CPSC 532D — 4. PAC LEARNING; INFINITE $\mathcal{H}$

Danica J. Sutherland
University of British Columbia, Vancouver
Fall 2023

As a reminder, in lecture 2 we proved the following:

PROPOSITION 1. *Suppose $\ell(z, h)$ is almost surely bounded in $[a, b]$, $\mathcal{H}$ is finite, and $\hat{h}_S$ is any empirical risk minimizer over the set $\mathcal{H}$ based on a sample $S = (z_1, \dots, z_m)$. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^m$ it holds that*

$$L_{\mathcal{D}}(\hat{h}_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) \le (b - a)\sqrt{\frac{2}{m} \log \frac{|\mathcal{H}| + 1}{\delta}}.$$

Proof. For any ERM and any $\mathcal{H}$, it holds that

$$\begin{aligned} L_{\mathcal{D}}(\hat{h}_S) &\le L_S(\hat{h}_S) + \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] \\ &\le L_S(h^*) + \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] \\ &\le L_{\mathcal{D}}(h^*) + \left[L_S(h^*) - L_{\mathcal{D}}(h^*)\right] + \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)]. \end{aligned} \tag{1}$$

The result follows by applying Hoeffding's inequality to $L_S(h^*) - L_{\mathcal{D}}(h^*)$ and $L_{\mathcal{D}}(h) - L_S(h)$ for all $h \in \mathcal{H}$. □

Another way to state this result is that with m samples, we can achieve statistical error at most ε with probability at least $(|\mathcal{H}| + 1) \exp\left(-\frac{m\varepsilon^2}{2(b-a)^2}\right)$.

Or, alternately, we can say that we can achieve excess error at most ε with probability at least $1 - \delta$ if we have at least $\frac{2(b-a)^2}{\varepsilon^2} \log \frac{|\mathcal{H}|+1}{\delta}$ samples. This last way establishes the *sample complexity* of learning to a given accuracy ε with a given confidence $1 - \delta$.

1 PAC LEARNING

This last way corresponds to one of the standard notions of learnability:

DEFINITION 2. An algorithm $\mathcal{A}$ *agnostically PAC learns* $\mathcal{H}$ with a loss ℓ if there exists a function $m : (0, 1)^2 \to \mathbb{N}$ such that, for every $\varepsilon, \delta \in (0, 1)$, for every distribution $\mathcal{D}$ over $\mathcal{Z}$, for any $m \ge m(\varepsilon, \delta)$, we have that

$$\Pr_{S \sim \mathcal{D}^m}\left(L_{\mathcal{D}}(\mathcal{A}(S)) > \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon\right) < \delta.$$

That is, $\mathcal{A}$ can *probably* get an *approximately correct* answer, where "correct" means the best possible error in $\mathcal{H}$.

If $\mathcal{A}$ runs in time polynomial in $1/\varepsilon$, $1/\delta$, n, and some notion of the size of h^*, then we say that A *efficiently agnostically PAC learns* $\mathcal{H}$.

For more, visit https://cs.ubc.ca/~dsuth/532D/23w1/.

DEFINITION 3. A hypothesis class $\mathcal{H}$ is *agnostically PAC learnable* if there exists an algorithm $\mathcal{A}$ which agnostically PAC learns $\mathcal{H}$.

So, ERM agnostically PAC-learns finite hypothesis classes, with the sample complexity $m(\varepsilon, \delta) = \frac{2(b-a)^2}{\varepsilon^2} \log \frac{|\mathcal{H}|+1}{\delta}$. Notice that in the definition of agnostic PAC learning, there's no limitation on the distribution – there needs to be an $m(\varepsilon, \delta)$ that works for *any* $\mathcal{D}$. Proposition 1 satisfies this, but in general, it's an extremely worst-case kind of notion.

Often it's nicer to think about cases where we can make some assumptions on $\mathcal{D}$. For example, maybe the number of samples you need depends on "how hard" the particular problem is. We'll talk about this more a little later in the course. For now, it's worth mentioning one common special case:

A1 Q3 was partly about this setting.

DEFINITION 4. Consider a nonnegative loss $\ell(h, z) \geq 0$. A distribution $\mathcal{D}$ is called *realizable* by $\mathcal{H}$ if there exists an $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$.

This version is the "privileged" version that doesn't need a modifier because it's the one that was introduced first [Val84].

DEFINITION 5. An algorithm $\mathcal{A}$ *PAC learns* $\mathcal{H}$ with a loss ℓ if there exists a function $m : (0, 1)^2 \to \mathbb{N}$ such that, for every $\varepsilon, \delta \in (0, 1)$, for every *realizable* distribution $\mathcal{D}$ over $\mathcal{Z}$, for any $m \geq m(\varepsilon, \delta)$, we have that

$$\Pr_{S \sim \mathcal{D}^m} (L_{\mathcal{D}}(\mathcal{A}(S)) > \varepsilon) < \delta.$$

That is, $\mathcal{A}$ can *probably* get an *approximately correct* answer, where "correct" means zero loss.

If $\mathcal{A}$ runs in time polynomial in $1/\varepsilon$, $1/\delta$, n, and some notion of the size of h^*, then we say that A *efficiently (realizably) PAC learns* $\mathcal{H}$.

DEFINITION 6. A hypothesis class $\mathcal{H}$ is *PAC learnable* if there exists an algorithm $\mathcal{A}$ which PAC learns $\mathcal{H}$.

Sometimes people say "realizable PAC learnable" or similar, to emphasize the difference versus agnostic PAC. The name "agnostic" is because the definition doesn't care whether there's a perfect h^* or not. (Notice that if $\mathcal{A}$ agnostically PAC learns $\mathcal{H}$, then it also PAC learns $\mathcal{H}$.)

The emphasis here on "how many samples for a given error" is also kind of a TCS-style framing, whereas statisticians more often ask "how much error for a given number of samples"; I tend to prefer the latter, but it's all equivalent.

If you read [SSBD] or other work by computational learning theorists, there tends to be a lot of focus on just being learnable versus not being learnable. That problem has been solved, though, as we'll see not too much later in class; recent work focuses much more on rates than on just learnability or not, and tends to be willing to make *some* assumptions on $\mathcal{D}$ rather than either being totally general or assuming only realizability.

2 Logistic Regression

We've shown that anything finite is agnostically PAC learnable. That's only an upper bound, though; it *doesn't* mean that infinite things aren't learnable. Which is good, because that's what we usually want to learn!

Lemma 6.1 of [SSBD] gives a really simple example of realizably PAC learning an infinite class, if you're curious to see that style of proof. I tried to do an agnostic version of that, but it was more complicated than I hoped, so let's do something more interesting instead.

In *logistic regression*, our data is in a subset of $\mathbb{R}^d$, our labels are in $\mathcal{Y} = \{-1, 1\}$ and we try to predict with a confidence score in $\widehat{\mathcal{Y}} = \mathbb{R}$. Our predictors are linear functions of the form $h_w(x) = w \cdot x$, and the logistic loss is given by

This is more convenient than $\mathcal{Y} = \{0, 1\}$ here...

You usually want an intercept term, $w \cdot x + w_0$, but you can achieve that by padding x with an always-one dimension.

$$\ell_{log}(h, (x, y)) = l_{log}(h(x), y) = \log(1 + \exp(-h(x)y)). \tag{2}$$

We'll use the hypothesis class $\mathcal{H} = \{h_w = x \mapsto w \cdot x : w \in \mathbb{R}^d, \|w\| \leq \mathrm{B}\}$ for some constant B; this avoids overfitting by using really-really complex w, and is basically equivalent to doing L_2-regularized logistic regression (we'll talk about this more later). This $\mathcal{H}$ is still infinite, but it has finite volume.

Now, our analysis is going to be based on the idea that if w and v are similar predictors, i.e. $h_w(x) \approx h_v(x)$ for all x, then they'll behave similarly: $\mathrm{L}_{\mathcal{D}}(h_w) \approx \mathrm{L}_{\mathcal{D}}(h_v)$ and $\mathrm{L}_{\mathrm{S}}(h_w) \approx \mathrm{L}_{\mathrm{S}}(h_v)$. Thus we don't have to do a totally separate concentration bound on their empirical risks; we can exploit that they're similar.

To formalize that, we'll want to bound

$$|\mathrm{L}_{\mathcal{D}}(h_w) - \mathrm{L}_{\mathcal{D}}(h_v)| \leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left| l(h_w(x), y) - l(h_v(x), y) \right|. \tag{3}$$

We can use the following result about the *Lipschitz constant* of l_{log}:

LEMMA 7. *For any $y \in \{-1, 1\}$ and $\hat{y}_1, \hat{y}_2 \in \mathbb{R}$, $\left| l_{log}(\hat{y}_1, y) - l_{log}(\hat{y}_2, y) \right| \leq \left| \hat{y}_1 - \hat{y}_2 \right|$.*

Proof. Let $l_y(\hat{y}) = l_{log}(\hat{y}, y)$. l_y is differentiable, and

$$\begin{aligned} \left| l_y'(\hat{y}) \right| &= \left| \frac{\mathrm{d}}{\mathrm{d}\hat{y}} \log(1 + \exp(-y\hat{y})) \right| = \left| \frac{1}{1 + \exp(-y\hat{y})} \exp(-y\hat{y})(-y) \right| \\ &= \left| \frac{\exp(-y\hat{y})}{1 + \exp(-y\hat{y})} \times \frac{\exp(y\hat{y})}{\exp(y\hat{y})} \right| \left| -y \right| = \left| \frac{1}{1 + \exp(y\hat{y})} \right| \leq 1. \end{aligned}$$

Thus l_y is 1-Lipschitz:

$$\left| l_y(\hat{y}_2) - l_y(\hat{y}_1) \right| = \left| \int_{\hat{y}_1}^{\hat{y}_2} l_y'(t) \mathrm{d}t \right| \leq \int_{\hat{y}_1}^{\hat{y}_2} \left| l_y'(t) \right| \mathrm{d}t \leq \int_{\hat{y}_1}^{\hat{y}_2} \mathrm{d}t = \left| \hat{y}_1 - \hat{y}_2 \right|. \qquad \square$$

Plugging this into (3), we get

$$|\mathrm{L}_{\mathcal{D}}(h_w) - \mathrm{L}_{\mathcal{D}}(h_v)| \leq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} |h_w(x) - h_v(x)|.$$

If the predictions are similar, the losses are too.

We can further say that if w and v are close, then their predictions are similar:

$$|h_w(x) - h_v(x)| = |w \cdot x - v \cdot x| = |(w - v) \cdot x| \leq \|w - v\| \, \|x\|$$

by Cauchy-Schwarz. Thus

$$|\mathrm{L}_{\mathcal{D}}(h_w) - \mathrm{L}_{\mathcal{D}}(h_v)| \leq \left(\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \|x\| \right) \|w - v\|.$$

For simplicity, let's assume that $\Pr_{(x,y)\sim\mathcal{D}}(\|x\| > C) = 0$, obtaining

$$|L_\mathcal{D}(h_w) - L_\mathcal{D}(h_v)| \le C\,\|w - v\|. \tag{4}$$

The same kind of thing is true for L_S; we could repeat the argument with averages instead of $\mathbb{E}$, or we could use the *empirical distribution* $\hat{\mathcal{D}}$ corresponding to S, the discrete distribution that puts $1/m$ probability at each member of S, and note that expectations over $\hat{\mathcal{D}}$ are exactly averages over S. Either way,

$$|L_S(h_w) - L_S(h_v)| \le \left(\frac{1}{m}\sum_{i=1}^{m}\|x_i\|\right)\|w - v\| \le C\,\|w - v\|. \tag{5}$$

Now, how do we exploit that similar hypotheses have similar losses? We'll use the following concept:

DEFINITION 8. A ρ-cover of a set U is a set $T \subseteq U$ such that, for all $u \in U$, there is a $t \in T$ with $\operatorname{dist}(t, u) \le \rho$.

We're going to use a set cover for $\{w : \|w\| \le B\}$ based on the Euclidean distance, and then use (4) and (5) to turn that into a set cover for $\mathcal{H}$.

Let $N(B, \rho)$ be the size of the smallest cover for $\mathcal{H}$. We have the following result (proved in Section 2.1):

For $\rho \ge B$, you immediately get $N(B, \rho) = 1$.

LEMMA 9. *Let $B \ge \rho > 0$. The covering number $N(B, \rho)$ of the radius-B Euclidean ball in $\mathbb{R}^d$, $\{x \in \mathbb{R}^d : \|x\| \le B\}$, satisfies $N(B, \rho) \le (3B/\rho)^d$.*

We now have all the tools we need for the following result.

PROPOSITION 10. *Let $h_w(x) = w \cdot x$ and $\mathcal{H} = \{h_w : \|w\| \le B\}$ for some $B > 0$. Consider the logistic loss given by (2), and assume that $\|x\| \le C$ almost surely under $\mathcal{D}$. Assume for simplicity $BC \ge 1$. Then, with probability at least $1 - \delta$,*

$$\sup_{h\in\mathcal{H}} L_\mathcal{D}(h) - L_S(h) \le \frac{2BC}{\sqrt{m}}\left[1 + \sqrt{\log\frac{1}{\delta} + \frac{d}{2}\log(9m)}\right].$$

Proof. Our proof will be of the form sometimes called an "ε-net argument." We will choose a ρ-cover $T = \{w_1, \dots, w_{N(B,\rho)}\} \subset \{w \in \mathbb{R}^d : \|w\| \le B\}$, where ρ is a parameter to be set later. Then, for any $h_w \in \mathcal{H}$, let j be the index of the w_j closest to w, which can't be further than ρ away. Thus,

$$\begin{aligned}
\sup_{h\in\mathcal{H}} L_\mathcal{D}(h) - L_S(h) &= \sup_{h\in\mathcal{H}} L_\mathcal{D}(h) - L_\mathcal{D}(h_j) + L_\mathcal{D}(h_j) - L_S(h_j) + L_S(h_j) - L_S(h) \\
&\le \underbrace{\sup_{h\in\mathcal{H}} L_\mathcal{D}(h) - L_\mathcal{D}(h_j)}_{\text{bound with (4)}} + \underbrace{\sup_{h_j\in T} L_\mathcal{D}(h_j) - L_S(h_j)}_{\text{as in Proposition 1}} + \underbrace{\sup_{h\in\mathcal{H}} L_S(h_j) - L_S(h)}_{\text{bound with (5)}}.
\end{aligned}$$

The first and last terms are each $C\rho$.

The middle term is uniform convergence over a finite $\mathcal{H}$, as in Proposition 1. There's one catch, though: the logistic loss isn't "naturally" bounded. But given that $\|x\| \le C$

and $\|w\| \le B$, we know that $|h(x)| = |w \cdot x| \le BC$. Thus

$$\left|\ell(h,(x,y))\right| = \left|\log(1+\exp(-yh(x))\right| \le \log(1+\exp(BC)) \le BC+1. \tag{6}$$

Then we can apply Hoeffding to each element of T, giving it a failure probability of $\delta/N(B,\rho)$, and obtaining that with probability at least $1-\delta$,

$$\begin{aligned}\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] &\le 2C\rho + (BC+1)\sqrt{\frac{1}{2m}\log\frac{N(B,\rho)}{\delta}} \\ &\le 2C\rho + (BC+1)\sqrt{\frac{1}{2m}\left[\log\frac{1}{\delta} + d\log\frac{3B}{\rho}\right]}.\end{aligned}$$

Now, we could try to exactly optimize the value of ρ by setting the derivative to zero, but I think we won't be able to solve that equation. Instead, let's notice that if ρ is $o(1/\sqrt{m})$, the first term being smaller doesn't really help in rate since the other two are $1/\sqrt{m}$ anyway – but choosing a smaller ρ makes the $\log\frac{1}{\rho}$ worse. Also, the dependence on ρ there is only in a log term, so it's probably okay-ish to choose $\rho = \alpha/\sqrt{m}$, giving

$$\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] \le \frac{1}{\sqrt{m}}\left[2C\alpha + \frac{BC+1}{\sqrt{2}}\sqrt{\log\frac{1}{\delta} + d\log\frac{3B\sqrt{m}}{\alpha}}\right].$$

Picking $\alpha = B$ gives

$$\sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(h) - L_S(h)] \le \frac{BC}{\sqrt{m}}\left[2 + \frac{1+1/(BC)}{\sqrt{2}}\sqrt{\log\frac{1}{\delta} + \frac{d}{2}\log(9m)}\right],$$

and the desired result follows from $1/(BC) \le 1$ and $2/\sqrt{2} < 2$. □

Treating everything but m as a constant, the rate is $\mathcal{O}_p\left(\sqrt{\frac{\log m}{m}}\right)$. That $\sqrt{\log m}$ factor is actually unnecessary, but getting rid of it with covering number-type arguments requires some more advanced machinery (called "chaining"; we *might* cover it later in class). Instead, next time we'll see a simpler way to show a $\mathcal{O}_p(1/\sqrt{m})$ rate that will also be very generally applicable.

We only wrote this proof here for $\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$, but since the loss is a.s. bounded, this implies exactly as in (1) an upper bound on the generalization error of any ERM $\hat{h}_S$:

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le (BC+1)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}} + \frac{2BC}{\sqrt{m}}\left[1 + \sqrt{\log\frac{2}{\delta} + \frac{d}{2}\log(9m)}\right],$$

which using the assumption $BC \ge 1$, $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, and $1/\sqrt{2} < 1$ we can simplify further as

$$L_{\mathcal{D}}(\hat{h}_S) - L_{\mathcal{D}}(h^*) \le \frac{2BC}{\sqrt{m}}\left[1 + 2\sqrt{\log\frac{2}{\delta}} + \sqrt{\frac{d}{2}\log(9m)}\right].$$

General case We needed the following properties about the problem to get this result:

- A bounded loss, for Hoeffding, here implied by (6).

- A Lipschitz loss, to get (4) and (5).
- A parameterization for $\mathcal{H}$ with a covering number bound.

The first point is general, and could e.g. be immediately weakened to sub-Gaussianity if you have that another way than boundedness. If you have some other way to show concentration for a finite number of points, or for expectation bounds, you don't necessarily need this.

The second point, of a Lipschitz loss, is definitely necessary in some form. You could potentially use a locally Lipschitz loss (where the constant varies through space), but then you have to be more careful.

The third point, of a covering number bound on $\mathcal{H}$, is also important. We framed this as covering the parameter set, but you could also think of it as defining a distance metric on $\mathcal{H}$ (by $\operatorname{dist}(h_w, h_v) = \|w - v\|$) and then covering $\mathcal{H}$. This generality is often useful, e.g. for nonparametric $\mathcal{H}$.

2.1 *Bounds on covering numbers*

We'll now prove our upper bound on covering numbers. Recall their definition:

DEFINITION 8. A ρ-cover of a set U is a set $T \subseteq U$ such that, for all $u \in U$, there is a $t \in T$ with $\operatorname{dist}(t, u) \le \rho$.

We used $N(B, \rho)$ to be the size of the smallest ρ-cover for the B-ball $\{w \in \mathbb{R}^d : \|w\| \le B\}$.

We'll also need the idea of *packing numbers*: how many balls can we squeeze into a set T?

DEFINITION 11. A ρ-packing of a set U is a set $T \subseteq U$ such that, for all $t, t' \in T$ with $t \ne t'$, we have $\operatorname{dist}(t, t') \ge \rho$.

Let $M(B, \rho)$ be the size of the largest possible ρ-packing of the B-ball.

PROPOSITION 12. *A maximal ρ-packing of a set U is also a ρ-cover of* T.

Proof. Suppose there were some point $u \in U$ such that $\operatorname{dist}(u, t) > \rho$ for all $t \in T$. Then we could add u to the ρ-packing, producing a packing of size one larger; this contradicts that T was maximal. □

We're now ready to prove the result:

For $\rho \ge B$, you immediately get $N(B, \rho) = 1$.

LEMMA 9. *Let $B \ge \rho > 0$. The covering number $N(B, \rho)$ of the radius-B Euclidean ball in $\mathbb{R}^d$, $\{x \in \mathbb{R}^d : \|x\| \le B\}$, satisfies $N(B, \rho) \le (3B/\rho)^d$.*

Proof. By Proposition 12, we have that $N(B, \rho) \le M(B, \rho)$; we'll actually prove the result about M.

Let T be a maximal ρ-packing of the B-ball $\{w \in \mathbb{R}^d : \|w\| \le B\}$. Thus the open $\rho/2$-balls centered at each $t \in T$, $\{w \in \mathbb{R}^d : \|w - t\| < \rho/2\}$, are disjoint: if they weren't, you could get from one t to another in distance less than ρ. These balls are also all contained within the ball of radius $(B + \rho/2)$, since each t has norm at most B. Thus we have that

$$\sum_{t \in T} \operatorname{vol}\left(\{w \in \mathbb{R}^d : \|w - t\| < \rho/2\}\right) \leq \operatorname{vol}\left(\{w \in \mathbb{R}^d : \|w\| < B + \rho/2\}\right).$$

But we know that the volume of a ball of radius R in d dimensions is $R^d V_1$, where $V_1 = \operatorname{vol}(\{w \in \mathbb{R}^d : \|w\| < 1\})$. Thus

$$\sum_{t \in T} \left(\frac{\rho}{2}\right)^d V_1 = M(B, \rho)\left(\frac{\rho}{2}\right)^d V_1 \leq \left(B + \frac{\rho}{2}\right)^d V_1,$$

and so

$$M(B, \rho) \leq \left(\frac{2B}{\rho} + 1\right)^d = \left(\frac{2B + \rho}{\rho}\right)^d \leq \left(\frac{3B}{\rho}\right)^d,$$

using at the end that $\rho \leq B$ to get a simpler form. □

References

[SSBD] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[Val84] Leslie G. Valiant. "A Theory of the Learnable." *Communications of the ACM* 27.11 (1984), pages 1134–1142.