CSCI 5525 Machine Learning Fall 2019

Lecture 11: Learning Theory

Oct 8th 2019

Lecturer: Steven Wu *Scribe: Steven Wu*

Now let's do some theory. Recall that in supervised learning, we have data (x_i, y_i)'s drawn from a distribution P over labelled examples (X, Y). For any predictor f in function class $\mathcal{F}$ (e.g. linear models, neural networks), the empirical risk is defined as

$$\hat{\mathcal{R}}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

However, since we care about the predictive performance on future instances, the primary objective should be *true or population risk* of f:

$$\mathcal{R}(f) = \mathop{\mathbf{E}}_{P} [\ell(Y, f(X))]$$

The *excess risk* of f is then defined as

$$\mathcal{R}(f) - \hat{\mathcal{R}}(f)$$

Last lecture, we gave some algorithms for solving the ERM problem: $\min_f \hat{\mathcal{R}}(f)$. Now let's say the algorithm returns $\hat{f}$, which approximately solves the problem. Let f^* be the minimizer of the true risk. Then the differencde between the true risks of $\hat{f}$ and f^* can be written as

$$\begin{aligned} \mathcal{R}(\hat{f}) - \mathcal{R}(f^*) &= \hat{\mathcal{R}}(f^*) - \mathcal{R}(f^*) && \text{(sampling error)} \\ &+ \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f^*) && \text{(approximation error)} \\ &+ \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) && \text{(generalization error)} \end{aligned}$$

How do we bound these things?

- First, sampling error is easiest. We know that $(x_1, y_1), \ldots, (x_n, y_n)$ are drawn i.i.d. from P. By law of large numbers,
$$\hat{\mathcal{R}}(f^*) \to \mathcal{R}(f^*)$$
We will be formal about how fast the convergence is in a bit.
- Secondly, approximation error depends on how good the ERM algorithm is.
- The tricky part is the generalization error. It's tempted to direclty apply the same reasoning of the law of large numbers. Not quite. The output $\hat{f}$ is also a random variable that depends on the data. Recall the following predictor with no restriction.
$$\hat{f}(X) = \begin{cases} y_i, & \text{if } X = x_i \\ \text{"Gopher!"}, & \text{otherwise} \end{cases}$$

For this extreme predictor, we don't really expect fast convergence of $\hat{\mathcal{R}}(\hat{f})$ to $\mathcal{R}(f)$. The main issue is that we are not restricting the *complexity* of the predictor function at all.

1 Bounding Sampling Error

Now let's be formal about how we bound the sampling error. We will focus on the special case where the loss ℓ corresponds to the the zero-one loss in classification, and so for each fixed predictor f,

$$\hat{\mathcal{R}}_{01}(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[f(x_i) \neq y_i] \qquad \mathcal{R}_{01}(f) = \mathop{\mathbf{E}}_{(X,Y)\sim P}[\mathbf{1}[f(X) \neq Y]]$$

How fast does $\hat{\mathcal{R}}_{01}$ converge to $\mathcal{R}_{01}$ as a function of n? We will use the following *concentration inequality*:

Theorem 1.1 (Chernoff/Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n \in [a, b]$ be i.i.d. real-valued random variables drawn from a distribution D. Then*

$$\mathbf{Pr}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i - \mathbf{E}\left[Z_i\right] \geq \epsilon\right] \leq \exp\left(\frac{-2n\epsilon^2}{(b-a)^2}\right)$$

Equivalently, we can also state that for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\mathbf{E}\left[Z\right] \leq \left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) + (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}$$

You can have each Z_i as $\mathbf{1}[f(x_i) \neq y_i]$, and so we can have the following probabilistic statement: with probability at least $1 - \delta$ over the i.i.d. draws of the examples $(x_1, y_1), \ldots, (x_n, y_n)$, the following holds

$$\mathcal{R}_{01}(f) \leq \hat{\mathcal{R}}_{01}(f) + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

Here we see the most basic complexity measure of a function class $\mathcal{F}$, which is the cardinality $|\mathcal{F}|$.

2 Finite class $\mathcal{F}$

Now suppose that we would like to get the same concentration bound for the predictor $\hat{f}$ returned by the ERM method. We can no longer directly use the Chernoff bound since conditioned the ERM output being $\hat{f}$, the collection of examples $(x_1, y_1), \ldots, (x_n, y_n)$ are no longer i.i.d. To avoid this, we will instead ask for something stronger called "uniform convergence", and show that the concentration holds for every predictor f in the function class $\mathcal{F}$. A useful tool is the *union bound*: for any two arbitrary events A and B,

$$\mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B]$$

Theorem 2.1 (Uniform convergence over finite class)**.** *Let $\mathcal{F}$ be a finite class of predictor functions. Then with probability $1-\delta$ over the i.i.d. draws of $(x_1, y_1) \dots (x_n, y_n)$, for all $f \in \mathcal{F}$*

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{\ln(|\mathcal{F}|/\delta)}{2n}}$$

Proof. For each $f \in \mathcal{F}$, let E_f be the event such that $\mathcal{R}(f) > \hat{\mathcal{R}}(f) + \epsilon_f$, where $\epsilon_f = \sqrt{\frac{\ln(1/\delta_f)}{2n}}$. By Chernoff bound, for each fixed f, $\mathbf{Pr}[E_f] \leq \delta_f$. Then $\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \epsilon_f$ for all f if the complements of all events E_f hold simultaneously.

$$\begin{aligned}
\mathbf{Pr}[\forall f,\ \mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \epsilon_f] &= 1 - \mathbf{Pr}[\exists f, E_f] \\
&= 1 - \mathbf{Pr}[\cup_f E_f] \\
&\geq 1 - \sum_f \mathbf{Pr}[E_f] \\
&\geq 1 - \sum_f \delta_f
\end{aligned}$$

To finish the proof, it suffices to set each $\delta_f = \delta/|\mathcal{F}|$. □

3 VC Dimension

What if $|\mathcal{F}|$ is infinite? In this case, we will replace $\ln(|\mathcal{F}|)$ by some complexity measure of the class $\mathcal{F}$. We will introduce the *Vapnik-Chervonenkis dimension* (VC dimension) of $\mathcal{F}$. We say that $\mathcal{F}$ *shatters* a set of points $x_1, \dots, x_n \in \mathcal{X}$ if $\mathcal{F}$ realizes all labelings over these n points. The VC dimension of $\mathcal{F}$ is the largest number of points $\mathcal{F}$ can shatter:

$$\text{VCD}(\mathcal{F}) = \max\{n \in \mathbb{Z} \colon \exists (x_1, \dots, x_n) \in \mathcal{X}^n, \forall (y_1, \dots, y_n) \in \{0,1\}^n, \exists f \in \mathcal{F}, f(x_i) = y_i\}$$

Claim 3.1. *If $\mathcal{F}$ is finite, then*

$$\textit{VCD}(\mathcal{F}) \leq \log(|\mathcal{F}|)$$

In homework and exam problems, you should prove two things to establish a VC dimension bound for a function class $\mathcal{F}$ is VCD($\mathcal{F}$): 1) an **upper bound** showing that no set of more than VCD($\mathcal{F}$)$+1$ points can be shattered by $\mathcal{F}$ and 2) a **lower bound** given by a set of VCD($\mathcal{F}$) points that can be shattered by $\mathcal{F}$. A couple examples are in order.

Example 3.2 (Intervals)**.** *The class of all intervals on the real line $\mathcal{F} = \{\mathbf{1}[x \in [a,b]] \mid a, b \in \mathbb{R}\}$ has VC dimension 2.*

Example 3.3 (Affine classifier)**.** *The class of all intervals on the real line $\mathcal{F} = \{\mathbf{1}[\langle a, x\rangle + b \geq 0] \mid a \in \mathbb{R}^d, b \in \mathbb{R}\}$ has VC dimension $d+1$. **Upper bound:** In convex analysis, Radon's theorem shows that any set of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint sets whose convex hulls intersect. **Lower bound:** the standard basis and the origin.*

With VC dimension as a complexity measure, we can obtain a uniform convergence result for infinite function classes $\mathcal{F}$.

Theorem 3.4 (Uniform convergence over bounded VC class)**.** *Suppose that the function class has bounded VC dimension. Then with probability* $1-\delta$ *over the i.i.d. draws of* $(x_1, y_1), \ldots, (x_n, y_n)$*, for all* $f \in \mathcal{F}$*,*

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \tilde{O}\left(\sqrt{\frac{\textit{VCD}(\mathcal{F}) + \ln(1/\delta)}{n}}\right)$$

where $\tilde{O}$ *hides some dependences on* $\log(\textit{VCD}(\mathcal{F}))$ *and* $\log(n)$*.*