36-710: Advanced Statistical Theory II **Fall 2019**

Lecture #2: August 28

Lecturer: Alessandro Rinaldo *Scribes: Mikaela Meyer*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Quick announcements

- If you're still on the waitlist, talk to Ale
- We voted in class today to do a final project instead of a final exam. More details to come.

When we left off in the spring in 36-709, we were discussing VC theory. This lecture serves as a recap of the elements of VC theory we covered then. Thus, proofs of theorems will only be sketched in these notes; full proofs can either be found in last semester's notes or chapter 4 of [W].

2.0 Rademacher Complexity

Recall that we are interested in $\|P_n - P\|_{\mathcal{F}}$, which is the supremum of the empirical process. Note that $\mathcal{F}$ is a class of uniformly bounded, real-valued functions, and P_n is the empirical measure. In order to control $\|P_n - P\|_{\mathcal{F}}$, we need to be able to control the Rademacher complexity of the class of functions, $\mathcal{F}$.

Definition: Let $x_1^n = (x_1, ..., x_n) \in \mathcal{X}^n$ be arbitrary and set $\mathcal{F} = \{(f(x_1), ..., f(x_n)), f \in \mathcal{F}\} \subseteq \mathbb{R}^n$. Let $\epsilon_1, ..., \epsilon_n \overset{iid}{\sim}$ Rademacher (ie, $P(\epsilon_1 = 1) = P(\epsilon_1 = -1) = 1/2$)

The **empirical Rademacher complexity** of $\mathcal{F}$ at x_1^n is

$$\mathcal{R}_n(\mathcal{F}(x_1^n)) = \mathbb{E}_{\underline{\epsilon}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right]$$

where $\underline{\epsilon} = \epsilon_1, ..., \epsilon_n$. The **Rademacher complexity** of $\mathcal{F}$ with respect to P is

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\underline{x},\epsilon}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right]$$

where $\underline{x} = (x_1, ..., x_n) \overset{iid}{\sim} P$. Essentially, the Rademacher complexity tells us how well we can fit functions of $\mathcal{F}$ to random noise.

The punch line of all of this is that: $\|P_n - P\|_{\mathcal{F}} \overset{p/a.e.}{\to} 0$ iff $\mathcal{R}_n(\mathcal{F}) \to 0$ as $n \to \infty$. The following theorem (theorem 4.10 in [W]) says that the supremum of the empirical process concentrates around $2\mathcal{R}_n$.

Theorem 2.1 *Let $\mathcal{F}$ be a class of functions on $\mathcal{X}$ taking values in $\mathbb{R}$ such that $\|f\|_\infty = \sup_x f(x) \leq b \forall f \in \mathcal{F}$ and let $x_1, ..., x_n \overset{iid}{\sim} P$, where P is a probability measure on $\mathcal{X}$. Then, $\forall t > 0$,*

$$P(\|P_n - P\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + t) \geq 1 - exp\left(\frac{-nt^2}{2b^2}\right)$$

Proof sketch:

1. Show that $\|P_n - P\|_{\mathcal{F}}$ concentrates around its mean, $\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right]$ in a sub-Gaussian fashion. To do this, use the bounded difference inequality.
2. From the symmetrization lemma (given below; proposition 4.11 in [W]), $\mathbb{E}\left[\|P_n - P\|_{\mathcal{F}}\right] \leq 2\mathcal{R}_n(\mathcal{F})$.

Lemma 2.2 *Let $\mathcal{F}$ be a class of integrable functions with respect to P on $\mathcal{X}$, and let $\|R_n\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \frac{1}{n} \left|\sum_{i=1}^n \epsilon_i f(x_i)\right|$ and $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\underline{x},\underline{\epsilon}}\left[\|\mathcal{R}_n\|_{\mathcal{F}}\right]$. Then, for any convex, non-decreasing function, ϕ,*

$$\mathbb{E}_{\underline{x},\underline{\epsilon}}\left[\phi\left(\frac{1}{2}\|\mathcal{R}_n\|_{\bar{\mathcal{F}}}\right)\right] \leq \mathbb{E}_{\underline{x}}\left[\phi\left(\|P_n - P\|_{\mathcal{F}}\right)\right] \leq \mathbb{E}_{\underline{x},\underline{\epsilon}}\left[\phi\left(2\|\mathcal{R}_n\|_{\mathcal{F}}\right)\right]$$

where $\bar{\mathcal{F}} = \{f - \mathbb{E}\left[f(x)\right], f \in \mathcal{F}\}$

It is also possible to show that, with probability $\geq 1 - \exp(\frac{-nt^2}{2b^2})$,

$$\|P_n - P\|_{\mathcal{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(x)]|}{\sqrt{n}} - t$$

2.1 VC Theory

Definition: $\mathcal{F}$ has **polynomial discrimination** with parameter $\nu \geq 1$ if for each $n \geq 1$ and each x_1^n, $|\mathcal{F}(x_1^n)| = (f(x_1), ..., f(x_n)), f \in \mathcal{F} \subseteq \mathbb{R}^n \leq (n+1)^\nu$, where $|\mathcal{F}(x_1^n)|$ is the cardinality of the set.

This is an interesting property because this could hold for classes $\mathcal{F}$ that are infinitely large.

Lemma 2.3 *If $\mathcal{F}$ has polynomial discrimination with parameter ν,*

$$\mathbb{E}_{\underline{\epsilon}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \epsilon_i f(x_i)\right|\right] \leq 2D(x_1^n)\sqrt{\frac{\nu log(n)}{n}}$$

where $D(x_1^n) = \sup_f \sqrt{\frac{1}{n}\sum_{i=1}^n f^2(x_i)} \quad \forall x_1^n$

The quantity $D(x_1^n)$ is bounded above by b if functions are uniformly bounded by b, and we said at the beginning of the lecture that we are concerned with uniformly bounded functions.
This lemma implies that $\forall P$,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_{\underline{x}}\left[\mathbb{E}_{\underline{\epsilon}|\underline{x}}\left[\mathcal{R}_n(\mathcal{F}(x_1^n))\right]\right] \\
&= \mathbb{E}_{\underline{x}}\left[\mathbb{E}_{\underline{\epsilon}}\left[\mathcal{R}_n(\mathcal{F}(x_1^n)\right]\right] && \text{when } \underline{x} \text{ is fixed, } \underline{\epsilon} \text{ independent} \\
&\leq \mathbb{E}_{\underline{x}}\left[2b\sqrt{\frac{\nu log(n)}{n}}\right] && \text{lemma 2.3} \\
&= 2b\sqrt{\frac{\nu log(n)}{n}}
\end{aligned}$$

Example: Last semester, we looked at the class of functions $\mathcal{F} = \{\mathbb{I}_{(-\infty,z)}, z \in \mathbb{R}\}$. We know $\|P_n - P\|_{\mathcal{F}} = \sup_{z\in\mathbb{R}} \left|\hat{F}_n(z) - F(z)\right|$ for this class of functions, where F is the CDF. It is easy to see that $|\mathcal{F}(x_1^n)| \leq (n+1)\ \ \forall x_1^n$. In this case, $\nu = 1$, so

$$P\left(\sup_{z\in\mathbb{R}} \left|\hat{F}_n(z) - F(z)\right| \geq 4\sqrt{\frac{log(n+1)}{n}} + t\right) \leq 2\text{exp}\left\{\frac{-nt^2}{2}\right\}$$

Setting equal to $\frac{1}{n}$ and solving for t, we get that, with probability $\geq 1 - \frac{1}{n}$

$$\sup_{z\in\mathbb{R}} \left|\hat{F}_n(z) - F(z)\right| \leq c\sqrt{\frac{log(n)}{n}}$$

This result is cool because we're not fixing z since we are taking the supremum over all z.

NOTE: To get a tighter bound, use the DKW Inequality, which says
$P(\sup_{z\in\mathbf{R}} \left|\hat{F}_n(z) - F(z)\right| \geq t) \leq 2\text{exp}\left\{\frac{-nt^2}{2}\right\}$

2.1.1 VC Dimension

Let $\mathcal{F}$ be a class of Boolean functions (ie, take values in 0,1). Each $f \in \mathcal{F}$ corresponds to a subset of $\mathcal{X} : \{x \in \mathcal{X} : f(x) = 1\}$. We will develop the theory for $\mathcal{A}$, a collection of subsets of $\mathcal{X}$.

Then define for each x_1^n, $\mathcal{A}(x_1^n) = \{x_1^n \cap A, A \in \mathcal{A}\}$. We should note that we can think of $\mathcal{A}(x_1^n)$ as $\mathcal{F}(x_1^n)$. Clearly, $|\mathcal{A}(x_1^n)| \leq 2^n$ for any x_1^n (the element x_i can either be in the subset or not, which means the cardinality of this set cannot be greater than 2^n, which is the number of subsets you can make from n elements). Informal: If $\mathcal{A}$ is a class of sets of finite dimension, ν, then $|\mathcal{A}(x_1^n)| = 2^n$.

Definition: A n-tuple of points, x_1^n, is said to be **shattered by** $\mathcal{A}$ if $|\mathcal{A}(x_1^n)| = 2^n$

Definition: The **VC Dimension** of $\mathcal{A}$, $\nu = \nu(\mathcal{A})$, is the largest integer n such that <u>some</u> n-tuple x_1^n is shattered by $\mathcal{A}$.

Simply put, a set of points is shattered if all possible combinations of points in the set can be "picked out" by $\mathcal{A}$. Thus, the VC Dimension refers to the largest number of points that $\mathcal{A}$ can "pick out". Let's look at some examples.

Examples:

1. $\mathcal{A} = \{(-\infty, z], z \in \mathbb{R}\}$
   The VC Dimension is 1; given any two numbers in $\mathbb{R}$, $\mathcal{A}$ cannot "pick out" the largest valued number without also including the other number in the subset.
2. $\mathcal{A} = \{(a, b], a < b\}$
   The VC Dimension is 2; given any three numbers in $\mathbb{R}$, $\mathcal{A}$ cannot "pick out" only the largest and smallest numbers. Any subset that includes the largest and smallest numbers would have to include the third, in-between number.
3. $\mathcal{A} =$ the set of polygons in $\mathbb{R}^2$ with arbitrarily many edges
   The VC Dimension is infinity. Imagine putting a large number of points on a circle; no matter how many points there are, a polygon can be drawn to connect them.