Machine Learning Theory (CSC 482A/581B) - Lecture 9

Nishant Mehta

1 Recap of risk bounds for VC classes

Let's begin by recasting the risk bounds we established in the last few lectures in a minimax framework. In the bound below, the outer infimum serves as the "min" player and the supremum serves as the "max" player. Let $\mathcal{F}$ be a class for which $\mathrm{VCdim}(\mathcal{F}) = V$.

In the agnostic learning setting, we have

$$\inf_{\hat{f}} \sup_{P} \Pr\left(R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) > \sqrt{\frac{32\left(V \log \frac{en}{V} + \log \frac{8}{\delta}\right)}{n}} \right) \leq \delta,$$

where

- the probability is with respect to the training sample $(X_1, Y_1), \ldots, (X_n, Y_n) \stackrel{\text{iid}}{\sim} P$;
- the infimum is over all learning methods that output a hypothesis $\hat{f} \in \mathcal{F}$ that depends on the training sample;
- the supremum is over all probability distributions over $\mathcal{X} \times \mathcal{Y}$.

On the other hand, in the realizable case (i.e. PAC learning), we have

$$\inf_{\hat{f}} \sup_{P \in \mathcal{P}_{\mathcal{F}}} \Pr\left(R(\hat{f}) > \frac{2\left(V \log \frac{2en}{V} + \log \frac{2}{\delta}\right)}{n} \right) \leq \delta, \tag{1}$$

where the probability and infimum are as before, but now the supremum is restricted to $\mathcal{P}_{\mathcal{F}}$, the set of all distributions P over $\mathcal{X} \times \mathcal{Y}$ for which the label $Y = c(X)$ for some $c \in \mathcal{F}$.

Each of the above bounds was established by showing that a particular learning method, empirical risk minimization, obtains low risk with high probability no matter the distribution generating the data.[1] Thus, if $\mathcal{F}$ has finite dimension, a problem is "learnable" in that, no matter the distribution, the gap between the error our learning method achieves and the best possible error using $\mathcal{F}$ converges to zero as the sample size increases. One might then ask if there is a converse:

Is it *necessary* for the VC dimension to be finite in order for a problem to be learnable?

As we will see today, the answer is yes. The VC dimension thus *characterizes* the classes $\mathcal{F}$ for which learnability holds.

[1] Interestingly, the "min" player could perform well even though it was straightjacketed (so to speak) by being forced to be a proper learner (which restricts $\hat{f}$ to lie in $\mathcal{F}$); we could have entertained e.g. allowing predictions according to weighted majority votes over $\mathcal{F}$, but the above bounds hold without broadening the infimum to this larger class.

2 A minimax lower bound for the realizable case

Ignoring logarithmic factors, the upper bound (1) is essentially unimprovable. In all the bounds below, the learning method $\hat{f}$ can be *any* learning method, not necessarily one restricted to taking values in the set $\mathcal{F}$.

Theorem 1. *Let $\mathcal{F}$ satisfy* $\mathrm{VCdim}(\mathcal{F}) = V + 1$*. Then in the realizable case, for $n \geq 15$,*

$$\inf_{\hat{f}} \sup_{P \in \mathcal{P}_{\mathcal{F}}} \Pr\left(R(\hat{f}) \geq \frac{V-1}{12n}\right) \geq \frac{1}{10}.$$

We will not prove the above result (for a proof, see Theorem 14.2 of the book of Devroye, Györfi, and Lugosi (1996)). Instead, we'll prove a related lower bound on the *expected* risk, where the expectation is over the training sample:

Theorem 2. *Let $\mathcal{F}$ be a class for which* $\mathrm{VCdim}(\mathcal{F}) = V + 1$*. Then for any $n \geq V$,*

$$\inf_{\hat{f}} \sup_{P} \mathsf{E}\left[R(\hat{f})\right] \geq \frac{V}{2en}\left(1 - \frac{1}{n}\right).$$

Note that the choice $V + 1$ (instead of V) is to slightly simply the proof.

Proof. We begin by constructing a special family of probability distributions. Observe that since $\mathrm{VCdim}(\mathcal{F}) = V + 1$, there exists a set of points $\{x_0, x_1, \ldots, x_V\}$ that is shattered by $\mathcal{F}$. Let $\mathcal{P}_V = \{P_b : b \in \{0,1\}^V\}$ be a family of 2^V probability distributions. Let $\varepsilon > 0$ be some constant to be determined later. We take all the probability distributions to have the same marginal distribution over $\mathcal{X}$ which is supported on $\{x_0, x_1, \ldots, x_V\}$. Under this distribution, $\Pr(X = x_j) = \varepsilon$ for $j \in [V]$, and $\Pr(X = x_0) = 1 - V\varepsilon$. Under distribution P_b, let $Y = f_b(X)$, with f_b defined as

$$f_b(X) = \begin{cases} b_j & \text{if } j \in [V], \\ 0 & \text{if } j = 0. \end{cases}$$

The idea behind this construction is to let one of these 2^V distributions be the one that generates the data. Learner will then need to identify the correct $b \in \{0,1\}^V$ in order to perform well; for every bit b_j that Learner misses, it pays additional risk ε. However, most of the probability mass is on the "garbage" point x_0, which reveals no information about b. Only samples falling in the set $\{x_1, \ldots, x_V\}$ reveal information about which distribution is correct, and this set has probability only $V\varepsilon$. Now, onwards with the proof.

Let $Z^n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, and let $\hat{f}_{Z^n}$ be an arbitrary classifier (that depends on Z^n). The first step is to lower bound the supremum over b by the expectation over a random variable B distributed uniformly over $\{0,1\}^V$:

$$\sup_{b \in \{0,1\}^V} \mathsf{E}_{Z^n}\left[R(\hat{f}_{Z^n})\right] \geq \mathsf{E}_B\left[\mathsf{E}_{Z^n}\left[R(\hat{f}_{Z^n})\right]\right].$$

It will be useful to rewrite the RHS in terms of a conditional probability, as we then can leverage properties of the Bayes risk of a decision problem:

$$\begin{aligned} \mathsf{E}_B\Big[\mathsf{E}_{Z^n}\left[R(\hat{f}_{Z^n})\right]\Big] &= \mathsf{E}\Big[\mathsf{E}\left[\mathbf{1}\left[\hat{f}_{Z^n}(X) \neq f_B(X)\right] \,\Big|\, Z^n, X\right]\Big] \\ &= \mathsf{E}\Big[\Pr\left(\hat{f}_{Z^n}(X) \neq f_B(X) \,\Big|\, Z^n, X\right)\Big]. \end{aligned} \tag{2}$$

Next, we analyze the conditional probability inside the expectation:

$$
\begin{aligned}
&\Pr\left(\hat{f}_{Z^n}(X) \neq f_B(X) \mid Z^n, X\right) \\
&= \mathbf{1}\left[\hat{f}_{Z^n}(X) = 0\right] \cdot \Pr\left(f_B(X) = 1 \mid Z^n, X\right) \\
&\quad + \mathbf{1}\left[\hat{f}_{Z^n}(X) = 1\right] \cdot \Pr\left(f_B(X) = 0 \mid Z^n, X\right) \\
&\geq \min\{\Pr\left(f_B(X) = 1 \mid Z^n, X\right), 1 - \Pr\left(f_B(X) = 1 \mid Z^n, X\right)\} \\
&= \min\{\eta(Z^n, X), 1 - \eta(Z^n, X)\},
\end{aligned} \tag{3}
$$

where $\eta(Z^n, X) = \Pr\left(f_B(X) = 1 \mid Z^n, X\right))$. From the last line above, we can see that we have arrived at a quantity that is completely analogous to the (conditional) Bayes risk, where the conditioning is on X (as usual) but now also Z^n.

It remains to lower bound the expectation of (3); let's first get a handle on $\eta(Z^n, X)$. Suppose that $X \in \{X_1, \dots, X_n, x_0\}$; then the label of X is known and hence $\eta(Z^n, X)$ is equal to either 0 or 1. On the other hand, if $X \notin \{X_1, \dots, X_n, x_0\}$, then, among the distributions in $\mathcal{P}_V$ that are consistent with the labeling of $X_1, \dots, X_n$, precisely half label X as 1 and half label X as 0, so in this case we have $\eta(Z^n, X) = \frac{1}{2}$. It therefore follows that (3) is equal to

$$
\frac{1}{2}\,\mathbf{1}\left[X \notin \{X_1, \dots, X_n, x_0\}\right],
$$

and hence (2) is equal to

$$
\frac{1}{2}\Pr\left(X \notin \{X_1, \dots, X_n, x_0\}\right).
$$

Considering the V possible values of X (as x_0 is excluded in the above event), this probability is

$$
\frac{1}{2}\sum_{j=1}^{V}\Pr(X = x_j)\prod_{i=1}^{n}\Pr(X_i \neq x_j) = \frac{1}{2}\sum_{j=1}^{V}\varepsilon(1-\varepsilon)^n = \frac{V}{2}\varepsilon(1-\varepsilon)^n.
$$

Next, setting $\varepsilon = \frac{1}{n}$ yields

$$
\frac{V}{2n}\left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n}\right)^{n-1}.
$$

The result follows since $\left(1 - \frac{1}{n}\right)^{n-1} \geq \frac{1}{e}$. To see this, note that this inequality is equivalent to

$$
(n-1)\log\left(1 - \frac{1}{n}\right) \geq -1 \quad \iff \quad \frac{1}{n-1} \geq \log\left(\frac{n}{n-1}\right) \quad \iff \quad e^{\frac{1}{n-1}} \geq \frac{n}{n-1},
$$

and the claim follows from $\frac{n}{n-1} = 1 + \frac{1}{n-1}$ and the inequality $e^x \geq 1 + x$. □

3 Lower bound, agnostic setting

A similar lower bound can be worked out in the agnostic case.

Theorem 3. *There are constants $c_1, c_2 > 0$ such that, for any $\mathcal{F}$ satisfying* $\mathrm{VCdim}(\mathcal{F}) = V$*, for any learning method $\hat{f}$, there exists a distribution P over $\mathcal{X} \times \mathcal{Y}$ for which*

$$\Pr\left(R(\hat{f}) - R(f^*) > c_1\sqrt{\frac{V}{n}}\right) > c_2.$$

4 Lower bounds on the expected risk actually tell you a lot of about the achievable high probability upper bounds on the risk

Suppose someone approaches you on the street and says that they have a learning algorithm for which, under any distribution $P \in \mathcal{P}_{\mathcal{F}}$ (i.e. the realizable case), satisfies for some $A, c > 0$

$$\Pr\left(R(\hat{f}) > \varepsilon\right) \leq Ae^{-cn\varepsilon}. \tag{4}$$

Should you believe them? Well, if their claim is true, then, for any $\gamma \geq 0$,

$$\begin{aligned}
\mathsf{E}[R(\hat{f})] &= \int_0^1 \Pr(R(\hat{f}) > \varepsilon)d\varepsilon \\
&\leq \gamma + \int_\gamma^1 \Pr(R(\hat{f}) > \varepsilon)d\varepsilon \\
&\leq \gamma + A\int_\gamma^1 e^{-cn\varepsilon}d\varepsilon \\
&= \gamma + \frac{A}{cn}\left(e^{-cn\gamma} - e^{-n}\right) \\
&\leq \gamma + \frac{A}{cn}e^{-cn\gamma}.
\end{aligned}$$

Taking $\gamma = \frac{\log A}{cn}$ yields the upper bound

$$\mathsf{E}[R(\hat{f})] \leq \frac{\log A}{cn} + \frac{1}{cn}.$$

In light of Theorem 2, it must be the case that

$$\frac{\log A}{cn} + \frac{1}{cn} \geq \frac{V}{2en}\left(1 - \frac{1}{n}\right),$$

and so

$$A \geq \exp\left(\frac{cV}{2e}\left(1 - \frac{1}{n}\right) - 1\right).$$

Thus, unavoidably, A must depend on the VC dimension in a bound of the form (4). That is to say, if the street person did not choose A to be exponential large in V, then they must be lying!

References

Luc Devroye, László Györfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 1996.