VC-Dimension

Aaron Anderson

July 10 - 13, 2024

Shattering Sets

Let X be a set, and let $\mathcal{F}$ be a family of subsets of X. If $A \subseteq X$ is a subset of X, then we define

$$\mathcal{F} \cap A = \{S \cap A : S \in \mathcal{F}\}.$$

We say that $\mathcal{F}$ *shatters* A if $\mathcal{F} \cap A$ is the whole powerset of A.

Problem 1. For some of the following families of sets, how big of a set A can you find that is shattered by $\mathcal{F}$?

- $X = \mathbb{R}$, $\mathcal{F}$ is the set of intervals
- X is infinite, and $\mathcal{F}$ is the set of subsets of X of size at most d (where d is some natural number)
- $X = \mathbb{R}^2$, $\mathcal{F}$ is the set of half-planes - one side of a line
- $X = \mathbb{R}^2$, $\mathcal{F}$ is the set of convex sets (a set S is convex when any line segment containing two points of S is also contained in S)
- $X = \mathbb{R}^2$, $\mathcal{F}$ is the set of axis-aligned rectangles (sides are parallel to x- and y-axes)
- $X = \mathbb{R}^d$, $\mathcal{F}$ is the set of half-spaces (a half-space is the the solution set of an inequality $a_1x_1 + \cdots + a_dx_d \leq c$ for some numbers $a_1, \ldots, a_d, c$)

Hint: See Radon's Theorem below.

Call the largest d such that $\mathcal{F}$ shatters a set of size d the *VC-dimension* of $\mathcal{F}$. If there's no such d, we say the VC-dimension of $\mathcal{F}$ is ∞.

Shatter Functions

We can also measure a degree of shattering with the *shatter function.* Define

$$\pi_{\mathcal{F}}(A) = |\mathcal{F} \cap A|$$

and for $n \in \mathbb{N}$,

$$\pi_{\mathcal{F}}(n) = \max_{|A|=n} |\mathcal{F} \cap A|.$$

Note that $\mathcal{F}$ shatters a set of size n if and only if $\pi_{\mathcal{F}}(n) = 2^n$.

Problem 2. Let $X = \mathbb{R}$, and let $\mathcal{F} = \{(a, b) : a < b\}$ be the set of intervals. Calculate $\pi_{\mathcal{F}}(n)$ for all n.

Problem 3. Let $X = \mathbb{R}^2$, and let $\mathcal{F}$ be the set of half-planes. Calculate $\pi_{\mathcal{F}}(A)$ where A consists of n points arranged around a circle.

Problem 4. Suppose X is a set and $\mathcal{F}$ is a finite set of subsets of X, with $|\mathcal{F}| = N$. What can we say about the shatter function $\pi_{\mathcal{F}}(n)$?

Dual Set Families

We can turn around any set family $\mathcal{F}$ on a set X, and get a new set system, called its *dual.*

Definition. Let $\mathcal{F}^*$, the dual of $\mathcal{F}$, be the set family on the set $\mathcal{F}$, consisting of the sets

$$\mathcal{F}^* = \{\{S \in \mathcal{F} : x \in S\} : x \in X\}.$$

Problem 5. Use a dual set family to explain what the lazy caterer's problem (from the lecture at the beginning of class) has to do with shattering.

Problem 6. Show that if $\mathcal{F}$ has finite VC-dimension, then so does $\mathcal{F}^*$, and vice versa.

Specifically, if we know that $\mathcal{F}$ has VC-dimension at most d, what do we know about the VC-dimension of $\mathcal{F}^*$?

Background: Radon's Theorem

Definition 0.1. The *convex hull* of a finite set $A = \{a_1, \dots, a_n\}$ in $\mathbb{R}^d$ is the intersection of all half-spaces containing A.

We can also define the convex hull as consisting of all points $c_1a_1 + \cdots + c_na_n$ where each $c_i \geq 0$, and $c_1 + \cdots + c_n = 1$. (To calculate $c_1a_1 + \cdots + c_na_n$ where the cs are real numbers and the as are points in space, multiply and add coordinatewise.)

Theorem 0.2 (Radon)**.** *Any set of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two nonempty pieces whose convex hulls intersect.*

Extremely Optional Problems:

Problem 7. Prove that the two definitions of convex hull are equivalent.

Problem 8. Prove Radon's Theorem, starting with the following linear algebra fact:

For any $A = \{a_1, \dots, a_n\}$ in $\mathbb{R}^d$ with $n > d$, there are real numbers $c_1, \dots, c_n$, at least one of which is nonzero, such that

$$\sum_{i=1}^{n} c_1a_1 + \cdots + c_na_n = 0.$$

Shatter Functions - review from yesterday

We can measure a degree of shattering with the *shatter function.* Define

$$\pi_{\mathcal{F}}(A) = |\mathcal{F} \cap A|$$

and for $n \in \mathbb{N}$,

$$\pi_{\mathcal{F}}(n) = \max_{|A|=n} |\mathcal{F} \cap A|.$$

Note that $\mathcal{F}$ shatters a set of size n if and only if $\pi_{\mathcal{F}}(n) = 2^n$.

Problem 9. Let $X = \mathbb{R}$, and let $\mathcal{F} = \{(a, b) : a < b\}$ be the set of intervals. Calculate $\pi_{\mathcal{F}}(n)$ for all n.

Problem 10. Let $X = \mathbb{R}^2$, and let $\mathcal{F}$ be the set of half-planes. Calculate $\pi_{\mathcal{F}}(A)$ where A consists of n points arranged around a circle.

Problem 11. Suppose X is a set and $\mathcal{F}$ is a finite set of subsets of X, with $|\mathcal{F}| = N$. What can we say about the shatter function $\pi_{\mathcal{F}}(n)$?

The Big Lemmas

If we know every value of the shatter function $\pi_{\mathcal{F}}(n)$, then we know the VC-dimension of $\mathcal{F}$. Now let's figure out what the VC-dimension tells us about the shatter function.

Let's start with the easy case.

Problem 12. Suppose $\mathcal{F}$ has infinite VC-dimension. What is $\pi_{\mathcal{F}}(n)$?

The harder direction is seeing what happens to the shatter function when the VC-dimension is finite. We will find an upper bound on $\pi_{\mathcal{F}}(n)$ that holds whenever we assume the VC-dimension of $\mathcal{F}$ is at most d. This upper bound, together with our understanding of the infinite VC-dimension case, are known as the *Sauer-Shelah Lemma*.

To prove this bound, let's start by defining a new family of sets, which we can count more easily: If $\mathcal{F}$ is a family of subsets of X, let $\mathrm{SH}(\mathcal{F})$ be the set of subsets of X shattered by $\mathcal{F}$.

Problem 13. Suppose $\mathcal{F}$ has VC-dimension at most d, and X is finite with $|X| = n$. Find an upper bound for $|\mathrm{SH}(\mathcal{F})|$. Can you find a family $\mathcal{F}$ that satisfies that bound exactly?

Let $f_d(n)$ be the bound from that Problem 13. In the rest of this worksheet, we'll prove this lemma:

Lemma 0.3 (Sauer-Shelah)**.** *Let $\mathcal{F}$ be a family of subsets of X with VC-dimension at most d. Show that $\pi_{\mathcal{F}}(n) \leq f_d(n)$.*

Problem 14. For each d, find a family $\mathcal{F}$ of VC-dimension d such that $\pi_{\mathcal{F}}(n) = f_d(n)$ for all d.

This problem sets up the strategy for proving the Sauer-Shelah lemma.

Problem 15. Assume that whenever $\mathcal{F}$ has VC-dimension at most d, and X is finite with $|X| = n$, $|\mathcal{F}|$ satisfies the upper bound from Problem 13.

Show that for any set X, if $\mathcal{F}$ has VC-dimension at most d, then $\pi_{\mathcal{F}}(n)$ satisfies that bound. Find a family $\mathcal{F}$ that satisfies that bound exactly.

To complete the lemma, we need to show that whenever $|X| = n$ and $\mathcal{F}$ has VC-dimension at most d, we have $|\mathcal{F}| \leq f_d(n)$. Here are two approaches, you can choose either one:

Sauer-Shelah Lemma: Algorithmic Approach

Our first approach to proving the upper bound we want will be by incrementally swapping out the family $\mathcal{F}$ for a different one. We assume that $|X| = n$ and $\mathcal{F}$ has VC-dimension at most d, and we will try to bound $|\mathcal{F}|$.

Problem 16. Pick some $x \in X$. For every $S \in \mathcal{F}$, if $x \in S$ but $S \setminus \{x\} \notin \mathcal{F}$, replace S with $S \setminus \{x\}$.

Show that $\mathrm{SH}(\mathcal{F}') \leq \mathrm{SH}(\mathcal{F})$.

Problem 17. Iterate the process from the previous problem until you can't anymore, and call that family of sets $\mathcal{G}$.

Show that $\mathcal{G} = \mathrm{SH}(\mathcal{G})$, and conclude that $|\mathcal{G}|$ satisfies our bound. What does this tell us about $|\mathcal{F}|$?

Pajor's Lemma

There's another way to prove the Sauer-Shelah Lemma, by proving an even stronger lemma.

Problem 18 (Pajor's Lemma)**.** Show that if X is finite, then $\mathcal{F}$ shatters at least $|\mathcal{F}|$ sets, so $|\mathcal{F}| \leq |\mathrm{SH}(\mathcal{F})|$.

Hint: Induction on $|X|$.

Problem 19 (Sauer-Shelah Lemma)**.** Use Pajor's Lemma to prove Sauer-Shelah.

Examples

Problem 20. Let X be a set, and let $\mathcal{F}_1, \mathcal{F}_2$ be two different families of subsets of X, each with finite VC-dimension. Let $\mathcal{F}$ be the set of all sets $S_1 \cap S_2$ where $S_1 \in \mathcal{F}_1$ and $S_2 \in \mathcal{F}_2$. Show that $\mathcal{F}$ has finite VC-dimension.

What other families of sets can you build out of $\mathcal{F}_1$ and $\mathcal{F}_2$ and get the same shatter function bound?

As we've seen, any set family with finite VC-dimension has a polynomial upper bound on $\pi_{\mathcal{F}}(n)$. We define the *VC-density* of $\mathcal{F}$ as the least $d \in \mathbb{R}$ (if it exists) such that there's some constant C with $\pi_{\mathcal{F}}(n) \leq Cn^d$ for all n.

Problem 21. If we know the VC-dimension of $\mathcal{F}$, what do we know about its VC-density? If we know the VC-density of $\mathcal{F}$, is there anything we can say about its VC-dimension?

Problem 22. Let X be a set, and let $\mathcal{F}_1, \mathcal{F}_2$ be two different families of subsets of X, with VC-densities d_1 and d_2. Let $\mathcal{F}$ be the set of all sets $S_1 \cap S_2$ where $S_1 \in \mathcal{F}_1$ and $S_2 \in \mathcal{F}_2$.

What do we know about the VC-density of $\mathcal{F}$?

Problem 23. Let X be a set, and let $\mathcal{F}_1, \mathcal{F}_2$ be two different families of subsets of X, with VC-densities d_1 and d_2.

What is the VC-density of $\mathcal{F}_1 \cup \mathcal{F}_2$?

Problem 24. If P is a set of points in $\mathbb{R}^2$, and L is a set of lines, let

$$I(P, L) = \{(p, \ell) : p \in \ell\}$$

be the set of *incidences*. Let $I(n)$ be the maximum of $|I(P, L)|$ over all sets with $|P| = |L| = n$.

The Szemerédi-Trotter theorem says that there are constants C_1 and C_2 that

$$C_1 n^{4/3} \leq |I(P, L)| \leq C_2 n^{4/3}.$$

Find a set family $\mathcal{F}$ with VC-density $\frac{4}{3}$ - this shows it doesn't have to be an integer. In fact, VC-density can be any real number d with $d \geq 1$, or 0.

Problem 25. If you $\mathcal{F}_1, \mathcal{F}_2$ be two different families of subsets of X, with VC-densities d_1 and d_2, can you construct a set family with VC-density $d_1 + d_2$?

PAC Learning

For an application of VC-Dimension, let's look at learning theory!

Let's fix a set X, with a set family $\mathcal{F}$. We have some way of generating random points in X.

I'm thinking of a secret set $S \in \mathcal{F}$, and you'd like to guess S. In order to you to make your guess, we get to pick n random points in X, and I have to tell you if they're in S. It's unlikely that your guess will be correct - there may be infinitely many sets in $\mathcal{F}$ that would give the same answer! However, it'll be good enough if your guess is *approximately* correct - but what's a good notion of error?

We're going to say that the error of a guess S' is the probability that a random point in X will give a different answer between S and S'. That is - it's the probability of getting a point in the *symmetric difference* $S \Delta S' = (S \setminus S') \cup (S' \setminus S)$. We have to decide how much of an error we're ok accepting - say we pick some $\varepsilon > 0$, and we'll say you're *approximately correct* when the probability of $S \Delta S'$ is at most ε.

We also can't completely guarantee that you're approximately correct, because you could just get really unlucky with the randomly-selected points. So we'll settle for *probably approximately correct* learning - we pick some $\delta > 0$, and you're doing well if you can be approximately correct with probability $\geq 1 - \delta$.

This depends on δ, ε, and the source of random points, but we'll say $\mathcal{F}$ can be PAC learned when for every δ, ε, there is some number n of guesses that'll work, regardless of the source of random points.

Error Family

Tomorrow, we'll show that every family with finite VC-dimension can be PAC learned, and we'll even give an upper bound on the number of samples we need. To do that, we're going to look first at another related family - the *error family* of deviations from my secret set.

Definition. Suppose I'm thinking of a set $S \in \mathcal{F}$. Define the set family $\Delta(S) = \{S' \Delta S : S' \in \mathcal{F}\}$.

This actually shatters the same sets as $\mathcal{F}$! Suppose that $A \in \mathrm{SH}(\mathcal{F})$. Intuitively, we know that each subset of A is produced by intersecting with something in $\mathcal{F}$ - by taking the symmetric differences with S, these sets swap which subsets of A they carve out, so all subsets are still produced.

More precisely, for each set $B \subseteq A$, look at $(B \Delta S) \cap A$ - this is also a subset of A, so there is $S' \in \mathcal{F}$ with $S' \cap A = (B \Delta S) \cap A$. Now look at $(S' \Delta S) \cap A$. This is

$$(S' \cap A)\Delta(S \cap A) = ((B \Delta S) \cap A)\Delta(S \cap A) = ((B \Delta S) \cap S) \cap A = B \cap A = B.$$

We'll see that if we understand the shatter function of a set family, then any random sample of enough points is (with high probability) going to intersect any large (high probability) sets in the family. Applying this logic to the error family, we see that in the PAC learning setup, there's a high probability that our random sample will notice any large error sets, allowing us to avoid them and be approximately correct.

PAC Learning Lower Bound

What does PAC learning have to do with shattering and VC-dimension? Suppose $\mathcal{F}$ shatters a set $\{x_1, \ldots, x_d\}$. We're going to see that this situation makes $\mathcal{F}$ hard to learn. More specifically, we're going to see that it makes learning hard for *some probability distribution* - namely, when we choose points uniformly from $\{x_1, \ldots, x_d\}$.

In this situation, the error is $\frac{1}{d}|(S \Delta S') \cap \{x_1, \ldots, x_d\}|$. In particular, you only need to get a similar intersection with $\{x_1, \ldots, x_d\}$, so we could just assume $\mathcal{F}$ *is* the power set of $\{x_1, \ldots, x_d\}$.

We're going to see that if we shoot for error less than $\varepsilon \leq \frac{1}{8}$ with probability $\delta \leq \frac{1}{8}$, then no matter what your guessing procedure is, you'll have a bad probability of being approximately correct. I'm going to choose a set in $\mathcal{F}$ uniformly at random, and we'll see that there's a high probability you're not doing well.

Suppose you try to guess my random set from $\frac{d}{2}$ data points. These data points are also random, so there's a chance of repetition - say we actually get $m \leq \frac{d}{2}$ distinct points. So, maybe you know whether $x_1, \dots, x_m$ are in my set, but you have no info at all about the rest of the d points, so you get each of them wrong or right by coin flip. On average, you'll get $\frac{d-m}{2} \geq \frac{d}{4}$ of them wrong, so your average error is at least $\frac{1}{4}$. But what's the probability that your error is more than $\frac{1}{8}$?

Suppose the probability the error is more than $\frac{1}{8}$ is P. No matter what, the error is at most 1. So the average is at most $\frac{1}{8}(1-P) + P \leq \frac{1}{8} + P$, but we know the average is at least $\frac{1}{4}$, so $P \geq \frac{1}{8}$.

Because this is the *average* chance of error over all of my choices for $S \in \mathcal{F}$, there must be one of them that's at least this hard to guess! This tells us that to PAC learn $\mathcal{F}$, you need at least $\frac{d}{2}$ data points. If $\mathcal{F}$ has infinite VC-dimension, you'll never get there.

PAC Learning Upper Bound and Epsilon Nets

Meanwhile, if $\mathcal{F}$ has finite VC-dimension, your strategy will be simple: just guess some S' compatible with all n random data points. There's not much else you can do. There always will be at least one - my secret set S. We want to find out the probability that the n random data points are good enough that you're guaranteed to be approximately correct if you follow them.

That is, our data points are *good* when for every "big" error set $S \Delta S' \in \Delta(S)$ with $\mathbb{P}[S \Delta S'] \geq \varepsilon$, there is some data point in $S \Delta S'$, keeping us from picking S'.

Definition. We call a set of points $A \subseteq X$ an *ε-net* for $\mathcal{F}$ if $A \cap S \neq \emptyset$ for every $S \in \mathcal{F}$ of probability $\geq \varepsilon$.

To finish finding a strategy for PAC learning, we just need to figure out how big n needs to be for the probability of n random samples being an ε-net to be at least $1 - \delta$.

Theorem. *There is a constant C such that if $\mathcal{F}$ has VC-dimension at most d, then*

$$C\left(\frac{1}{\varepsilon}\log\frac{1}{\delta} + \frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\right)$$

random points have a probability at least $1 - \delta$ of being a ε-net.

Proof. We'll basically prove that *some* number works, and skip some details of calculating how many we need.

Let n be the number of samples we want to draw. First we draw n random points, $A = \{x_1, \dots, x_n\}$, and then we draw *another* n random points, $B = \{x_{n+1}, \dots, x_{2n}\}$. If A fails to be a ε-net, then there is some $S \in \mathcal{F}$ with $\mathbb{P}[S] \geq \varepsilon$ that we've missed. We'll see that the probability of A failing is at most twice the probability that

- A fails
- B hits some big S that A missed, $\frac{\varepsilon n}{2}$ times.

We'll call this probability P.

If A fails, then for *any* big S that A missed, the *average* number of times that B hits A is $\mathbb{P}[S]n \geq \varepsilon n$. By the probability lemma below, $n \geq \frac{8}{\varepsilon}$, this number is at least half its average with probability $\frac{1}{2}$. This means that the probability A fails is at most $2P$.

We can then show P is small - this scenario isn't very likely, by starting instead by picking $2n$ points $C = \{x_1, \dots, x_{2n}\}$, and then sorting them randomly into A and B. In fact, for *any* multiset C, this will be unlikely - so let C be fixed.

For this scenario to happen, from this perspective, we need

- some $S \in \mathcal{F}$ with $\mathbb{P}[S] \geq \varepsilon$
- $S \cap C$ contains at least $\frac{\varepsilon n}{2}$ points
- when we choose which half of the points go to A, we take none of the points of $S \cap C$.

Note that there are many choices of S that could make this happen - we can't possibly account for all of them individually. But actually, this only depends on $S \cap C$ - and there are only $\pi_{\mathcal{F}}(2n)$ of these!

Thus we can bound P by adding up the probabilities from each $S \cap C$ - call this subset D, because we don't actually depend which S gave us $S \cap C$. The probability that when we randomly split C into A and B, we end up with $D \cap A = \emptyset$, is

$$\frac{\binom{n}{|D|}}{\binom{2n}{|D|}} \leq \prod_{i=0}^{|D|-1} \frac{n-1}{2n-1} \leq \frac{1}{2^{|D|}}.$$

Because we only care about D with $|D| \geq \frac{\varepsilon n}{2}$, we see that the total over all choices of D is at most

$$\binom{2n}{\leq d} 2^{-\frac{\varepsilon n}{2}} \leq C n^d 2^{-\frac{\varepsilon n}{2}}$$

for some choice of C. The exponential shrinks faster than the polynomial grows, so for large enough n, this is less than 2δ, so the probability A fails is less than δ.

Problem 26. This "exercise left to the reader" isn't that enlightening, but if you want, you can calculate from what we have so far that

$$n = C \left(\frac{1}{\varepsilon} \log \frac{1}{\delta} + \frac{d}{\varepsilon} \log \frac{1}{\varepsilon} \right)$$

is big enough for some C.

□

HW: Probability Lemmas

I left out the proof of the following lemma, which is an example of a *Chernoff bound*:

Lemma. *Let T be a random trial, which succeeds with probability p. We try T n times, independently, and $np \geq 8$. The probability that more than $\frac{1}{2}np$ trials succeed is at least $\frac{1}{2}$.*

The intuition for this is as follows: We expect to succeed np times on average, and as n gets bigger, the answer will get closer and closer to the average, by the law of large numbers. Being less than half of the average is indeed rare.

The following exercises prove it, by first proving a bunch of classical probability theory results about the average of a random variable, or as we usually say in probability theory, the *expectation*. We'll write $\mathbb{P}[A]$ to be the probability of an event A, and $\mathbb{E}[X]$ to be the expectation of the random variable X.

Theorem (Markov's Inequality)**.** *Let X be a random variable that only takes nonnegative values, and let $a > 0$. Then*

$$\mathbb{P}[X \geq a\mathbb{E}[x]] \leq \frac{1}{a}.$$

Problem 27. Let's prove Markov's Inequality. To do so, use another random variable Y, defined by

$$Y = \begin{cases} 0 & X < a\mathbb{E}[x] \\ a\mathbb{E}[x] & X \geq a\mathbb{E}[x] \end{cases}.$$

How does $\mathbb{E}[X]$ compare to $\mathbb{E}[Y]$?

Our further theorems and exercises will use the notion of *variance*, which measures how much a random variable *varies* away from its expectation.

Definition. The *variance* of X is

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Chebyshev's Inequality says that we can use the variance of a random variable to bound how often it's particularly different than its average.

Theorem (Chebyshev's Inequality)**.** *Let X be a random variable with finite variance. Then*

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq k\sqrt{\mathrm{Var}[X]}\right] \leq \frac{1}{k^2}.$$

Problem 28. Prove Chebyshev's Inequality.

One reason variance is so good to work with is that if we add two *independent* random variables, the variances add:

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Problem 29. Prove the above equation from the fact that if X and Y are independent,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Problem 30. Suppose $X_1, \ldots, X_n$ are independent random variables, each of which is 1 with probability p and 0 with probability $1 - p$, and let $X = X_1 + \cdots + X_n$.

Calculate $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

Problem 31. Prove our probability lemma.