CSCI 5525 Machine Learning Fall 2019

Lecture 14: Learning Theory (Part 3)

March 2020

Lecturer: Steven Wu *Scribe: Steven Wu*

1 Uniform Convergence

Previously, we talked about how to bound the generalization error of the ERM output. The key is to obtain *uniform convergence*.

Theorem 1.1 (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 $f \in \mathcal{F}$*

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

We can derive a similar result for the case where $|\mathcal{F}|$ is infinite, by essentially replacing $\ln(|\mathcal{F}|)$ by some complexity measure of the class $\mathcal{F}$. The complexity measure is called *Vapnik-Chervonenkis dimension* (VC dimension) of $\mathcal{F}$, which is the largest number of points $\mathcal{F}$ can shatter:

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

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

Theorem 1.2 (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 $(x_1, y_1), \dots, (x_n, y_n)$, 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 $\log(n)$.

During lecture 13, we saw two simple example function classes and their VC dimensions.

Example 1.3 (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 1.4 (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 $d+1$.*

We can also obtain VC dimension bound for neural networks, which depends on the choices of activation functions.

Example 1.5 (Neural networks)**.** *Consider the classifier given by neural networks: for each feature vector x, the prediction is given by*

$$f(x, \theta) = sgn\left[\sigma_L\left(W_L(\dots W_2\sigma_1(W_1x + b_1) + b_2 \dots) + b_L\right)\right]$$

Let ρ be the number of parameters (weights and biases), L be the number of layers, and m be the number of nodes. If we use the same activation for all σ_i, we can obtain:

- *Binary activation $\sigma(z) = \mathbf{1}[z \geq 0]$, VCD $= O(\rho \ln \rho)$.*
  (See Theorem 4 of this paper for a proof.)
- *ReLU activation $\sigma(z) = \max(0, z)$, VCD $= O(\rho L \ln(\rho L))$*
  (See Theorem 6 of this paper for a proof.)

Roughly speaking, the VC-dimension of a neural network scales with the number of parameters defining class $\mathcal{F}$. However, in practice, the number of parameters might exceed the number of training examples, so the generalization bound derived from VC dimension is often not very useful for deep nets.

Here is a simple example for which the VC dimension is very different from the number of parameters. Consider, for example, the domain $\mathcal{X} = \mathbb{R}$, and the class $\mathcal{F} = \{h_\theta \colon \theta \in \mathbb{R}$ where $h_\theta \colon \mathcal{X} \to \{0, 1\}$ is defined by $h_\theta(x) = \lceil 0.5 \sin(\theta x) \rceil$. It is possible to prove that $\text{VCD}(\mathcal{F}) = \infty$.

2 Rademacher Complexity

Well, VC dimension is designed for binary classification. How about other learning problems including multi-class classification and regression? There is actually a more general complexity measure.

Given a set of examples $S = \{z_1 \dots, z_n\}$ and function class $\mathcal{F}$, the *Rademacher complexity* is defined as

$$\text{Rad}(\mathcal{F}, S) = \mathop{\mathbf{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(z_i)$$

where each $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Rademacher random variables: $\mathbf{Pr}[\epsilon_i = 1] = \mathbf{Pr}[\epsilon_i = -1] = 1/2$. Why does Rademacher complexity capture the complexity of a function class? One intuition is that it captures the ability of $\mathcal{F}$ to fit random signs given by the Rademacher random variables. For any loss function $\ell \colon \mathcal{Y} \times \mathcal{Y}$ and predictor $f \in \mathcal{F}$, let $\ell \circ f$ be a function such that for any example $z = (x, y)$

$$\ell \circ f(z) = \ell(y, f(x))$$

Let the corresponding function class $\ell \circ \mathcal{F} = \{\ell \circ f \mid f \in \mathcal{F}\}$. Now we can derive the following generalization bound using Rademacher complexity.

Theorem 2.1. *Assume that for all $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and $f \in \mathcal{F}$ we have $|\ell(y, f(x)| \leq c$. Let $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ be i.i.d. draws from the underlying distribution P. Then with probability at least $1 - \delta$, for all $f \in \mathcal{F}$*

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + 2\textit{Rad}(\ell \circ \mathcal{F}, S) + 4c\sqrt{\frac{2\ln(4/\delta)}{n}}$$

Moreover, if ℓ is γ-Lipschitz in the second argument for all y, then $\textit{Rad}(\ell \circ \mathcal{F}, S) \leq \gamma\textit{Rad}(\mathcal{F}, S)$, and so

$$\mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + 2\gamma\textit{Rad}(\mathcal{F}, S) + 4c\sqrt{\frac{2\ln(4/\delta)}{n}}$$

Note that Rademacher complexity depends on the underlying data distribution. For simple function classes, we can obtain complexity bounds only by assuming boundedness in the data.

Example 2.2 (Linear predictors)**.** *Consider two classes of linear functions:*

$$\mathcal{F}_1 = \{x \to w^\intercal x \colon w \in \mathbb{R}^d, \|w\|_1 \leq W_1\}$$

$$\mathcal{F}_2 = \{x \to w^\intercal x \colon w \in \mathbb{R}^d, \|w\|_2 \leq W_2\}$$

Let $S = (x_1, ..., x_n)$ be vectors in $\mathbb{R}^d$.

$$\textit{Rad}(\mathcal{F}_1, S) \leq (\max_i \|x_i\|_\infty) W_1 \sqrt{\frac{2\log(2d)}{n}}$$

$$\textit{Rad}(\mathcal{F}_2, S) \leq (\max_i \|x_i\|_2) W_2 \sqrt{\frac{1}{n}}$$

For linear functions, a nice feature of Rademacher complexity is that it picks up explicit dependence on the norm bounds of the weight vectors. In comparison, the VC dimension for the class of affine functions is just $d + 1$.