36-755: Advanced Statistics Theory I **Fall 2017 Nov. 15**

Lecture 21: VC Theory for Functions

Lecturer: Alessandro Rinaldo *Scribes: Yufei Yi*

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

21.1 Extension of VC Theory to Functions on $\mathbb{R}^d$

Lemma 21.1 *Let $\mathcal{F}$ be a class of real-valued function such that $f : \mathcal{X} \to [0,1], \forall f \in \mathcal{F}$. Then, for $X, X_i, i = 1, ..., n \in \mathcal{X}$:*

$$||P_n - P||_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)| \leq \sup_{f\in\mathcal{F}, t\in(0,1)} |\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{f(X_i) > t\} - \mathbb{P}\{f(X_i) > t\}|$$

Proof:

$$\text{Recall :} \qquad X = \int_0^\infty \mathbb{I}\{X > t\}dt \quad , \quad \mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt$$

Fix $f \in \mathcal{F}$,

$$\begin{aligned}
\frac{1}{n}|\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)| &= \frac{1}{n}\Big|\int_0^\infty \sum_{i=1}^n [\mathbb{I}\{f(X_i) > t\} - \mathbb{P}\{f(X_i) > t\}]dt\Big| \\
&\leq \int_0^1 \frac{1}{n}\Big|\sum_{i=1}^n [\mathbb{I}\{f(X_i) > t\} - \mathbb{P}\{f(X_i) > t\}]\Big|dt \\
&\leq \sup_{t\in[0,1]} \frac{1}{n}\Big|\sum_{i=1}^n [\mathbb{I}\{f(X_i) > t\} - \mathbb{P}\{f(X_i) > t\}]\Big| \\
&\leq \sup_{f\in\mathcal{F}, t\in(0,1)} |\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{f(X_i) > t\} - \mathbb{P}\{f(X_i) > t\}|
\end{aligned}$$

■

Remark 21.2 *For $f \in \mathcal{F}$ and $t \in (0,1)$, let*

$$A_{f,t} = \{X \in (0,1), f(X) > t\} \quad \textit{and} \quad \mathcal{A} = \{A_{f,t}, f \in \mathcal{F}, t \in (0,1)\}$$

$$\textit{Then,} \qquad ||P_n - P||_{\mathcal{F}} \leq] \sup_{A\in\mathcal{A}} |P_n(A) - P(A)|$$

The VC dimension of $\mathcal{F}$ is the VC dimension of $\mathcal{A}$.

21.2 Approach Based on Covering Numbers

Definition 21.3 ***(Review: δ-cover and δ-covering number)*** *For $\delta > 0$, a δ-cover of $\mathcal{F}$ w.r.t. metric d on $\mathcal{F}$ is a subset $\{f_1, ..., f_M\} \subset \mathcal{F}$ such that $\forall f \in \mathcal{F}, \exists i \in [1, ..., M]$ satisfying $d(f, f_i) \leq \delta$. The δ-covering number of F w.r.t. metric d is the size of the smallest δ-cover of $\mathcal{F}$.*

Definition 21.4 ***(Empirical Metric $d_{1,P_n}(,)$)*** *For $X_1, ..., X_n$ i.i.d. $\sim P$ and $f, g \in \mathcal{F}$. Define*

$$d_{1,P_n}(f,g) = \frac{1}{n}\sum_{i=1}^{n} |f(X_i) - g(X_i)|$$

$$\textit{Notice} \qquad \mathbb{E}[d_{1,P_n}(f,g)] = \int |f(X) - g(X)| dP_n \quad (L_1 \textit{ distance w.r.t. empirical measure})$$

Theorem 21.5 *For $\lambda \in (0, 1)$,*

$$\mathbb{P}(||P_n - P||_{\mathcal{F}} \geq \lambda) \leq \mathbb{E}[N_1(\frac{\lambda}{8}, \mathcal{F}, P_{2n})] \exp\{-\frac{n\lambda^2}{32}\}$$

$$\textit{where} \quad N_1(\frac{\lambda}{8}, \mathcal{F}, P_{2n}) \textit{ is the cardinality of } \frac{\lambda}{8}\textit{-cover of } \mathcal{F} \textit{ w.r.t. } d_{1,P_{2n}}(f,g).$$

Proof:

1. For $\lambda \in (0, 1)$, if $n \geq \frac{2}{\lambda^2}$, then

$$\begin{aligned}
\mathbb{P}\{||P_n - P||_{\mathcal{F}} \geq \lambda\} &\leq 2\mathbb{P}\{||P_n - P'_n||_{\mathcal{F}} \geq \frac{\lambda}{2}\} \qquad (HW6) \\
&\leq 2\mathbb{E}_{X,X'}\Big\{ \mathbb{P}_\epsilon\Big[\frac{1}{n}\sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} \epsilon_i (f(X_i) - f(X'_i))| \geq \frac{\lambda}{2} \Big| X, X' \Big] \Big\} \qquad (21.1)
\end{aligned}$$

2. Discretization:
   Let $C_{\frac{\lambda}{8}}$ be a smallest $\frac{\lambda}{8}$-cover of $\mathcal{F}$ w.r.t. metric $d_{1,P_{2n}}$. Then,

$$|C_{\frac{\lambda}{8}}| = N_1(\frac{\lambda}{8}, \mathcal{F}, P_{2n}) \qquad \text{and} \qquad \mathcal{F} \subset \bigcup_{g \in C_{\frac{\lambda}{8}}} B(g, \frac{\lambda}{8})$$

$$\textit{where } B(g, \frac{\lambda}{8}) = \{f \in \mathcal{F}, d_{1,P_{2n}}(f,g) \leq \frac{\lambda}{8}\}$$

   Notice that for any $f \in \mathcal{F}$ in $B(g, \frac{\lambda}{8})$,

$$\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - g(X_i)| + \frac{1}{n}\sum_{i=1}^{n} |f(X'_i) - g(X'_i)| \leq \frac{\lambda}{4} \qquad (21.2)$$

Then, (21.1) is upper bounded by

$$\begin{aligned}
&\mathbb{P}_\epsilon\Big[\sup_{g\in C_{\frac{\lambda}{8}}}\sup_{f\in B(g,\frac{\lambda}{8})}\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i(f(X_i)-f(X_i'))\Big|\geq\frac{\lambda}{2}\Big|X,X'\Big]\\
&\leq \sum_{g\in C_{\frac{\lambda}{8}}}\mathbb{P}_\epsilon\Big[\sup_{f\in B(g,\frac{\lambda}{8})}\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i(f(X_i)-f(X_i'))\Big|\geq\frac{\lambda}{2}\Big|X,X'\Big]\\
&\leq \sum_{g\in C_{\frac{\lambda}{8}}}\mathbb{P}_\epsilon\Big[\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i(g(X_i)-g(X_i'))\Big|\geq\frac{\lambda}{4}\Big|X,X'\Big] && \text{By (21.2)}\\
&\leq N_1(\frac{\lambda}{8},\mathcal{F},P_{2n})2\exp\Big\{-\frac{n^2\lambda^2}{32\sum_{i=1}^n[g(X_i)-g(X_i')]^2}\Big\} && \text{Hoeffding}\\
&= N_1(\frac{\lambda}{8},\mathcal{F},P_{2n})2\exp\Big\{-\frac{n\lambda^2}{32}\Big\} && (21.3)
\end{aligned}$$

3. Take expectation w.r.t. X, X' (get rid of the randomness of $N_1(\frac{\lambda}{8},\mathcal{F},P_{2n})$).
   Combine (21.3), (21.1), we get

$$\mathbb{P}\{||P_n-P||_{\mathcal{F}}\geq\lambda\}\leq\mathbb{E}[N_1(\frac{\lambda}{8},\mathcal{F},P_{2n})]2\exp\Big\{-\frac{n\lambda^2}{32}\Big\}$$

∎

Remark 21.6 ***How to bound the random covering number?***

1. *Bound it by* $N_\infty(\frac{\lambda}{8},\mathcal{F})$*, where* $||f-g||_\infty=\sup_{X\in\mathcal{X}}|f(X)-g(X)|$. *(HW 6)*

2. *More generally:*

   Theorem 21.7 *Let* $\mathcal{F}$ *be a class of functions taking values in* $[0,B]$ *with VC dimension* ν*. Then for* $\lambda<\frac{B}{4}$, $p>1$ *:*

$$N_p(\lambda,\mathcal{F})\leq 3\Big[\frac{2eB^p}{\lambda^p}\log(\frac{3eB^p}{ep})\Big]^\nu$$

where $N_p(\lambda,\mathcal{F})$ *is the* λ*-covering number w.r.t.* $L_p-distance:(\int|f(X)-g(X)|^pdP)^{\frac{1}{p}}$

21.3 Good to Know

21.3.1 Measurability :

$\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n f(X_i)-\mathbb{E}f(x)|$ is measurable under the assumption that $\mathcal{F}$ is separable.

Definition 21.8 ***(Separability)*** *A class of function* $\mathcal{F}$ *on* $\mathcal{X}$ *is separable if there exists a countable subset* $\mathcal{F}_0\subset\mathcal{F}$ *such that* $\forall f\in\mathcal{F}$, $\exists\{f_n\}\subset\mathcal{F}_0$ *satisfying*

$$f(X)=\lim_{n\to\infty}f_n(X),\quad\forall X\in\mathcal{X}$$

21.3.2 Talagrand's Inequality for the Suprema of Empirical Processes

Theorem 21.9 *Let $\mathcal{F}$ be a class of real-valued function such that $f : \mathcal{X} \to [0,1], \forall f \in \mathcal{F}$. Then, for $t > 0$ and $\sigma^2(\mathcal{F}) = \sigma^2(\mathcal{F}, P) = \sup_{f \in \mathcal{F}} Var[f(X)]$:*

$$\mathbb{P}\Big\{ ||P_n - P||_{\mathcal{F}} \geq \mathbb{E}||P_n - P||_{\mathcal{F}} + \sqrt{\frac{2t}{n}\sigma^2(\mathcal{F}) + 2\mathbb{E}||P_n - P||_{\mathcal{F}}} + \frac{t}{3n} \Big\} \leq e^{-t}, \quad \textit{and}$$
$$\mathbb{P}\Big\{ ||P_n - P||_{\mathcal{F}} \leq \mathbb{E}||P_n - P||_{\mathcal{F}} - \sqrt{\frac{2t}{n}\sigma^2(\mathcal{F}) + 2\mathbb{E}||P_n - P||_{\mathcal{F}}} - \frac{t}{n} \Big\} \leq e^{-t}$$

Remark 21.10 *We can bound $\mathbb{E}||P_n - P||_{\mathcal{F}}$ by Rademacher complexity of $\mathcal{F}$.*

Next Time: suprema of sub-Gaussian processes.