mathcal{H}}(S) \leq 2^{|S|}$. A set S for which *equality* is achieved above is said to be *shattered* by $\mathcal{H}$. I.e.,

3.1 definition. $\mathcal{H}$ is said to "shatter" a point set S if $\sigma_{\mathcal{H}}(S) = 2^{|S|}$.

Next, we define the VC dimension – named after the authors who introduced it – Vapnik and Chervonenkis.

3.2 definition (VC dimension). The VC dimension of a class $\mathcal{H}$ is the size of the largest S it shatters.

Any non-empty class trivially shatters a set of size 0, thus the VC dimension is non-negative. Also, the VC dimension is equal to zero iff $\mathcal{H}$ has precisely one hypothesis – a constant function.

3.3 example. Let $\mathcal{H}$ be the set of all half spaces in the two-dimensional plane. What is its VC dimension?

We want the largest m for which there is a set of m points that is shattered. It is clear that $\mathcal{H}$ can shatter a set of 3 points, so the VC dimension is ≥ 3. Could there *exist* a set of 4 points that are shattered?

This illustrates the difficulty in reasoning about the VC dimension – while it is often easy to show that the VC dimension is at *least* a certain value (by exhibiting a set that is shattered), it is trickier to prove that the dimension is *less than* a certain value. To show this, we need to show that *no set* of that size can be shattered.

It is easy to lower bound the VC dimension and harder to upper bound it.

In the case of lines, this can be done by observing that any set of 4 points in the plane are in a "convex" configuration (Fig 2a) or as a triangle with one point in the interior (Fig 2b). In either case, we can show that it is impossible to attain all possible sign patterns using half spaces. Thus, in this example, the VC-dimension is 3.

3.4 example. Consider the example of convex polygons in a plane. (Insert image) I.e., the class $\mathcal{H}$ consists of all convex polygons in a 2D plane, with the interior labeled $+$ and the exterior $-$ (each $h \in \mathcal{H}$ gives a classifier for the plane). What is the largest set that can be shattered? If we take any K points that are all in convex position then if we pick any subset of these they can form a convex polygon and thus we can make precisely this subset $+$ and the rest $-$. We can also easily get the "all $-$" hypothesis. Since we can do this for any integer $K \geq 3$, the VC-dim of this $\mathcal{H}$ is ∞.

4 Relation to the Growth Function

We next prove a rather surprising theorem, relating the VC dimension with the growth function. It says that a hypothesis class of VC dimension d has its growth function $\tau_{\mathcal{H}}(m) = O(m^d)$. I.e., for any fixed d, it is polynomial in m. Formally,

4.1 theorem (Sauer, Shelah, Vapnik-Chervonenkis). *Let $\mathcal{H}$ be a hypothesis class of finite VC dimension d. Then for every m, we have*

$$\tau_{\mathcal{H}}(m) \leq \binom{m}{0} + \binom{m}{1} + ... + \binom{m}{d} [= O(m^d)] \tag{1}$$

This is a remarkable theorem, as it implies that "asymptotically", the growth function of any $\mathcal{H}$ is either 2^m, or it is m^d, for a fixed d. An easy consequence is the so-called fundamental theorem of statistical learning theory which we

will discuss now.

5 Fundamental theorem of stat learning theory

5.1 theorem. *Class $\mathcal{H}$ is PAC learnable $\iff$ $\mathcal{H}$ is agnostically PAC learnable $\iff$ $\mathcal{H}$ has a finite VC dimension.*

It basically says that the VC-dimension characterizes the notion of PAC learnability. The proof is via the following two claims.

1. If $\mathcal{H}$ has VC-dim ∞, then $\mathcal{H}$ is not PAC learnable.
2. If VC-dim is d, then $\mathcal{H}$ is agnostically PAC learnable.

This means the growth function is 2^m, and so by the "no-free-lunch" theorem we saw last class, it implies that $\mathcal{H}$ is not PAC learnable.

The second claim follows by first using Sauer's lemma to conclude that $\tau_{\mathcal{H}}(m) \leq O(m^d)$, and then appealing to the result we saw in the last lecture, namely

5.2 theorem. *Let $\mathcal{H}$ be a hypothesis class. Then for every $\mathcal{D}$, and every $\delta \in (0,1)$, we have that w.p. at least 1 - δ over the choice of S,*

$$\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_s(h)| \leq \frac{O\left(\sqrt{\log(\tau_{\mathcal{H}}(2m))}\right)}{\delta\sqrt{m}}. \tag{2}$$

If we use the fact that $\tau_{\mathcal{H}}(m) = O(m^d)$ in the above, the RHS simplifies to $\sqrt{\frac{d\log m}{m\delta^2}}$. Thus, for (ϵ,δ)-learning, we need $\approx d\text{poly}(1/\delta, 1/\epsilon)$ training samples.

The bound above gives a bound on the number of training samples we need to learn a hypothesis class. The moral is: small VC-dim implies that fewer training e.g.s suffice.

6 Proof of Theorem 4.1

Let us now prove Sauer's Lemma (stated earlier), as it is key to the proof of the Fundamental theorem.

Proof. Given a hypothesis class $\mathcal{H}$ of VC dimension d, we need to prove that for any set S of size $m \geq 1$,

$$\sigma_{\mathcal{H}}(S) \leq \phi(m,d), \quad \text{where } \phi(m,d) := \binom{m}{0} + \binom{m}{1} + ... + \binom{m}{d}. \tag{3}$$

The proof proceeds by induction on $m + d$. As base cases, we consider $d = 0$ (and arbitrary m) and $m = 1$ (and arbitrary d).

Note that $d = 0$ means that the class $\mathcal{H}$ has just one hypothesis, as we observed earlier. Thus in this case, for any S, $\sigma_{\mathcal{H}}(S) = 1 = \phi(m,0)$.

So also, if $m = 1$, (and assuming $d \neq 0$, as that case is covered above) we have $\sigma_{\mathcal{H}}(S) \leq 2$ (as there are only two sign patterns possible), and $\phi(1,d) = 2$.

Now, consider some $m > 1$ and $d > 0$. Suppose that we have $S = s_1, s_2, s_3, ...s_m$ and we want to analyze how many sign patters $\mathcal{H}$ could induce on S. First, we can omit one of the points and see how many sign pattern $\mathcal{H}$ could give

on the rest of the points. Let us define $\mathcal{P}_1$ as the set of sign patterns induced by $\mathcal{H}$ on $\{s_2, s_3, ..., s_m\}$.

By the inductive hypothesis (noting that we have $m-1$ elements, and the VC dimension of $\mathcal{H}$ is d), we have $|\mathcal{P}_1| \leq \phi(m-1, d)$.

Now let's define $\mathcal{P}_2$ as the set of sign patterns σ on $s_2, s_3, ..., s_m$ that have the property that *both* 0σ and 1σ are valid sign patterns on $\{s_1, s_2, s_3, ..., s_m\}$.

6.1 example. Suppose m=3, and suppose that the sign patterns induced by $\mathcal{H}$ on some set are $P = \{011, 111, 101, 010\}$.
$\mathcal{P}_1$: is the restriction on 2nd and 3rd bit. Thus, $\mathcal{P}_1 = \{11, 01, 10\}$.
$\mathcal{P}_2$: is the subset of $\mathcal{P}_1$ with the property that both 0σ and 1σ are patterns in P (e.g for $\sigma = 01$, we want to see if 001 and 101 are in P). Thus, $\mathcal{P}_2 = \{11\}$.

6.2 claim. $|\tau_{\mathcal{H}}(S)| = |\mathcal{P}_1| + |\mathcal{P}_2|$.

In our example $\mathcal{P}_1 = 3$ and $\mathcal{P}_2 = 1$, and their sum is $|P|$. In general, if we look at a sign pattern σ on $s_2, s_3, \ldots, s_m$, and ask how many patterns in P (defined as above) have σ as the restriction to the last $m-1$ coordinates. If the answer is two, it means that 0σ and 1σ are both in P (in which case σ appears once in $\mathcal{P}_1$ and once in $\mathcal{P}_2$), and if the answer is 1, then σ appears only in $\mathcal{P}_1$. Thus the claim follows.

Now, how can we bound the size of $\mathcal{P}_2$? Let us abuse notation slightly and also think of $\mathcal{P}_2$ as a set of hypotheses on the domain $s_2, \ldots, s_m$.

6.3 claim. *VC dimension of $\mathcal{P}_2$ is $\leq d-1$.*

In particular, it is not d. The reason is that if $\mathcal{P}_2$ could shatter a set $T \subset \{s_2, \ldots, s_m\}$, then $\mathcal{H}$ can shatter the set $T \cup \{s_1\}$ (this is by the definition of $\mathcal{P}_2$), and thus if $|T| = d$, then the VC dimension of $\mathcal{H}$ is $d+1$ — a contradiction.

Thus we have $|\mathcal{P}_2| \leq \phi(m-1, d-1)$.

This implies that

$$\tau_{\mathcal{H}}(S) \leq \phi(m-1, d) + \phi(m-1, d-1).$$

Now, using Pascal's identity (recall Binomial theorem from high school), we have $\binom{m-1}{k-1} + \binom{m-1}{k} = \binom{m}{k}$, and using this repeatedly, we get $\phi(m-1, d-1) + \phi(m-1, d) = \phi(m, d)$. Together with the bound on $\tau_{\mathcal{H}}(S)$ above, this completes the inductive proof. □