Stats 300b: Theory of Statistics **Winter 2017**

Lecture 12 – February 16

Lecturer: John Duchi *Scribe: Youngsuk Park, Youngtak Sohn*

Warning: *these notes may contain factual errors*

Outline:

- Uniform Laws via Entropy numebrs
- Classes with finite entropies
  -Nonparametric classes
  -VC classes

Recap: Given $\mathcal{F}$ with distance d, $N(\mathcal{F},d,\epsilon) = min\{N \in \mathbb{N} \mid \exists\ \epsilon\text{-cover of } \mathcal{F} : \{f_i\}_{i=1}^N$ in distance $d\}$

Chaining: If $\{X_t\}$ is sub-Gaussian precess, $\log(\mathbb{E}\exp(\lambda(X_s - X_t))) \leq \frac{\lambda^2 d(s,t)^2}{2}$. Then, $\mathbb{E}(\sup_{t\in T} X_t) \leq C\mathcal{J}(\mathcal{F},d) = C\int_0^\infty \sqrt{\log N(\mathcal{F},d,\epsilon)}d\epsilon$.

Entropy number $\rightarrow$ uniform laws. For empirical distirbution P_n, let $L_p(P_n)$ be the L_p norm w.r.t. P_n, i.e., $\|f\|_{L_p(P_n)} = (\frac{1}{n}\sum|f(X_i)|^p)^{1/p}$.

Example 1: Often use $L_2(P_n)$ norm in symmetrized processes, i.e., if $Z_f := \frac{1}{\sqrt{n}}\sum\epsilon_i f(X_i)$ where $\epsilon_i \overset{i.i.d}{\sim} unif(-1.+1)$.

For fixed $X_1,\ldots,X_n$,

$$\log(\mathbb{E}\exp(\lambda(Z_f - Z_g))) = \log(\mathbb{E}\exp(\lambda\frac{1}{\sqrt{n}}\sum(\epsilon_i(f(X_i) - g(X_i))))$$
$$\leq \frac{\lambda^2}{2n}\sum(f(X_i) - g(X_i))^2 = \frac{\lambda^2}{2}\|f-g\|_{L_2(P_n)}^2.$$

i.e., $f \rightarrow \frac{1}{\sqrt{n}}\sum\epsilon_i f(X_i)$ is an $\|\cdot\|_{L_2(P_n)}^2$ sub-Gaussian process. so,

$$\mathbb{E}[\sup_{f\in\mathcal{F}}|\frac{1}{\sqrt{n}}\sum\epsilon_i f(X_i)| \mid X_1\ldots,X_n] \leq C\int_0^\infty\sqrt{\log N(\mathcal{F},L_2(P_n),\epsilon)}d\epsilon.$$

For $M < \infty$, let $f_M(x) = f(x)$ if $|f(x)| \leq M$ or 0 otherwise. Let $\mathcal{F}$ be a collection of functions on $\mathcal{X}$ with envelop F, i.e., $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and $F \in L_1(P)$. Define $\mathcal{F}_M := \{f_M\}_{f\in\mathcal{F}}$.

Theorem 1. *(ULLNs with entropies): If $\sqrt{\log N(\mathcal{F},L_1(P_n),\epsilon)} = o_p(n)$ for all $M < \infty, \epsilon > 0$, then $\|P_n - P\|_{\mathcal{F}} \overset{p}{\rightarrow} 0$, ie. $\mathcal{F}$ is G.C. class.*

Proof Let $P_n^0(f) := \frac{1}{n}\sum\epsilon_i f(X_i)$ where $\epsilon_i \overset{i.i.d}{\sim} unif(-1.+1)$. Then $|P_n^0 f| \leq P_n|f| = \|f\|_{L_1(P_n)}$.

$$\mathbb{E}[\|P_n - P\|] \leq 2\mathbb{E}[\sup_{f\in\mathcal{F}}||P_n^0 f|] \leq 2\mathbb{E}[\sup_{f\in\mathcal{F}}||P_n^0(f - f_M)|] + \mathbb{E}[\sup_{f\in\mathcal{F_M}}|P_n^0 f|]$$

Note that, $2\mathbb{E}[\sup_{f\in\mathcal{F}} ||P_n^0(f - f_M)|] \leq 2\mathbb{E}[F1(F \geq M)]$.

Now control $E[\sup_f |P_n^0 f|]$. Let $\mathcal{G}$ be a minimal ϵ - cover of $\mathcal{F}_m$ in $L_1(P_n)$ norm. Then, $card(\mathcal{G}) = N(\mathcal{F}_M, L_1(P_n), \epsilon)$.
Therefore, $\sup_{f\in\mathcal{F}} ||P_n^0 f| \leq \max_{g\in\mathcal{G}} |P_n^0 g| + \epsilon$ (by triangular inequality).

Note that w.l.o.g. $|g(x)| \leq M$, all $g \in \mathcal{G}$, so $P_n^0 g$ is $\frac{M^2}{n}$ sub- Gaussian. So

$$\mathbb{E}[\max_{g\in\mathcal{G}} |P_n^0 g| \mid X] \leq 2\sqrt{2\frac{M^2}{n}\log N(\mathcal{F}, L_1(P_n), \epsilon)}$$

LHS is less than M, and the RHS is $\sqrt{\frac{M}{n} o_p(n)} = o_p(1))$

$$\mathbb{E}[\mathbb{E}[\max_g |P_n^0 g|]] \leq \mathbb{E}[min(M, o_p(1))] \to 0 \text{ as } n \to \infty.$$

$$\Rightarrow \mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \leq 2\mathbb{E}[F1(F \geq M)] + o(1) + \epsilon$$

Since, $\mathbb{E}[F1(F \geq M)] \to 0$ as $n \to \infty$, the proof is done. □

Understand uniform entropies: Often random covering numbers such as $N(\mathcal{F}, L_r(P_n), \epsilon)$ are a bit annoying. so try to give conditions such that $\sup_P N(\mathcal{F}, L_r(P_n), \epsilon)$ can be controlled.

Let's look at some examples in non-parametric function classes.

Example 2: Let $\mathcal{F}$ be the collection of $1-$ Lipschitz functions on $[0,1]$ with $f(0) = 0$. Fix $\epsilon > 0$, consider $\|f\|_\infty := \sup_{x\in[0,1]} |f(x)|$. By dividing the unit intervals by intervals with length ϵ and moving along x axis by epsilon with 3 choice of directions, namely up(45 degree angle), staright, down(45 degree angle) Packing $\#s \geq 3^{\frac{1}{\epsilon}}$(1-Lipschitz function's height change associated widch change of ϵ is also at most ϵ). Thus, $\sup_{P:supp\, P=[0,1]} \log N(\mathcal{F}, L_r(P), \epsilon) \leq \frac{C}{\epsilon}$ where $c < \infty$ is absolute constant. So

$$\mathbb{E}[\sup |P_n - Pf|] \leq 2\mathbb{E}[\sup_f |P_n^o f|] \leq \frac{c}{\sqrt{n}}\mathbb{E}[\int_0^1 \sqrt{\log N(\mathcal{F}, L_2(P_n), \epsilon)}d\epsilon] \leq \frac{c}{\sqrt{n}}\int_0^1 \frac{1}{\sqrt{\epsilon}}d\epsilon \leq \frac{c}{\sqrt{n}}$$

In 2+ dimensions, divided boxes with length ϵ has $\frac{1}{\epsilon^2}$ boxes, (or $(\frac{1}{\epsilon})^d$ in d dimensions), so

$$\log N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \geq \frac{c}{\epsilon^2} \Rightarrow \mathcal{J}(\mathcal{F}, \|\cdot\|_\infty) = \int_0^1 \frac{1}{\epsilon} = +\infty.$$

Vapnik- Chervonenkis (VC classes) Collections of functions or sets with nice combinatorial structure allowing uniform entropy/covering number bounds.

Definition 0.1. *Let $\mathcal{C}$ be a collection of sets and $X = \{X_1, \ldots, X_n\}$ be a collection of points . A vector $y \in \{+1, -1\}^n$ is a labeling of X. Say $\mathcal{C}$ shatters X if for all labelings y of X, $\exists$ a set $A \in \mathcal{C}$, i.e., $X_i \in A$ if $y_i = 1$ and $X_i \notin A$ if $y_i = -1$.*

Equivalently, $\{x_1, \ldots, x_n\} \cap \mathcal{C} = \{A \cap \{x_1, \ldots, x_n \mid A \in \mathcal{C}\}\} = 2^X$.

Example 3: Let $x_1, x_2, x_3 \in \mathbb{R}^2$, not collinear. $\mathcal{C} = \{\text{half space in } \mathbb{R}^2\}$. Than $\mathcal{C}$ shatters $\{x_1, x_2, x_3\}$

Definition 0.2. *: The VC- dimension VC($\mathcal{C}$) is the size of the largest set $\{x_1, \ldots, x_n\}$ s.t. $\mathcal{C}$ shatters $\{x_1, \ldots, x_n\}$.*

Definition 0.3. $\Delta_n(\mathcal{C}, \{x_1, \ldots, x_n\})$ *:= the number of labelings $\mathcal{C}$ realizes on $\{x_i\}$. Then VC($\mathcal{C}$)* $:= \sup\{n \in \mathbb{N} \mid \max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, \{x_i\}) = 2^n\}$.

Example 4: Half-spaces in $\mathbb{R}^d$ have VC($\mathcal{C}$) $= d+1$, Think of $\mathbb{R}^2$. Then $VC(\mathcal{C}) \geq 3$. To do rigorously requires arguing (by geometry) that we would have to have the situation where diagonal labeling does not work.

Lemma 2. *(Sauer- Shelah) for any class $\mathcal{C}$,*

$$\max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, \{x_i\}) \leq \sum_{k=0}^{VC(\mathcal{C})} \binom{n}{k} = O(n^{VC(\mathcal{C})}).$$

Consequence: If $\sup_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}, \{x_i\}) < 2^n$, then $\Delta_n(\mathcal{C}, \{x_i\})$ is polynomial in n.

Let $L_r(P)$ norm on sets $A \subset \mathcal{X}$ be defined by $\|1_A\|_{L_r(P)} = \left(\int 1(x \in A)^r dP(x)\right)^{1/r}$

Theorem 3. *: $\exists$ a universal constant $K < \infty$ s.t. $\forall \epsilon > 0$,*

$$\sup_P N(\mathcal{C}, L_r(P), \epsilon) \leq K \cdot VC(\mathcal{C}) \cdot (4e)^{VC(\mathcal{C})} (\frac{1}{\epsilon})^{r \cdot VC(\mathcal{C})}$$

$\Rightarrow \log N(\mathcal{C}, L_r(P), \epsilon) \leq c \cdot r \cdot VC(\mathcal{C}) \cdot \log(1/\epsilon)$