Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 7

Yun Yang

- Uniform laws of large numbers

Rademacher complexity

For any fixed collection $x_1^n = (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathcal{F}(x_1^n) = \left\{ \big(f(x_1), \ldots, f(x_n)\big) \,\middle|\, f \in \mathcal{F} \right\}.$$

Recall that the Ramemacher complexity of this set (rescaled by n^{-1}) is defined by

$$\mathcal{R}\big(\mathcal{F}(x_1^n)/n\big) = \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right],$$

which is called the empirical Rademacher complexity.

Definition

Given random samples $X_1^n = (X_1, \ldots, X_n)$, the Rademacher complexity of the function class $\mathcal{F}$ is defined as

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_X\big[\mathcal{R}\big(\mathcal{F}(x_1^n)/n\big)\big] = \mathbb{E}_{X,\,\varepsilon}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right].$$

A uniform law via Rademacher complexity

Rademacher complexity characterizes the typical largest correlation between a random noise vector and any function in the class $\mathcal{F}$, thereby the "complexity" of $\mathcal{F}$.

Theorem

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ that is uniformly bounded by $b > 0$. Then for all $n > 0$ and $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\,\mathcal{R}_n(\mathcal{F}) + \delta$$

with $\mathbb{P}$ probability at least $1 - 2\,\exp\left(-\frac{n\delta^2}{8b^2}\right)$. Consequently, $\mathcal{R}_n(\mathcal{F}) = o(1)$ implies $\mathcal{F}$ to be Glivenko-Cantelli.

Proof step one: Concentration around mean

Consider the function

$$G(x_1, \dots, x_n) = \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^{n} f(x_i) \Big|.$$

It satisfies the bounded difference property: for all $x_1, \dots, x_n, x'_k \in \mathbb{R}$,

$$\big| G(x_1, \dots, x_n) - G(x_1, \dots, x_{k-1}, x'_k, x_{k+1}, \dots, x_n) \big| \le \frac{2\|f\|_\infty}{n} \le \frac{2b}{n}.$$

Therefore, the bounded difference inequality implies the following holds with probability at least $1 - 2\exp\big(-\frac{n\,t^2}{8b^2}\big)$,

$$\Big| \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \Big| \le t, \quad \text{for any } t > 0.$$

Proof step two: Upper bound on mean

Applying the symmetrization technique.

Let $(Y_1, \ldots, Y_n)$ be a second independent copy of $(X_1, \ldots, X_n)$. Then

$$
\begin{aligned}
\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &= \mathbb{E}_X\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]\}\Big|\Big] \\
&= \mathbb{E}_X\Big[\sup_{f\in\mathcal{F}}\Big|\mathbb{E}_Y\big[\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\big]\Big|\Big] \\
&\leq \mathbb{E}_{X,Y}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\{f(X_i) - f(Y_i)\}\Big|\Big],
\end{aligned}
$$

where the last step is due to Jensen's inequality.

Proof step two: Upper bound on mean

Let ε_i be i.i.d. Rademacher random variables.

For any $f \in \mathcal{F}$, random variable $\varepsilon_i(f(X_i) - f(Y_i))$ has the same distribution as $f(X_i) - f(Y_i)$. Consequently,

$$\begin{aligned}\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &\leq \mathbb{E}_{X,Y}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\big\{f(X_i) - f(Y_i)\big\}\Big|\Big] \\ &\leq 2\mathbb{E}_{X,\,\varepsilon}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\Big|\Big] = 2\mathcal{R}_n(\mathcal{F}).\end{aligned}$$

Necessary conditions with Rademacher complexity

In the proof, we relates $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ with its symmetrized version

$$\|R_n\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|.$$

The stochastic process R_n over $\mathcal{F}$ is known as the Rademacher process. What is lost in moving from $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ to $\|R_n\|_{\mathcal{F}}$? Essentially nothing!

Theorem

For any convex non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}_{X,\varepsilon}\left[\Phi\left(\frac{1}{2}\|R_n\|_{\bar{\mathcal{F}}}\right)\right] \leq \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \leq \mathbb{E}_{X,\varepsilon}\left[\Phi\left(2\|R_n\|_{\mathcal{F}}\right)\right],$$

where $\bar{\mathcal{F}} = \left\{f - \mathbb{E}[f] : f \in \mathcal{F}\right\}$ is the re-centered function class.

Necessary conditions with Rademacher complexity

Corollary

Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ that is uniformly bounded by $b > 0$. Then for all $n > 0$ and $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{b}{2\sqrt{n}} - \delta$$

with $\mathbb{P}$ probability at least $1 - 2\,\exp\left(-\frac{n\delta^2}{8b^2}\right)$.

Combined with the previous result, we obtain a two sided bound: with probability at least $1 - 2\,\exp\left(-\frac{n\delta^2}{8b^2}\right)$,

$$\frac{1}{2}\mathcal{R}_n(\mathcal{F}) - 2\delta \leq \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathcal{R}_n(\mathcal{F}) + \delta, \text{ for all } \delta > 0.$$

Some upper bounds on Rademacher complexity

Given $x_1^n = (x_1, \ldots, x_n)$, the size of

$$\mathcal{F}(x_1^n) = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}$$

provides a sample-dependent measure of the complexity of $\mathcal{F}$.

Definition

For a class $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$, the growth function is

$$\Pi_{\mathcal{F}}(n) = \max\left\{|\mathcal{F}(x_1^n)| : x_1, \ldots, x_n \in \mathcal{X}\right\}.$$

A class $\mathcal{F}$ is said to have polynomial growth of order $\nu \geq 1$ if

$$\Pi_{\mathcal{F}}(n) \leq (n+1)^{\nu}, \quad \text{for all } n \geq 1.$$

Controlling Rademacher complexity: growth function

Theorem

For any $x_1^n = (x_1, \ldots, x_n)$*, let* $D(x_1^n) = \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ *denote the* ℓ_2 *radius of* $\mathcal{F}(x_1^n)/\sqrt{n}$*. Then*

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = \mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right] \leq D(x_1^n) \sqrt{\frac{2 \log(2\, \Pi_{\mathcal{F}}(n))}{n}}.$$

In particular, if $\mathcal{F}$ *is uniformly bounded by* $b > 0$*, and has polynomial growth of order* $\nu \geq 1$*, then*

$$\mathcal{R}_n(\mathcal{F}) \leq b \sqrt{\frac{2\nu \log(2(n+1))}{n}}.$$

Proof is left as a homework problem.

Application: Classical Glivenko-Cantelli theorem

Recall the classical Glivenko-Cantelli theorem on the uniform convergence of CDFs:

$$\|\widehat{F}_n - F\|_\infty \overset{\text{a.s.}}{\to} 0,$$

Corollary

Let F be the cdf and $\widehat{F}_n$ the empirical CDF, then

$$\mathbb{P}\left[\|\widehat{F}_n - F\|_\infty \geq \sqrt{\frac{2\log(2(n+1))}{n}} + \delta\right] \leq 2e^{-\frac{n\delta^2}{8}} \quad \text{for all } \delta > 0,$$

and hence $\|\widehat{F}_n - F\|_\infty \overset{\text{a.s.}}{\to} 0$.

Proof: Take $\mathcal{F} = \{(-\infty, t] : t \in \mathbb{R}\}$, then $\mathcal{F}$ is uniformly bounded by 1, and has polynomial growth of order 1.

The bound is not tight (the $\log(n+1)$ factor can be removed).

Vapnik-Chervonenkis (VC) dimension

Definition

A class $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ shatters $(x_1, \ldots, x_d) \subset \mathcal{X}$ means $|\mathcal{F}(x_1^d)| = 2^d$.

The VC-dimension $d_{VC}(\mathcal{F})$ is defined as the largest integer d for which there is some $(x_1, \ldots, x_d) \subset \mathcal{X}$ of d points that can be shattered by $\mathcal{F}$.

Examples

- $\mathcal{F}_{\text{left}} = \{(-\infty, t] : t \in \mathbb{R}\}$ has VC-dim 1. It has polynomial growth of order 1.
- $\mathcal{F}_{\text{two}} = \{(s, t] : s, t \in \mathbb{R}\}$ has VC-dim 2. It has polynomial growth of order 2 (why?).

Vapnik-Chervonenkis (VC) dimension

Theorem (Sauer's Lemma)

If $d_{VC}(\mathcal{F}) \leq d$*, then*

$$\Pi_{\mathcal{F}}(n) \leq \sum_{k=1}^{d} \binom{n}{k} \leq (n+1)^d.$$

Consequently, if $d_{VC}(\mathcal{F}) < \infty$ *(called VC class), then* $\mathcal{F}$ *has polynomial growth of order* $d_{VC}(\mathcal{F})$*.*

Proof: See "Weak convergence and empirical processes: with applications to statistics", Section 2.6.1.

Some useful results on Rademacher complexity

Properties

1. $\mathcal{F}_1 \subset \mathcal{F}_2$ implies $\mathcal{R}_n(\mathcal{F}_1) \leq \mathcal{R}_n(\mathcal{F}_2)$.
2. For any constant $c \in \mathbb{R}$, $\mathcal{R}_n(c\,\mathcal{F}) = |c|\,\mathcal{R}_n(\mathcal{F})$.
3. For any fixed bounded function g (bounded by b), $\left|\mathcal{R}_n(\mathcal{F}+g) - \mathcal{R}_n(\mathcal{F})\right| \leq b\sqrt{2\log 2/n}$.
4. $\mathcal{R}_n(\mathrm{conv}(\mathcal{F})) = \mathcal{R}_n(\mathcal{F})$, where $\mathrm{conv}(\mathcal{F})$ is the convex hull of $\mathcal{F}$.
5. If $\phi : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz continuous and satisfies $\phi(0) = 0$, then $\mathcal{R}(\phi(\mathcal{F})) \leq 2\mathcal{R}(\mathcal{F})$.

For a proof of the last claim, see “Probability in Banach Spaces” by Michel Ledoux and Michel Talagrand, Theorem 4.12.