Machine Learning Theory (CSC 431/531) - Lecture 9

Nishant Mehta

1 Learnability and VC dimension

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 $\text{VCdim}(\mathcal{F}) = V$.

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 $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P$;
- 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[1]

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

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.[2] 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 (when outputting hypotheses in $\mathcal{F}$) 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?

The answer is yes! The VC dimension thus *characterizes* the classes $\mathcal{F}$ for which learnability holds.

[1] I did not prove this result in class, but you can find a proof in the previous set of lecture notes.

[2] 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 Minimax lower bounds in the realizable and agnostic cases

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 see a proof of this result here, but I'll provide some additional lecture notes if you are interested in seeing a proof of a related result (a lower bound on the expected risk). Seeing how lower bounds are proved can be very valuable; the techniques employed often greatly differ from those used to prove upper bounds.

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

Theorem 2. *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*

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

3 Compression schemes

Definition 1. A compression scheme with kernel size k for a concept class $\mathcal{C}$ is specified by a choice of *compression map* κ and a *reconstruction map* ρ where:

- κ takes as input a labeled sample of size n (for any n) that is labeled by a concept in $\mathcal{C}$, and κ then outputs a subsequence of at most k examples;
- ρ takes as input a labeled sample of size at most k labeled by a concept in $\mathcal{C}$, and ρ then outputs a hypothesis in $\{0,1\}^{\mathcal{X}}$;
- we further have that for any sample S labeled by a concept $c \in \mathcal{C}$, the hypothesis $\hat{f} = \rho(\kappa(S))$ is consistent with S.

Theorem 3. *Let (κ, ρ) be a compression scheme for $\mathcal{C}$ with kernel size k. Suppose that $X_1, \ldots, X_n$ are drawn independently from P and labeled according to some concept $c \in \mathcal{C}$, yielding a labeled sample S. Let $\hat{f}_{\kappa,\rho}$ denote the hypothesis defined as $\hat{f}_{\kappa,\rho} = \rho(\kappa(S))$. Then with probability at least $1 - \delta$,*

$$R(\hat{f}_{\rho,\kappa}) \leq \frac{k\log(ne/k) + \log\frac{1}{\delta}}{n-k}.$$

Proof. Throughout the proof, we use the notation Z_j to denote labeled example (X_j, Y_j). Note that $S = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$.

Let $\mathcal{T}$ be the set of all subsets of $\{1, \ldots, n\}$ whose cardinality is at most k. Consider a fixed subset $I \in \mathcal{T}$, and let $\mathbf{Z}_I$ denote the subsequence of examples $(Z_j)_{j \in I}$. Corresponding to I is a hypothesis

$$\hat{f}_{I,\rho} = \rho(\mathbf{Z}_I).$$

Observe that $\hat{f}_{I,\rho}$ depends only on $\mathbf{Z}_I$. Since $\mathbf{Z}_I$ is independent of $\mathbf{Z}_{[n]\setminus I}$, it also is true that $\hat{f}_{I,\rho}$ is independent of $\mathbf{Z}_{[n]\setminus I}$. Now, for each fixed I, the probability that $\hat{f}_{I,\rho}$ has risk more than ε and yet is consistent with c on $\mathbf{X}_{[n]\setminus I}$ is at most

$$(1-\varepsilon)^{n-k}. \tag{2}$$

Next, we extend the above argument to hold simultaneously for all possible choices of I from $\mathcal{T}$, which includes the random, data-dependent choice that results froms our actual training sample. To this end, note that the cardinality of $\mathcal{T}$ is $\sum_{j=0}^{k}\binom{n}{j} \leq \left(\frac{ne}{k}\right)^k$ from Lemma 2 in the previous set of lecture notes. Therefore, the probability that $\hat{f}_{\kappa,\rho}(x)$ has risk more than ε and yet is consistent with $\mathbf{Z}^n$ is at most

$$\left(\frac{ne}{k}\right)^k (1-\varepsilon)^{n-k} \leq \left(\frac{ne}{k}\right)^k e^{-(n-k)\varepsilon}.$$

The result follows by inversion (set the RHS to δ). □

Definition 2. An agnostic compression scheme with kernel size k for a concept class $\mathcal{C}$ is specified by a choice of *compression map* κ and a *reconstruction map* ρ where:

- κ takes as input a labeled sample of size n (for any n), and κ then outputs a subsequence of at most k examples;
- ρ takes as input a labeled sample of size at most k, and ρ then outputs a hypothesis in $\{0,1\}^{\mathcal{X}}$;
- we further have that for any sample S, the hypothesis $\hat{f} = \rho(\kappa(S))$ satisfies $\hat{R}(\hat{f}$ is at most $\min_{f\in\mathcal{C}} \hat{R}(f)$. That is, the empirical risk of $\hat{f}$ on the sample S is no larger than that of the empirical risk minmizer over $\mathcal{C}$.

Lemma 1. *If there is a compression scheme for $\mathcal{C}$, then there is also an agnostic compression scheme.*

Proof. Let (κ, ρ) be a compression scheme for $\mathcal{C}$. Given a labeled sample of n examples, use ERM over $\mathcal{C}$ to obtain $\hat{f}_{\mathrm{ERM}}$. Discard the examples where ERM is incorrect. On the remaining examples, ERM is consistent, and so we may apply our compression scheme for $\mathcal{C}$ on these examples. The resulting hypothesis $\hat{f}_{\rho,\kappa}$ will also be correct on these examples, and it can only possibly do better than ERM on the removed examples. Hence, its empirical risk is upper bounded by the empirical risk of ERM. (In addition, we use a proper compression scheme, then it must be the case that the empirical risk of $\hat{f}_{\rho,\kappa}$ is equal to the empirical risk of ERM!). □