CSC 588: Machine learning theory **Spring 2022**

Lecture 10: Uniform convergence for VC classes; Rademacher complexity

Lecturer: Chicheng Zhang *Scribe: Yuanyuan Sun*

1 Uniform Convergence Theorems

Theorem 1: Given a hypothesis class $\mathcal{H}$ with $VC(\mathcal{H}) = d$, a set of n iid training samples $(x_1, y_1), \ldots, (x_n, y_n)$ from D, with probability $1 - \delta$,

$$\sup_{h \in \mathcal{H}} |\operatorname{err}(h, S) - \operatorname{err}(h, D)| \leq c_1 \sqrt{\frac{\ln S(\mathcal{H}, n) + \ln \frac{1}{\delta}}{n}} \leq c_2 \sqrt{\frac{d \ln \frac{n}{d} + \ln \frac{1}{\delta}}{n}}$$

for constant $c_1, c_2 > 0$.

Theorem 2: Suppose $\mathcal{F} \subseteq (\mathcal{Z} \to \{0, 1\})$, $S = (z_1, \ldots, z_n)$ iid samples from distribution D, with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} |\mathbb{E}_S[f(Z)] - \mathbb{E}_D[f(Z)]| \leq \sqrt{\frac{32(\ln S(\mathcal{F}, n) + \ln \frac{4}{\delta})}{n}}$$

Reminder: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, z = (x, y)$, $\mathcal{F} = \{l_h : h \in \mathcal{H}\}$ with $l_h(x, y) = I(h(x) \neq y)$.

2 Example: Glivenko-Cantelli Theorem

Given the distribution D over $\mathbb{R}$ and iid samples $z_1, \ldots, z_n$ drawn from D. The empirical cumulative distribution function (CDF) is defined as

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(z_i \leq t)$$

The population CDF is defined as:

$$F(t) = \mathbb{P}_{z \sim D}(z \leq t)$$

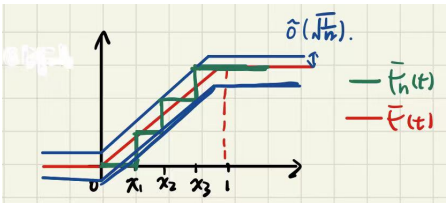

Figure 1: CDF and population CDF

Theorem 2 can be used to bound $\sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$ with probability $1 - \delta$.

Define $\mathcal{F} = \{h_t(x) = I(x \leq t) : t \in \mathbb{R}\}$. Then

$$F_n(t) = \mathbb{E}_S[h_t(Z)]$$
$$F(t) = \mathbb{E}_D[h_t(Z)]$$

According to Theorem 2

$$\begin{aligned}\sup_{t\in\mathbb{R}}|F_n(t)-F(t)| &= \sup_{t\in\mathbb{R}}|\mathbb{E}_S f(z)-\mathbb{E}_D f(z)| \\ &\le O\left(\sqrt{\frac{\ln\frac{1}{\delta}+\ln S(\mathcal{F},n)}{n}}\right) \\ &\le \tilde{O}\left(\sqrt{\frac{1}{n}}\right)\end{aligned}$$

3 Proof of uniform convergence theorem

Here we only show the upper concentration bound that with probability $1-\frac{\delta}{2}$,

$$\sup_{f\in\mathcal{F}}\mathbb{E}_S f(Z)-\mathbb{E}_D f(Z)\le\sqrt{\frac{32\left(\ln\frac{4}{\delta}+\ln S(\mathcal{F},n)\right)}{n}}=:\epsilon(\mathcal{F},n)$$

Exercise: The similar logic applies to show the lower concentration bound which is with probability $1-\delta/2$

$$\sup_{f\in\mathcal{F}}(-\mathbb{E}_S f(z))-(-\mathbb{E}_D f(z))\le\epsilon(\mathcal{F},n)$$

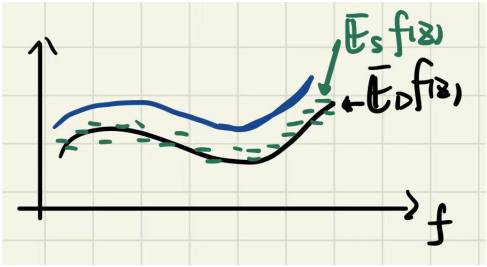

Figure 2: upper concentration bound

Lemma 1 With probability $1-\frac{\delta}{2}$:

$$\sup_{f\in\mathcal{F}}\mathbb{E}_S f(Z)-\mathbb{E}_D f(Z)\le\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_S f(Z)-\mathbb{E}_D f(Z)\right]+\sqrt{\frac{\ln\frac{4}{\delta}}{2n}}$$

Notes: LHS is sample-based uniform deviation which can be considered as r.v. depending on sample S; RHS consists of expected uniform deviation $\mathbb{E}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_S f(z)-\mathbb{E}_D f(z)\right]$ and a term $\sqrt{\frac{\ln\frac{4}{\delta}}{2n}}$ which is indepedent of the complexity of $\mathcal{F}$.

Lemma 2 (Symmetrization)

$$\begin{aligned}\mathbb{E}_{S\sim D^n}\left[\sup_{f\in\mathcal{F}}\mathbb{E}_S f(z)-\mathbb{E}_D f(z)\right] &\le \frac{2}{n}\mathbb{E}_{S\sim D^n}\mathbb{E}_{\sigma\sim U(\pm1)^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}f(z_i)\sigma_i\right] \\ &=: 2\,\mathrm{Rad}_n(\mathcal{F})\end{aligned}$$

Remarks:

1. With the complexity of $\mathcal{F}$ increases, both the LHS and the RHS increase.

2. LHS = expectation over the maximum of an infinite collection of r.v.'s; RHS = expectation over the maximum of a finite collection of r.v.'s. So we have reduced bounding something hard to reason about to something much easier to reason about.

3. LHS: $\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)$ has asymmetric distribution; RHS: $\sum_{i=1}^n f(z_i)\sigma_i$ has symmetric distribution(with PDF symmetric around 0).

Definition (Rademacher Complexity) With class $\mathcal{F}$ and sample set S of size n, the Empirical Rademacher Complexity of $\mathcal{F}$ is defined as

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\sigma\sim U(\pm 1)^n}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n f(z_i)\sigma_i\right]$$

The Population/distribution Rademacher Complexity of $\mathcal{F}$ is defined as

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}_{S\sim D^n}\,\mathrm{Rad}_S(\mathcal{F})$$

Here $U(\pm 1)$, the uniform distribution over $\{\pm 1\}$ is called the Rademacher distribution, hence the name Rademacher complexity.
e.g.

1. If $\mathcal{F} = \{f\}, \mathrm{Rad}_S(\mathcal{F}) = 0$.
2. If $\mathcal{F} = \{$ all functions from $\mathcal{Z}$ to $\{\pm 1\}\}$, $\mathrm{Rad}_S(\mathcal{F}) = 1$.
3. If $\mathcal{F} = \{$ all functions from $\mathcal{Z}$ to $\{0, 1\}\}$, $\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{2}$. (This is an exercise.)

Lemma 3 (Relating Rademacher complexity to growth function) For any sample set S of size n:

$$\mathrm{Rad}_S(\mathcal{F}) \le \sqrt{\frac{2\ln S(\mathcal{F}, n)}{n}}$$

Consequently

$$\mathrm{Rad}_n(\mathcal{F}) \le \sqrt{\frac{2\ln S(\mathcal{F}, n)}{n}}$$

Proof of Theorem 2:

$$\begin{aligned}\sup_{f\in\mathcal{F}} \mathbb{E}_S f(Z) - \mathbb{E}_D f(Z) &\le \sqrt{\frac{\ln\frac{4}{\delta}}{2n}} + 2\,\mathrm{Rad}_n(\mathcal{F}) \\ &\le \sqrt{\frac{\ln\frac{4}{\delta}}{2n}} + 2\sqrt{\frac{2\ln S(\mathcal{F}, n)}{n}} \\ &\le \epsilon(\mathcal{F}, n)\end{aligned}$$

4 Proof of Lemma 3

For all $(b_1, \ldots, b_n) \in \Pi_{\mathcal{F}}(S)$, there exists an f from $\mathcal{F}$, such that it achieves this labeling on $S = (z_1, \ldots, z_n)$. Therefore, $\mathrm{Rad}_S(\mathcal{F})$ can be equivalently written as:

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\sigma\sim U(\pm 1)^n}\left[\sup_{\vec{b}=(b_1,\ldots,b_n)\in\Pi_{\mathcal{F}}(S)}\sum_{i=1}^n b_i\sigma_i\right]$$

with $|\Pi_{\mathcal{F}}(S)| \leq S(\mathcal{F}, n)$. Denote the random variable $X_b = \sum_{i=1}^{n} b_i \sigma_i$. Therefore, the quantity we take expectation over is the maximum of at most $S(\mathcal{F}, n)$ zero-mean random variables.

Massart's Finite lemma Suppose $X_1, \ldots, X_N$ are zero mean, σ^2-SG, then

$$\mathbb{E}\left[\max_{i=1}^{N} X_i\right] \leq \sigma \cdot \sqrt{2 \ln N}$$

One proof idea of Massart's Lemma (This is an exercise): We aim to bound $\mathbb{E}\left[\max_{i=1}^{N} |X_i|\right]$. As for general nonnegative random variables Y, $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y \geq z)\, \mathrm{d}z$, it suffices to control the probability tail of random variable $\max_{i=1}^{N} |X_i|$. To this end, recall that if X_i is σ^2-SG, then $\forall z, \mathbb{P}(|X_i| \geq z) \leq 2 \exp\left(-\frac{z^2}{2\sigma^2}\right)$. Hence $\mathbb{P}(\max_i |X_i| \geq z) \leq \min\left(2N \exp\left(-\frac{z^2}{2\sigma^2}\right), 1\right)$. Therefore,

$$\mathbb{E}\left[\max_{i=1}^{N} |X_i|\right] \leq \int_0^\infty \min\left(2N \exp\left(-\frac{z^2}{2\sigma^2}\right), 1\right) \mathrm{d}z.$$

Calculating the integral gives a weaker version of Massart's Lemma (with slightly looser constants.) □

Applying Massart's Lemma, X_b is n-SG, $N \leq S(\mathcal{F}, n)$, then:

$$\mathrm{Rad}_S(\mathcal{F}) \leq \frac{1}{n} \cdot \sqrt{n} \cdot \sqrt{2 \ln S(\mathcal{F}, n)} \leq \sqrt{\frac{2 \ln S(\mathcal{F}, n)}{n}}.$$

□