Machine Learning Theory (CSC 431/531A) - Lecture 6

Nishant Mehta

1 Effective size of a class

So far, we have seen how to obtain a uniform convergence result when $\mathcal{F}$ is finite. We will now "upgrade" this result to the case when $\mathcal{F}$ is infinite. It is worth thinking about whether our previous proof might already yield a useful bound for infinite $\mathcal{F}$. Unfortunately, the answer is no because the union bound for infinite $\mathcal{F}$ leads to an infinite upper bound. As it turns out, the right way to derive a good uniform convergence bound still relies on a union bound, but applied in a very clever way. For this, we need the notion of the "effective size" of $\mathcal{F}$.

A key idea we will use is that even though $\mathcal{F}$ may be infinite, there are only finitely many ways to classify a given training sample by picking different hypotheses from $\mathcal{F}$. Let's make this concrete. Given a sequence of inputs $\mathbf{x}_1^n = (x_1, \ldots, x_n)$, let $\mathcal{F}_{|\mathbf{x}_1^n}$ be the coordinate projection of $\mathcal{F}$ onto $\mathbf{x}_1^n$. That is,

$$\mathcal{F}_{|\mathbf{x}_1^n} := \left\{ \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{pmatrix} : f \in \mathcal{F} \right\}.$$

Since $\mathcal{F}$ is a set of classifiers, each of which takes values in $\{0,1\}$, we have that $\mathcal{F}_{|\mathbf{x}_1^n} \subseteq \{0,1\}^n$, and hence $|\mathcal{F}_{|\mathbf{x}_1^n}| \leq 2^n$.

Intuitively, even though our hypothesis space $\mathcal{F}$ is infinite, when it is viewed through the lens of the data, there are only finitely many distinct hypotheses.

Example 1 (Threshold functions)**.** Consider learning threshold functions over $\mathbb{R}$, so that we take $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{f_t : t \in \mathbb{R}\}$, where $f_t(x) = \mathbf{1}[x \geq t]$. Suppose that we have n distinct inputs $x_1 < x_2 < \ldots < x_n$. Then it is easy to see that there only $n+1$ distinct ways that $\mathcal{F}$ can classify this training sample, namely:

$$t \in (-\infty, x_1) \quad t \in (x_1, x_2) \quad t \in (x_2, x_3) \quad \cdots \quad t \in (x_{n-1}, x_n) \quad t \in (x_n, \infty).$$

Thus, in this case, $|\mathcal{F}_{|\mathbf{x}_1^n}| = n+1$, and for *any* training sample of size n, $|\mathcal{F}_{|\mathbf{x}_1^n}| \leq n+1$.

2 Growth function

Definition 1. The *growth function* of $\mathcal{F}$ is defined as

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

The growth function of $\mathcal{F}$ evaluated at n is the maximum number of ways a set of n points can be split using functions from $\mathcal{F}$. Often in the literature, you may also see $\Pi_{\mathcal{F}}(n)$ called the n^{th} shatter

coefficient of $\mathcal{F}$. The latter term stems from the notion of "shattering", a crucial component in defining the fundamental notion of Vapnik-Chervonenkis dimension (VC dimension); we will study both shattering and VC dimension in the next lecture.

Example 2 (Intervals)**.** Consider the class of intervals over $\mathbb{R}$. That is, let's take $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{f_{a,b} : -\infty \leq a \leq b \leq \infty\}$ for $f_{a,b}(x) = \mathbf{1}[a \leq x \leq b]$. Similar to the case of threshold functions over $\mathbb{R}$, in this case $\Pi_{\mathcal{F}}(n)$ is finite (this is obvious) and also once again strictly less than 2^n. We leave it as an exercise to work out the exact value of $\Pi_{\mathcal{F}}(n)$.

3 Uniform convergence for infinite classes

We now prove the following fundamental result, originally proved by Vapnik and Chervonenkis in 1971. The result is a uniform convergence result over infinite classes, where the complexity of the class is paid for via the growth function $\Pi_{\mathcal{F}}(n)$.

Before presenting the result, we establish some useful notation that appears frequently in the empirical process theory literature. For any function g from $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, let $P\,g$ denote the expectation of $g(Z)$ for $Z = (X, Y) \sim P$. For a training sample $Z_1, \ldots, Z_n$, the empirical distribution P_n (with respect to $Z_1, \ldots, Z_n$) is $\frac{1}{n}\sum_{j=1}^{n} \delta_{Z_j}(\cdot)$. Here, for some value z, the delta function $\delta_z(z')$ is equal to 1 if $z' = z$ and 0 otherwise. Let $P_n\,g$ denote the expectation of $g(Z)$ when Z is drawn from the empirical distribution P_n. We thus have

$$P\,g = \mathsf{E}_{Z\sim P}[g(Z)] \qquad\qquad P_n\,g = \frac{1}{n}\sum_{j=1}^{n} g(Z_j).$$

For any hypothesis $f \in \mathcal{F}$, define the loss-composed version of f as $g_f \colon (x, y) \mapsto \mathbf{1}[f(x) \neq y]$. So, while f is a random variable mapping from $\mathcal{X}$ to $\{0, 1\}$, the function g_f is a random variable mapping from $\mathcal{X} \times \mathcal{Y}$ to $\{0, 1\}$. From $\mathcal{F}$ we can generate the corresponding "loss-composed" class $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$. Throughout this section, we will use the notation $Z = (X, Y)$ and $Z_j = (X_j, Y_j)$.

Theorem 1 (Vapnik and Chervonenkis, 1971)**.** *For any probability distribution P and any hypothesis space $\mathcal{F} \subset \{0, 1\}^{\mathcal{X}}$, and any $\varepsilon > 0$,*

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

The proof of this result uses two key ideas, both of which are variants of a powerful argument known as symmetrization.

The first symmetrization is sometimes called "symmetrization by ghost sample". The idea is to shift from bounding the uniform deviation of an empirical expectation from the actual expectation to bounding the uniform deviation between two empirical expectations from independent samples of the same size. To this end, let $Z'_1, \ldots, Z'_n$ be an independent copy of $Z_1, \ldots, Z_n$, so that all $2n$ random variables are i.i.d. according to P. We call $Z'_1, \ldots, Z'_n$ a ghost sample, because this is a fictional sample that we do not actually have, but which, nevertheless, we will use in our analysis. Now, similar to P_n, let $P'_n\,g$ denote the empirical expectation of a function $g \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with respect to the ghost sample, so that

$$P'_n\,g = \frac{1}{n}\sum_{j=1}^{n} g(Z'_j).$$

With this notation in place, we establish the first symmetrization lemma.

Lemma 1. *Let $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$ be i.i.d. random variables distributed according to P. Then for any ε satisfying $n\varepsilon^2 \geq 2$,*

$$\Pr\left(\sup_{g\in\mathcal{G}} |(P - P_n)g| > \varepsilon\right) \leq 2\Pr\left(\sup_{g\in\mathcal{G}} |(P'_n - P_n)g| > \varepsilon/2\right)$$

Intuitively, the reason why we view this lemma as progress is because we now only care about viewing the function class $\mathcal{G}$ through the lens of a double sample (the original sample and the ghost sample). Thus, it is conceivable that we may be able to use the finiteness of $\Pi_{\mathcal{F}}(2n)$ if we are clever enough.

Proof. The proof involves a sequence of lower bounds on the probability in the RHS of the lemma.

Let g_n be a function in $\mathcal{G}$ for which $|(P - P_n)g_n| = \sup_{g\in\mathcal{G}} |(P - P_n)g|$ (if there is no such g_n, minor but tedious modifications allow essentially the same proof to go through). Then

$$\Pr\left(\sup_{g\in\mathcal{G}} |(P_n - P'_n)g| > \varepsilon/2\right) \geq \Pr\Big(|(P_n - P'_n)g_n| > \varepsilon/2\Big). \tag{1}$$

Next, observe that $\left[|(P - P_n)g_n| > \varepsilon\right]$ and $\left[|(P - P'_n)g_n| < \frac{\varepsilon}{2}\right]$ together imply $\left[|(P'_n - P_n)g_n| > \frac{\varepsilon}{2}\right]$. Hence, the above is at least

$$\begin{aligned}
&\Pr\Big(\big(|(P - P_n)g_n| > \varepsilon\big) \bigwedge \big(|(P - P'_n)g_n| < \varepsilon/2\big)\Big) \\
&= \mathsf{E}\Big[\mathbf{1}[|(P - P_n)g_n| > \varepsilon] \cdot \Pr\left(|(P - P'_n)g_n| < \varepsilon/2 \mid Z_1, \ldots, Z_n\right)\Big]
\end{aligned} \tag{2}$$

Now, since g_n depends only on $Z_1, \ldots, Z_n$, we can lower bound the conditional probability above using Chebyshev's inequality:

$$\begin{aligned}
\Pr_{Z'_1,\ldots,Z'_n}\left((P - P'_n)g_n \geq \varepsilon/2\right) &\leq \frac{\mathrm{Var}[g_n(Z'_1) \mid Z_1, \ldots, Z_n]}{n\varepsilon^2/4} \\
&\leq \frac{1}{n\varepsilon^2},
\end{aligned}$$

where we used the fact that the variance of a Bernoulli random variable can be at most $1/4$. Since we assumed that $n\varepsilon^2 \geq 2$, it follows that

$$\Pr\left(|(P - P'_n)g_n| < \varepsilon/2 \mid Z_1, \ldots, Z_n\right) \geq \frac{1}{2}.$$

But this implies that (2) (and hence the LHS of (1)) is lower bounded by

$$\frac{1}{2}\Pr\left(|(P - P_n)g_n| > \varepsilon\right).$$

□

The second key idea is another application of symmetrization, this time an argument often called "symmetrization by random signs". For this argument, we employ a sequence of independent Rademacher random variables $\sigma_1, \ldots, \sigma_n$. A Rademacher random variable σ is one which takes the values $\{-1, +1\}$ with equal probability, so $\Pr(\sigma = -1) = \Pr(\sigma = 1) = \frac{1}{2}$.

Lemma 2. *Let $Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$ be i.i.d. random variables and let $\sigma_1, \ldots, \sigma_n$ be independent Rademacher random variables. Then for any $\varepsilon > 0$,*

$$\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\Big(g(Z'_j) - g(Z_j)\Big)\right| > \varepsilon/2\right) \le 2\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right| > \varepsilon/4\right)$$

Proof. Observe that for any choice of sign variables $\sigma_1, \ldots, \sigma_n \in \{-1, +1\}$, the distribution of

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\Big(g(Z'_j) - g(Z_j)\Big)\right|$$

is identical to the distribution of

$$\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\Big(g(Z'_j) - g(Z_j)\Big)\right|.$$

Therefore, letting $\sigma_1, \ldots, \sigma_n$ be i.i.d. Rademacher random variables, it holds that

$$\begin{aligned}
&\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\Big(g(Z'_j) - g(Z_j)\Big)\right| > \varepsilon/2\right)\\
&= \Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\Big(g(Z'_j) - g(Z_j)\Big)\right| > \varepsilon/2\right)\\
&\le \Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z'_j)\right| > \varepsilon/4\right) + \Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right| > \varepsilon/4\right)\\
&= 2\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right| > \varepsilon/4\right)
\end{aligned}$$

□

Proof (of Theorem 1). Lemmas 1 and 2 together imply that for $n\varepsilon^2 \ge 2$,

$$\Pr\left(\sup_{g\in\mathcal{G}}|(P - P_n)g| > \varepsilon\right) \le 4\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right| > \varepsilon/4\right).$$

Next, observe that

$$\begin{aligned}
&\Pr\left(\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g(Z_j)\right| > \varepsilon/4 \mid Z_1, \ldots, Z_n\right)\\
&= \Pr\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j g_f(Z_j)\right| > \varepsilon/4 \mid Z_1, \ldots, Z_n\right)\\
&= \Pr\left(\max_{v\in\mathcal{F}_{|\mathbf{x}_1^n}}\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\,\mathbf{1}[v_j \ne Y_j]\right| > \varepsilon/4 \mid Z_1, \ldots, Z_n\right)\\
&\le \Pi_{\mathcal{F}}(n)\max_{v\in\mathcal{F}_{|\mathbf{x}_1^n}}\Pr\left(\left|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\,\mathbf{1}[v_j \ne Y_j]\right| > \varepsilon/4 \mid Z_1, \ldots, Z_n\right).
\end{aligned}$$

Finally, note that conditional on $(Z_1, \ldots, Z_n)$, each random variable $\sigma_j \mathbf{1}[v_j \neq Y_j]$ is a zero-mean random variable taking values in $[-1, 1]$. Applying Hoeffding's inequality with $[a_j, b_j] = [-1, 1]$ for $j \in [n]$, the above probability is at most

$$2\Pi_{\mathcal{F}}(n)e^{-n\varepsilon^2/32}.$$

The final bound follows, even without the condition $n\varepsilon^2 \geq 2$, since $8e^{-n\varepsilon^2/32} > 1$ for $n\varepsilon^2 < 2$. □