Stats 300b: Theory of Statistics	Winter 2017

Lecture 13 – February 21

Lecturer: John Duchi *Scribe: Zhimei Ren*

Warning: *these notes may contain factual errors*

Reading:

Outline

- VC classes
- General convergence in distribution:
  - in metric space
  - uniform laws in function spaces
  - compactness in function spaces

1 Recap

Definition 1.1. (Vapnik-Chervonenkis classes) *$\mathcal{C}$(= collection of sets) shatters $x_1, ..., x_n$ if for all labelings $y \in \{\pm 1\}^n$ of $\{x_i\}$ $\exists C \in \mathcal{C}$, s.t.*

$$\begin{cases} y_i = 1 & x_i \in C \\ y_i = 0 & x_i \notin C \end{cases}$$

VC($\mathcal{C}$)=size of largest set $x_1, ..., x_n$ shattered by $\mathcal{C}$.

2 VC Classes

Theorem 1. (Uniform covering numbers in $L_r(P)$) *For sets A,B, let dist(A,B)=$||\mathbb{1}_A - \mathbb{1}_B||_{L_r(P)} = (\int |\mathbb{1}_A - \mathbb{1}_B|^r dP)^{\frac{1}{r}}$. $\exists$ constant $K < \infty$,*

$$\sup_P N(\mathcal{C}, Lr(P), \epsilon) \leq KVC(\mathcal{C})(4e)^{VC(\mathcal{C})}(\frac{1}{\epsilon})^{rVC(\mathcal{C})}$$

i.e.

$$\log N(\mathcal{C}, Lr(P), \epsilon) \lesssim rVC(\mathcal{C}) \log(\frac{1}{\epsilon})$$

Example 1: Let $\mathcal{F} = \{f(x) = \mathbb{1}_{X \leq t}, t \in \mathbb{R}^d\}$. vc($\mathcal{F}$)=$O(d)$.

$$\sup_P \log N(\mathcal{F}, L_2(P), \epsilon) \leq Kd \log(\frac{1}{\epsilon})$$

As a consequence, we have the classical Glivenko Cantelli theorem:

$$
\begin{aligned}
\mathbb{E}[\sup_{t\in\mathbb{R}^d}|\mathbb{P}_n(X\le t)-\mathbb{P}(X\le t)|] &= \mathbb{E}[\sup_{f\in\mathcal{F}}|\mathbb{P}_n f-\mathbb{P}f|] \\
&\le \frac{2}{\sqrt{n}}\mathbb{E}[|\sup_{f\in\mathcal{F}|}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(X_i)](\epsilon_i \overset{i.i.d}{\sim}\{\pm1\}) \\
&\overset{Dudley}{\le} \frac{const}{\sqrt{n}}\int_0^1\sqrt{\log N(\mathcal{F},L_2(P_n),\epsilon)}d\epsilon \\
&= \frac{const\sqrt{d}}{\sqrt{n}}\int_0^1\sqrt{\log(1/\epsilon)}d\epsilon \\
&\le \frac{\tilde{const}\sqrt{d}}{\sqrt{n}}
\end{aligned}
$$

♣

Definition 2.1. *The subgraph of a function:* $\mathcal{X}\to\mathbb{R}$*:*

$$subf := \{(x,t): t<f(x)\} = (epif)^c$$

Note: sub$f\subseteq\mathcal{X}\subseteq\mathbb{R}$.

Definition 2.2. $\mathcal{F}$ *is a VC-class (VC-subgraph-class) if* $subf: f\in\mathcal{F}$ *is VC.*

Example 2: Let $\mathcal{F}=\{f=<\theta,x>:\theta\in\mathbb{R}^d\}$, then VC($\mathcal{F}$) $\le d+2$ ♣

Theorem 2. *If* $VC(\mathcal{F})<\infty$ *and* $\mathcal{F}$ *has envelope* $F:\mathcal{X}\to\mathbb{R}_+$*(i.e.*$F(x)\ge|f(X)|$*, all* $f\in\mathcal{F}$*).*

$$\sup_P N(\mathcal{F},L_r(P),||F||_{L_r(P)}\epsilon)\le constVC(\mathcal{F})(16e)^{VC(\mathcal{F})}(\frac{1}{\epsilon})^{rVC(\mathcal{F})}$$

Example 3: Classification problem. Given $(X_i,y_i)\overset{i.i.d}{\sim}P$,$X_i\in\mathbb{R}^d, y_i\in\{\pm1\}$, choose vector $\theta\in\mathbb{R}^d$ to make $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{sign(\theta^T x)\ne y_i}$ small. i.e. we learn/ fit classifier with rule

$$\hat{y}|x = sign(\theta^T x)$$

$\mathcal{F}=\{<\theta,x>,\theta\in\mathbb{R}^d\}$

Therefore

$$\mathbb{E}[\sup_{\theta\in\Theta}|\mathbb{P}_n(sign(\theta^T x)\ne y_i)-\mathbb{P}(sign(\theta^T x)\ne y_i)|]\le const\sqrt{\frac{d}{n}}(\text{by entropy integral})$$

If you have theta with good empirical error, you should have low error under $\mathbb{P}$. ♣

Properties of VC classes convenient for analysis:

1. if $\mathcal{F}$ is a linear space of functions, dim($\mathcal{F}<\infty$), then VC($\mathcal{F}$)=$O(dim(\mathcal{F}))$.
2. If $\mathcal{C}$ and $\mathcal{D}$ are VC classes of sets,

   $$\mathcal{C}\cup\mathcal{D} := \{C\cup D, C\in\mathcal{C}, D\in\mathcal{D}\}$$

   $$\mathcal{C}\cap\mathcal{D} := \{C\cap D, C\in\mathcal{C}, D\in\mathcal{D}\}$$

   are VC.
3. Comparison: if $\mathcal{F}$ is VC classof functions, then if $\phi:\mathbb{R}\to\mathbb{R}$ is monotone $\phi\circ f: f\in\mathcal{F}$ is VC.

3 Convergence in distribution in metric spaces and uniform CLTs

Our goal here is that given collections of functions $\mathcal{F}$, when is there a limiting Gaussian for $[\sqrt{n}(\mathbb{P}_n f - \mathbb{P}f)]_{f\in\mathcal{F}}$ in a uniform sense? Our starting point is Weak convergence/convergence in distribution.

Recall $X_n \xrightarrow{d} X$ is equivalent to either of the following:

1. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded and continuous f
2. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all Lipschitz and bounded function f.

1 and 2 makes sense even if X_n are in some metric space.

Let $\mathbb{D}$ be a metric space. Then X is a r.v. on $\mathbb{D}$ if $X : \Omega \to \mathbb{D}$. Say X is $\mathbb{D}$-valued. Then given sequence of $X_n : \Omega_n \to \mathbb{D}$. We say $X_n \to X$ if $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded and continuous f (even Lipschitz). (Recall $\mathbb{E}f(X_n) = \int_{\Omega_n} f(X_n(\omega))d\mu_n(\omega)) = \int f(x)\mathbb{P}_n(x))$

Example 4: (Continuous function in a compact set) Let (T, d) be a compact metric space. $L_\infty(T)$=bounded functions: $T \to \mathbb{R}$. For $f, g \in L_\infty(T)$, $||f - g||_\infty = \sup_{t\in T} |f(t) - g(t)|$. Let $l : T \times \mathcal{X} \to \mathbb{R}$ be continuous in t. Define

$$Z_n(.) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [l(., x) - \mathbb{E}l(., x)]$$

Then Z_n is a l_∞-valued random variable. (Because $t \to Z_n(t)$ is contiunous, so $\sup_{t\in T} |Z_n(t)| < \infty$)

Remark Note if T_0 is countable and dense subset of T, Z_n is completely determined by $\{Z_n(t), t \in T_0\}$.

$$(Z_n(t_i), ..., Z_n(t_k)) \xrightarrow{d} \mathcal{N}(0, \text{cov}(l(t_i, x), l(t_j, x))_{i,j=1}^{k})$$

for fixed $t_i, ..., t_k$. ♣

Definition 3.1. *A random variable $X : \Omega \to \mathbb{D}$ is tight if $\forall \epsilon > 0$, $\exists$ a compact set $K \subseteq \mathbb{D}$*

$$\limsup_{n\to\infty} \mathbb{P}(X_n \notin K^\delta) \leq \epsilon$$

all $\delta > 0$.

$K^\delta := \{x : dist(x, K) < \delta\}$

Theorem 3. *(Prohorov) Let $X_n : \Omega \to \mathbb{D}$ and $X : \Omega \to \mathbb{D}$*

1. *If $X_n \xrightarrow{d} X$, where X is tight, then $\{X_n\}$ is asymptotically tight.*
2. *If $\{X_n\}$ is asymptotically tight, then $\exists$ a subsequence $\{n_k\}$, tight $X : \Omega \to \mathbb{D}, s.t.$ $X_{n_k} \xrightarrow{d} X$.*

Compactness in functional space:

Evidently, if $X_n \in L_\infty(T)$, we must understand compactness in $L_\infty(T)$. For us, limits will be in $\mathcal{C}(T, R)$. $\mathcal{C}(T, R)$ =continuous function. $f : T \to \mathbb{R}$, $||f - g||_\infty$ is metric.

Standard compactness theorem: Arzela-Ascoli theorem

Definition 3.2. *For a function $f : T \to \mathbb{R}$, modulus of continuity is*

$$w_f(\delta) = \sup_{d(s,t)<\delta} \{|f(t) - f(s)|\}$$

Definition 3.3. *A collection $\mathcal{F}$ is uniformly equicontinuous if*

$$\lim_{s\downarrow 0} \sup_{f\in\mathcal{F}} w_f(\delta) = 0$$

Theorem 4. *Let (T, d) be a compact metric space. Then:*

1. *$\mathcal{F} \subset \mathcal{C}(T, \mathbb{R})$ is sequentially compact.*

2. *$\mathcal{F}$ is uniformly equicontinuous and $\exists t_0 \in T$, s.t. $\sup_{f\in\mathcal{F}} |f(t_0)| < \infty$*

are equivalent.