Machine Learning Theory (CSC 431/531) - Lecture 7

Nishant Mehta

1 The Vapnik-Chervonenkis dimension

Recall that the growth function is defined as

$$\Pi_{\mathcal{F}}(n) = \sup_{(x_1,\ldots,x_n)\in\mathcal{X}^n} \left|\mathcal{F}_{|\mathbf{x}_1^n}\right|.$$

Definition 1. We say that a concept class $\mathcal{F}$ *shatters* $\mathbf{x}_1^n$ if $\left|\mathcal{F}_{|\mathbf{x}_1^n}\right| = 2^n$.

Suppose that $\mathcal{F}$ shatters a set of n inputs. Then no matter their labels, there exists a hypothesis $f \in \mathcal{F}$ which matches the labels; the empirical risk is thus zero even though the Bayes risk (and hence the risk of ERM) might be well above zero! Hence, for n in this regime, uniform convergence of the empirical risk to the true risk is hopeless. Indeed, the upper bound in Theorem 1 of the last lecture is nontrivial only if $n \gg \log \Pi_{\mathcal{F}}(n)$, but since $\mathcal{F}$ shatters the inputs, $\log \Pi_{\mathcal{F}}(n) = n \log 2$. This observation motivates the following fundamental notion of complexity.

Definition 2. The Vapnik-Chervonenkis (VC) dimension of a set of classifiers $\mathcal{F}$, denoted by $\mathrm{VCdim}(\mathcal{F})$, is the cardinality of the largest set shattered by $\mathcal{F}$. If $\mathcal{F}$ can shatter sets of arbitrarily large size, then $\mathrm{VCdim}(\mathcal{F}) = \infty$. A class is a *VC class* if it has finite VC dimension.

Unpacking the definition. From the definition of shattering, an equivalent definition of the VC dimension of $\mathcal{F}$ is the largest $k \geq 1$ such that $\Pi_{\mathcal{F}}(k) = 2^k$; the VC dimension is infinite if $\Pi_{\mathcal{F}}(k) = 2^k$ for every $k \geq 1$. Unpacking once more, the VC dimension of $\mathcal{F}$ is the largest $k \geq 1$ such that there exists $x_1, \ldots, x_k \in \mathcal{X}$ for which, for any $\mathbf{b} \in \{0,1\}^k$, there exists $f \in \mathcal{F}$ satisfying

$$f(x_j) = b_j \qquad \text{for all } j \in [k].$$

That is,

$$\mathrm{VCdim}(\mathcal{F}) = \max\left\{k : \exists\, \mathbf{x} \in \mathcal{X}^k,\, \forall\, \mathbf{b} \in \{0,1\}^k,\, \exists\, f \in \mathcal{F},\, f_{|\mathbf{x}} = \mathbf{b}\right\}.$$

As we will see in Corollary 1, the VC dimension says much more about the growth function than its definition suggests. Remarkably, the growth function $\Pi_{\mathcal{F}}$ exhibits only two behaviors as n increases:

- $\Pi_F(n) = 2^n$;
- $\Pi_{\mathcal{F}}(n)$ grows polynomially.

There is nothing in between, and the VC dimension characterizes where the phase transition occurs. Suppose that $V = \mathrm{VCdim}(\mathcal{F})$. For $n \leq V$, we have $\Pi_{\mathcal{F}}(n) = 2^n$ by definition. However, as soon as $n > V$, it holds that $\Pi_{\mathcal{F}}(n) = O(n^V)$. One could imagine that for $n > V$, it holds that $\Pi_{\mathcal{F}}(n) = \Theta(2^{\sqrt{n}})$, but this is completely ruled out.

2 Examples

In the following examples, we often identify a hypothesis by the set of points it labels positive.

Threshold functions Let $\mathcal{F}$ be the class of threshold functions over $\mathbb{R}$. As we saw previously, $\Pi_{\mathcal{F}}(n) = n + 1$. Hence, $\text{VCdim}(\mathcal{F}) = 1$. This argument is overkill though, as there is no need to nail down the correct growth function for this class. Instead, observe that $\mathcal{F}$ can shatter any single point. However, $\mathcal{F}$ cannot shatter any set of size 2 since, if $x_1 < x_2$, any threshold function that labels x_1 positive must also label x_2 as positive.

Intervals Let $\mathcal{F}$ be the class of intervals over $\mathbb{R}$. It is easy to work out that $\Pi_{\mathcal{F}}(n) = \binom{n+1}{2} + 1$, so $\text{VCdim}(\mathcal{F}) = 2$ (as $\Pi_{\mathcal{F}}(3) = 7 < 2^3$). Again, arguing more simply, any set of two distinct points can be shattered by $\mathcal{F}$. On the other hand, for any set of points $x_1 < x_2 < x_3$, no hypothesis in $\mathcal{F}$ can assign label 0 to x_2 if it assigns label 1 to x_1 and x_3. Thus, $\mathcal{F}$ cannot shatter any set of size 3.

Axis-aligned rectangles Let $\mathcal{F}$ be the class of axis-aligned rectangles. These are all rectangles of the form $\prod_{j=1}^{d}[a_j, b_j]$ for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Note that the case of $d = 1$ is identical to the class of intervals over $\mathbb{R}$. Let's work out the case of $d = 2$ (the general case will be left as an exercise). First, there is a set of 4 points in the plane which are shattered by $\mathcal{F}$. *(I drew the example in class)*

Next, observe that if there is a point in a set for which none of its coordinates are extremal (with respect to the elements in the set), then it is contained in the tightest rectangle that contains the other points. For any set of 5 points, at least one point has no extremal coordinate, as there are two dimensions and hence 4 extremal coordinate values. Thus, in this case, $\text{VCdim}(\mathcal{F}) = 4$. It turns out that for the case of general d, we have $\text{VCdim}(\mathcal{F}) = 2d$.

Linear separators Let $\mathcal{F}$ be the class of linear separators, so that each hypothesis is of the form $f(x) = \text{sgn}(\langle w, x\rangle + b)$ for some $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. Let's look at the case of $d = 2$. It is easy to see that 3 points in general position (i.e. 3 non-collinear points) can be shattered by $\mathcal{F}$. *(I drew the example in class)*

Now, suppose that we have 4 points. We consider two exhaustive cases.

Case 1: Suppose that one point is contained in the convex hull of the other 3. Since any halfspace is a convex set, in this case it is impossible to label as negative the point in the interior of the convex hull if the other 3 points are labeled as positive.

Case 2: Suppose that all 4 points lie on the boundary of their convex hull. Consider taking a walk along the convex hull and picking (among the the original 4 points) the first and third point encountered. Suppose that there is a linear separator which labels these two points as positive and the other two points as negative. Next, note that for any two points with a common label, all the points in their convex hull (a line) must have the same label. Hence, there are two lines consisting of oppositely labeled points. But these two lines are not parallel (if they were, we are in Case 1, and the lines would actually have to be identical), and so they share a common point.[1] Since this common point cannot be labeled both positive and negative, there can be no such linear separator. So, the VC dimension for linear separators in $\mathbb{R}^2$ is 3.

It is a simple exercise to work out that the VC dimension of the class of linear separators in $\mathbb{R}$ is 2. As we saw above, the VC dimension of the class of linear separators in $\mathbb{R}^2$ is 3. What happens in general, when $\mathcal{F}$ is the class of linear separators in $\mathbb{R}^d$? We then have $\text{VCdim}(\mathcal{F}) = d + 1$.

[1] We are in Euclidean geometry!

3 Perles-Sauer-Shelah-Vapnik-Chervonenkis Lemma

As promised, we will now show the following fundamental result.

Lemma 1. *Let $\mathcal{F}$ be a class of VC dimension V. Then for all n,*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{j=0}^{V} \binom{n}{j}. \tag{1}$$

(The plan is to prove this next class (on Thursday), but a proof can also be found in the Understanding Machine Learning book of Shalev-Shwartz and Ben-David: see Lemma 6.10 therein. My proof will be based on the (nice) one in Nina Balcan's lecture notes: http://www.cs.cmu.edu/%7Eninamf/ML11/lect0922.pdf)

This result has a fascinating history. It originally was stated in a paper of Vapnik and Chervonenkis in 1968, without proof, although it is suspected that they already had a proof. Its proof was first published in a paper of Vapnik and Chervonenkis (1971) that was submitted in 1969. The result also was independently proved by both Sauer (1972) and Shelah (1972), with Shelah giving the credit for his result to Perles. For brevity, we will call this result Sauer's Lemma.[2]

Our interest, however, is in what the result buys us rather than who deserves what fraction of the credit. Clearly, the growth of $\Pi_{\mathcal{F}}(n)$ is $O(n^V)$, but let's get an explicit bound.

Lemma 2. *For all $n \geq V$,*

$$\sum_{j=0}^{V} \binom{n}{j} \leq \left(\frac{en}{V}\right)^V.$$

Proof. Since $n \geq V$, it holds that

$$\sum_{j=0}^{V} \binom{n}{j} \leq \sum_{j=0}^{V} \binom{n}{j} \left(\frac{V}{n}\right)^{j-V} \leq \sum_{j=0}^{n} \binom{n}{j} \left(\frac{V}{n}\right)^{j-V}.$$

By rewriting appropriately, we may apply the binomial theorem:

$$\left(\frac{n}{V}\right)^V \sum_{j=0}^{n} \binom{n}{j} \left(\frac{V}{n}\right)^j = \left(\frac{n}{V}\right)^V \left(1 + \frac{V}{n}\right)^n \leq \left(\frac{n}{V}\right)^V e^V.$$

□

The following corollary of Sauer's lemma is immediate.

Corollary 1. *Let $\mathcal{F}$ be a class of VC dimension V. Then for all $n \geq V$,*

$$\Pi_{\mathcal{F}}(n) \leq \left(\frac{en}{V}\right)^V.$$

[2] To pick Sauer's name is ironic as he actually proved an improved version of (1) where the summation goes up to only $V - 1$. Yet, perhaps this improvement is enough to break the symmetry in his favor.

4 Back to uniform convergence for infinite (VC) classes

Armed with Corollary 1, we may now apply our uniform convergence result that depended on $\Pi_{\mathcal{F}}(n)$ to get an explicit bound for VC classes.

Theorem 1 (Vapnik and Chervonenkis, 1971)**.** *Let $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ be a VC class with $\mathrm{VCdim}(\mathcal{F}) = V$. For any probability distribution P, any $n \geq V$, and any $\varepsilon > 0$,*

$$\Pr\left(\sup_{f\in\mathcal{F}} |(P - P_n)f| > \varepsilon\right) \leq 8\left(\frac{en}{V}\right)^V e^{-n\varepsilon^2/32}.$$

The above implies that for any probability distribution P and any $n \geq V$, with probability at least $1-\delta$

$$\sup_{f\in\mathcal{F}} |(P - P_n)f| \leq \sqrt{\frac{32\left(V\log\frac{en}{V} + \log\frac{8}{\delta}\right)}{n}}.$$

The proof of the first result is immediate from Theorem 1 from last class and Corollary 1. The second result follows from inversion (set the probability equal to δ and solve for ε).

5 Uniform convergence in the realizable case

We already have seen that agnostically learning is possible when $\mathcal{F}$ is a VC class, and the excess risk obtainable via ERM converges (ignoring logarithmic factors) at the rate

$$O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{F}) + \log\frac{1}{\delta}}{n}}\right).$$

However, in the PAC learning (i.e. realizable) setting, at least for finite classes we were able to obtain a better convergence rate in that the rate did not have a square root. The same is true for VC classes, as we will now see.

Theorem 2 (Vapnik and Chervonenkis (1971))**.** *Let $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ be a VC class with $\mathrm{VCdim}(\mathcal{F}) = V$, and let $\hat{f}$ be an ERM classifier (which, given a training sample, outputs a hypothesis in $\mathcal{F}$ that minimizes the empirical risk), and let P be an arbitrary probability distribution P over $\mathcal{X}\times Y$ that satisfies $Y = c(X)$ for some $c \in \mathcal{F}$.*

Then for any $n \geq V$, and any $\varepsilon > 0$.

$$\Pr\left(R(\hat{f}) > \varepsilon\right) \leq 2\left(\frac{2en}{V}\right)^V e^{-n\varepsilon/2}.$$

Equivalently, for any $n \geq V$, with probability at least $1-\delta$

$$R(\hat{f}) \leq \frac{2\left(V\log\frac{2en}{V} + \log\frac{2}{\delta}\right)}{n}.$$

Before proving this result, observe that we can reframe our goal in terms of the convergence of the empirical risk $\hat{R}(\hat{f})$ of ERM to its true risk. Since $R(\hat{f}) = R(\hat{f}) - \hat{R}(\hat{f})$ for ERM, it follows that a high probability bound on $\left|R(\hat{f}) - \hat{R}(\hat{f})\right|$ is exactly equivalent to a high probability bound on the risk $R(\hat{f})$.

The proof of Theorem 2 relies on the above observation and two lemmas. As before, we use the notation that $Z = (X, Y)$ (likewise for $Z_j = (X_j, Y_j)$ and $Z'_j = (X'_j, Y'_j)$). We again introduce a ghost sample $Z'_1, \ldots, Z'_n$; recall that each $Z_j = (X_j, Y_j)$ is a labeled sample drawn from probability distribution P.

Lemma 3. *If $n\varepsilon > 2$, then*

$$\Pr\left(\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j) - \mathsf{E}_{Z\sim P}[\ell_f(Z)]\right| > \varepsilon\right)$$
$$\leq 2\Pr\left(\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right| > \varepsilon/2\right).$$

We won't cover the proof of this result. However, the high-level argument is similar the one we used for the general (agnostic) case.

Lemma 4. *It holds that*

$$\Pr\left(\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right| > \varepsilon/2\right) \leq \Pi_{\mathcal{F}}(2n)2^{-n\varepsilon/2}.$$

Proof. Let $\pi(Z_1), \ldots, \pi(Z_n), \pi(Z'_1), \ldots, \pi(Z'_n)$ be an arbitrary permutation of $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$. Observe that from the i.i.d. property of the double sample, the distribution of the random variable

$$\sup_{f\in\mathcal{F}:\sum_{j=1}^{n}\ell_f(Z_j)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|$$

is equal to the distribution of the random variable

$$\sup_{f\in\mathcal{F}:\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j)) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|$$

Let $U(S_{2n})$ be the uniform distribution over the symmetric group S_{2n}, the set of all permutations over $2n$ items. It therefore holds that

$$\Pr\left(\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right| > \varepsilon/2\right)$$
$$= \mathsf{E}\left[\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right| > \varepsilon/2\right]\right]$$
$$= \mathsf{E}\left[\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sup_{f\in\mathcal{F}:\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j)) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right| > \varepsilon/2\right]\right]\right].$$

We will get a small upper bound just for the internal expectation over π. For a fixed double sample, let $\mathcal{F}_{2n} \subset \mathcal{F}$ be a class which, for each labeling of $X_1, \ldots, X_n, X'_1, \ldots, X'_n$ attainable by a hypothesis in $\mathcal{F}$, contains precisely one representative from f that obtains this labeling. Then the conditional expectation above (conditional on the double sample) is equal to

$$\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sup_{f\in\mathcal{F}_{2n}:\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j)) - \frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right| > \varepsilon/2\right]\right],$$

which is at most

$$\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sum_{f\in\mathcal{F}_{2n}:\,\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right]\right]$$
$$=\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sum_{f\in\mathcal{F}_{2n}}\mathbf{1}\left[\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0\right]\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right]\right]$$
$$=\sum_{f\in\mathcal{F}_{2n}}\Pr_{\pi\sim U(S_{2n})}\left(\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0\ \bigwedge\ \sum_{j=1}^{n}\ell_f(\pi(Z'_j))>n\varepsilon/2\right).$$

Now, suppose that there are at least $r=n\varepsilon/2$ mistakes among $2n$ points. How many permutations are there in which no mistakes occur in the first half of the permuted double sample? There are $n(n-1)\cdots(n-r+1)$ ways to arrange the r mistake points in the second half, and $(2n-r)(2n-r-1)\cdots 1$ ways to arrange the remaining points thereafter. On the other hand, if we are unrestricted in where the mistakes are placed, then the first product is $2n(2n-1)\cdots(2n-r+1)$. Therefore, the fraction of the permutations where no mistakes occur in the first half is at most

$$\frac{n}{2n}\frac{n-1}{2n-1}\cdots\frac{n-r+1}{2n-r+1}\leq 2^{-r}\leq 2^{-n\varepsilon/2}.$$

Therefore,

$$\Pr\left(\sup_{f\in\mathcal{F}:\,\hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)-\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|>\varepsilon/2\right)\leq\mathsf{E}\left[|\mathcal{F}_{2n}|2^{-n\varepsilon/2}\right]$$
$$\leq\Pi_{\mathcal{F}}(2n)2^{-n\varepsilon/2}.$$

□

References

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13 (1):145–147, 1972.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.