TTIC 31290: Machine Learning for Algorithm Design (Fall 2025)
Avrim Blum and Dravyansh Sharma

Lecture 7: 10/21/25 Lecturer: Dravyansh Sharma

Outline for today

- Pseudodimension Example
- Sauer's Lemma
- Sample complexity for infinite concept classes (double sampling argument)

1 Pseudodimension Example

Consider the set of functions $\mathcal{F} = \{f_{a,b}(x) = ax + b : a, b \in \mathbb{R}\}$ defined from $\mathbb{R} \to \mathbb{R}$. What is $\text{Pdim}(\mathcal{F})$?

To show $\text{Pdim}(\mathcal{F}) \geq 2$, we construct a set of points $x_1, x_2 \in \mathbb{R}$ which is pseudoshattered by $\mathcal{F}$. Let $x_1 = 1, x_2 = 2$. We set thresholds $r_1 = r_2 = 0$. For each above-below pattern, we have

- above r_1 on x_1, above r_2 on x_2. Choose $f_{1,0}$.
- above r_1 on x_1, below r_2 on x_2. Choose $f_{-1,\frac{3}{2}}$.
- below r_1 on x_1, above r_2 on x_2. Choose $f_{1,-\frac{3}{2}}$.
- below r_1 on x_1, below r_2 on x_2. Choose $f_{-1,0}$.

To show $\text{Pdim}(\mathcal{F}) \leq 2$, we show that three distinct points $x_1 < x_2 < x_3 \in \mathbb{R}$ cannot be pseudoshattered by $\mathcal{F}$. Let $r_1, r_2, r_3 \in \mathbb{R}$ be some thresholds.

- Consider the (above, below, above) pattern, achieved by f_{a_1,b_1} (say).
- Consider the (below, above, below) pattern, achieved, if possible, by f_{a_2,b_2}.

We will show that both the above actually cannot be achieved for any choice of thresholds $r_1, r_2, r_3 \in \mathbb{R}$.

Note that we have

$$\begin{aligned} f_{a_2,b_2}(x_1) < r_1 \leq f_{a_1,b_1}(x_1), \\ f_{a_2,b_2}(x_2) \geq r_2 > f_{a_1,b_1}(x_2), \\ f_{a_3,b_3}(x_3) < r_3 \leq f_{a_3,b_3}(x_3). \end{aligned}$$

Now $g(x) := f_{a_2,b_2}(x) - f_{a_1,b_1}(x)$ is a linear function in x. But we have $g(x_1), g(x_3) < 0$, while $g(x_2) > 0$ (a contradiction!)

2 Sauer's Lemma

To reason about generalization for infinite hypothesis classes, we use the number of distinct labelings realizable on a finite sample.

Definition 1 (Growth Function)**.** *For a hypothesis class H, define the* growth function

$$\Gamma_H(m) = \max_{S \subseteq \mathcal{X}, |S|=m} |\{(h(x_1), \ldots, h(x_m)) : h \in H\}|.$$

That is, $\Gamma_H(m)$ counts the maximum number of distinct labelings H can induce on m points.

The following result gives an upper bound on the growth function of function classes with a bounded VC dimension. By the definition of VC dimension, the growth function $\Gamma_H(m) = 2^m$ for any $m \leq d$, since H achieves all 2^m labelings on some set S of size d. However, this exponential growth with m is replaced by a polynomial growth (for fixed d) once we increase m to beyond d as we will see in the following well-known result.

Theorem 1 (Sauer's Lemma)**.** *Let $H \subseteq \{0,1\}^{\mathcal{X}}$ with* VCdim$(H) = d$*. Then for any integer $m \geq d$,*

$$\Gamma_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

In particular, $\Gamma_H(m) = O(m^d)$ when $m \geq d$.

Proof. We will proceed by induction on both m, d.

Base cases. (a) $d = 0$, any m. No set of points can be shattered, so all points can be labeled only in one way. $\Gamma_H(m) = 1 = \sum_{i=0}^{0} \binom{m}{i}$.

(b) $m = 0$, any d. RHS is 1, which bounds the number of labelings on a set of zero points.

Inductive step. $d > 0, m > 0$. Suppose the claim holds for all d', m' such that $d' + m' < d + m$. Let $S = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$. Let H_S be the set of hypotheses in H with domain restricted to S (these capture all the behaviors of H on S, in particular achieve $\Gamma_H(m)$ labelings of S).

Fix arbitrary $x \in S$, say x_m, and let $S' = S \setminus \{x_m\}$. Now we consider the different ways hypotheses in H_S label S'. For each distinct labeling of S' achieved by some function in H_S, we have either one or two distinct extensions to S (one if all such functions label x_m identically, two otherwise). Let H' be the subset of functions in H_S with a unique "representative" for each labeling of S' as follows: if there is a unique extension to S, we add that function to H', and if there are two extensions, we add the one that labels x_m as negative. Note that all the functions in $H_S \setminus H'$ label x_m as positive.

Now, we claim that VCDim$(H_S \setminus H') \leq d - 1$. Note that if $H_S \setminus H'$ shatters $T' \subseteq S'$, then H_S shatters $T' \cup \{x_m\}$. This is by construction of H' (Think why? For each $h \in H_S \setminus H'$ we have a "twin" in H' with the negative label for x_m. Also, $H_S \setminus H'$ can only label x_m positive.)

On the other hand, $H' \subseteq H_S$, and so VCDim$(H') \leq d$.

By the induction hypothesis:

$$\begin{aligned}
\Gamma_H(m) = |H_S| = |H'| + |H_S/H'| &\le \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\
&= \binom{m-1}{0} + \sum_{i=1}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) \\
&= \binom{m}{0} + \sum_{i=1}^{d} \binom{m}{i},
\end{aligned}$$

where we have used the Pascal identity $\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}$. This completes the induction. □

Sauer's Lemma shows that even though there may be infinitely many hypotheses, the number of distinct labelings on a finite sample grows only polynomially with m once VC dimension is bounded.

3 The Fundamental Theorem of Statistical Learning Theory

But why do we care about the growth function? Turns out that we can upper bound the sample complexity of agnostic PAC learning over H in terms of its growth function $\Gamma_H(m)$ (and by Sauer's lemma, therefore, also its VC dimension).

We first consider the realizable case (when the target concept c^* belongs to the class C).

Theorem 2. *Let C be an arbitrary concept space with VC dimension d. Then C is PAC learnable in the realizable setting with sample complexity*

$$m_C(\epsilon, \delta) = \frac{8}{\epsilon} \left(d \ln \left(\frac{16}{\epsilon} \right) + \ln \frac{2}{\delta} \right).$$

That is, for any D, any $\epsilon, \delta \in (0,1)$, and any target concept $c^* \in C$, given a sample of size at least $m_C(\epsilon, \delta)$ above, with probability at least $1-\delta$, all hypotheses h with error $\mathrm{err}_D(h) > \epsilon$ are inconsistent with the data.

Proof. We will first establish the following bound on the sample complexity in terms of the growth function:

$$m \ge \max \left\{ \frac{4}{\epsilon} \left(\ln \left(\Gamma_C(2m) \right) + \ln \frac{2}{\delta} \right), \frac{8}{\epsilon} \right\}$$

examples are sufficient for realizable PAC learning. We will then apply Sauer's lemma to get the above bound.

Define a "bad" event,
B_1: $\exists h \in C$ with $\mathrm{err}_S(h) = 0$ but $\mathrm{err}_D(h) > \epsilon$.

Suppose $S' \sim D^m$ is another sample ("ghost sample") drawn i.i.d. from D. Given B_1, the following event is likely for sufficiently large m,
B_2: $\exists h \in C$ with $\mathrm{err}_S(h) = 0$ but $\mathrm{err}_{S'}(h) > \epsilon/2$.

Lemma 1. *If $m \geq \frac{8}{\epsilon}$, then* $\Pr[B_1] < 2\Pr[B_2]$.

Proof. Given some $h \in C$ consistent with S but with $\mathrm{err}_D(h) > \epsilon$, by Chernoff's bounds, we have $\Pr[\mathrm{err}_{S'}(h) \leq \epsilon/2] \leq e^{-(1/2)^2 \cdot (m\epsilon/2)} < \frac{1}{2}$. Thus, $\Pr[B_2|B_1] < \frac{1}{2}$.

Or, $\Pr[B_2] \geq \Pr[B_2 \wedge B_1] = \Pr[B_2|B_1]\Pr[B_1] > \Pr[B_1]/2$. □

This introduction of S' is also called the "symmeterization trick". It allows us to focus on bounding $\Pr[B_2]$, where B_2 involves two finite samples.

Let's imagine a different but equivalent process generating S and S'. Suppose we draw a sample of size $2m$ from D, $U \sim D^{2m}$, and randomly partition it into two sets S, S' of size m each.

Now, for B_2 to happen, there must be some $h \in C_{|U}$ such that (a) for $M \geq m\epsilon/2$ examples in U, $h(x) \neq c^*(x)$, and (b) all these examples end up in S'. For any fixed $h \in C_{|U}$ that satisfies (a), (b) happens with probability at most $\binom{2m-M}{m-M}/\binom{2m}{m} \leq 2^{-M} \leq 2^{-m\epsilon/2}$. By a union bound,

$$\Pr[B_2] \leq \Gamma_C(2m)2^{-m\epsilon/2}.$$

Combined with Lemma 1, this gives us the bound on m in terms of the growth function. But both LHS and RHS depend on m!

By Sauer's Lemma, for $2m > d$,

$$\Gamma_C(2m) \leq \sum_{i=0}^{d} \binom{2m}{i} \leq \left(\frac{2em}{d}\right)^d.$$

We use the fact $\ln x \leq \alpha x - \ln \alpha - 1$ for all $\alpha, x > 0$ (the function $f(x) = \ln x - \alpha x$ is maximized at $x = \frac{1}{\alpha}$), to simplify the above bound as

$$\begin{aligned}
\frac{4}{\epsilon}\left(\ln\left(\Gamma_C(2m)\right) + \ln\frac{2}{\delta}\right) &\leq \frac{4}{\epsilon}\left(d\ln\left(\frac{2em}{d}\right) + \ln\frac{2}{\delta}\right) \\
&= \frac{4}{\epsilon}\left(d\ln m + d\ln\left(\frac{2e}{d}\right) + \ln\frac{2}{\delta}\right) \\
&\leq \frac{4}{\epsilon}\left(d\left(\frac{\epsilon}{8d}m - \ln\left(\frac{\epsilon}{8d}\right) - 1\right) + d\ln\left(\frac{2e}{d}\right) + \ln\frac{2}{\delta}\right) \\
&= \frac{m}{2} + \frac{4}{\epsilon}\left(d\ln\left(\frac{8d}{e\epsilon}\right) + d\ln\left(\frac{2e}{d}\right) + \ln\frac{2}{\delta}\right) \\
&= \frac{m}{2} + \frac{4}{\epsilon}\left(d\ln\left(\frac{16}{\epsilon}\right) + \ln\frac{2}{\delta}\right).
\end{aligned}$$

Thus, it is sufficient to have $m \geq \frac{m}{2} + \frac{4}{\epsilon}\left(d\ln\left(\frac{16}{\epsilon}\right) + \ln\frac{2}{\delta}\right)$, which gives the desired sample complexity. □

The above sample complexity bound is tight for arbitrary consistent learners (Auer and Ortner, 2007). It turns out the $\ln\frac{1}{\epsilon}$ factor can be removed in this sample complexity upper bound for some more sophisticated learners achieving the optimal samplexity of $\Theta(\frac{1}{\epsilon}(d + \log\frac{1}{\delta}))$ (Hanneke, 2015).

Additional Resources:

- Peter Auer and Ronald Ortner. "A new PAC bound for intersection-closed concept classes." Machine Learning 66, no. 2 (2007): 151-163.
- Steve Hanneke. "The optimal sample complexity of PAC learning." Journal of Machine Learning Research 17, no. 38 (2016): 1-15.