36-755: Advanced Statistical Theory I	**Fall 2016**

Lecture 16: October 24

Lecturer: Alessandro Rinaldo *Scribes: Ilmun Kim*

16.1 A uniform law via Rademacher complexity

16.1.1 Classes with polynomial discrimination

Recall if $\mathcal{F}$ is a class of functions that are uniformly bounded by $b > 0$, then we have

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

where

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{X,\epsilon}\left[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\Big|\right] \quad \text{and} \quad ||P_n - P||_{\mathcal{F}} = \sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n \left(f(X_i) - \mathbb{E}\left[f(X_i)\right]\right)\Big|. \tag{16.1}$$

To bound $\mathcal{R}_n(\mathcal{F})$, we will use the VC theory and later we will develop more general tools.

Remark 16.1 (Concentration property of Rademacher complexity) *Let $g : \mathbb{R} \to \mathbb{R}$ such that $|g(x) - g(y)| \leq L|x - y|$, $g(0) = 0$ and set $g \circ \mathcal{F} = \{g \circ f, f \in \mathcal{F}\}$. Then*

$$\mathcal{R}_n(g \circ \mathcal{F}) \leq 2L\mathcal{R}_n(\mathcal{F}).$$

Example 16.2 *Suppose the function $f(x) = x^2$ is defined on $[-L, L]$. Then,*

$$\mathbb{E}_{X,\epsilon}\left[\sum_{f\in\mathcal{F}}\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i f^2(X_i)\Big|\right] \leq 4L\mathbb{E}_{X,\epsilon}\left[\sum_{f\in\mathcal{F}}\frac{1}{n}\Big|\sum_{i=1}^n \epsilon_i f(X_i)\Big|\right].$$

Definition 16.3 (Polynomial discrimination) *A class $\mathcal{F}$ of functions defined on the domain $\mathcal{X}$ such that $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{R}\}$ has polynomial discrimination with parameter $\nu \geq 1$ if for each positive integer n and collection $x_1^n = \{x_1, \cdots, x_n\}$ of n points in $\mathcal{X}$, the set*

$$\mathcal{F}(x_1^n) = \{(f(x_1), \cdots, f(x_n)) \in \mathbb{R}^n, f \in \mathcal{F}\}$$

has cardinality upper bounded as

$$|\mathcal{F}(x_1^n)| \leq (n+1)^\nu.$$

Lemma 16.4 *If $\mathcal{F}$ has polynomial discrimination with paramter ν, then for all n and any x_1^n, we have*

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

where $D(x_1^n) = \sup_{f\in\mathcal{F}}\sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$.

Corollary 16.5 *As a corollary, we obtain the following results.*

1) $\mathcal{R}_n(\mathcal{F}) \le 2\mathbb{E}_X\left[D(x_1^n)\right]\sqrt{\frac{\nu log(n+1)}{n}}$

2) If $||f||_\infty = \sup_{x\in\mathcal{X}}|f(x)| < b$ *for all* $f\in\mathcal{F}$, *then* $\mathcal{R}_n(\mathcal{F}) \le 2b\sqrt{\frac{\nu log(n+1)}{n}}$.

16.1.2 Uniform convergence of CDFs

Consider the function class $\mathcal{F} = \{(-\infty, t] : t\in\mathbb{R}\}$. In this case, $||P_n - P||_\mathcal{F}$ defines the Kolmogorov-Smirnov statistic as

$$||P_n - P||_\mathcal{F} = \sup_t\left|\hat{F}_n(t) - F(t)\right|.$$

Note that for a fixed $x_1^n = (x_1,\cdots,x_n)\in\mathbb{R}^n$, the ordered samples

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n-1)} \le x_{(n)}$$

split the real line into at most $n+1$ intervals including $(-\infty, x_{(1)}]$ and $(x_{(n)}, \infty]$ and for a given t, the indicator function $I_{[t,\infty)}$ takes the value 1 for all $x_{(i)} \ge t$, and 0 for all other samples. It follows that for any given sample x_1^n, we have $|\mathcal{F}(x_1^n)| \le n+1$ and $\mathcal{F}$ has poloynomial discrimination with $\nu = 1$. Consequently, we can show a quantitative version of Glivenko-Cantelli theorem as follows.

Corollary 16.6 *(Classical Glivenko-Cantelli) Let $F(t) = \mathbb{P}(X\le t)$ be the CDF of a random variable X, and let $\hat{F}_n(t)$ be the empirical CDF based on n i.i.d. samples $X_i\sim\mathbb{P}$. Then,*

$$\mathbb{P}\left(||\hat{F}_n - F||_\infty \ge 4\log\sqrt{\frac{\log(n+1)}{n}} + t\right) \le \exp\left(-\frac{nt^2}{2}\right)$$

for all $t\ge 0$, and hence $||\hat{F}_n - F||_\infty \overset{a.s.}{\to} 0$.

Proof: *The claim follows from Eq.(16.1) and Corollary 16.5.* ■

Dvoretzky-Kiefer-Wolfowitz (DKW) inequality provides a shaper tail bound of $||\hat{F}_n - F||_\infty$ as

$$\mathbb{P}\left(||\hat{F}_n - F||_\infty > t\right) \le 2\exp\left(-\frac{nt^2}{2}\right)$$

for every $t > 0$.

16.2 Vapnik-Chervonenkis (VC) dimension

Let us assume $\mathcal{F}$ is a collection of $\{0,1\}$ functions and represent this class using the collection $\mathcal{A}$ of subsets in $\mathcal{X}$ as follows.

$$f \in \mathcal{F} \Longleftrightarrow \mathcal{A} = \{x \in \mathcal{X} : f(x) = 1\}$$

Then, for a fixed x_1^n, we have

$$\mathcal{F}(x_1^n) = \mathcal{A}(x_1^n) = \{A \cap x_1^n : A \in \mathcal{A}\}.$$

Clearly, we can see that $|\mathcal{A}(x_1^n)| \leq 2^n$. A VC-class of sets is a class such that $|\mathcal{A}(x_1^n)|$ grows only polynomially in n.

Definition 16.7 (Shattering and VC dimension) *The class $\mathcal{A}$ shatters the n-tuple x_1^n if $|\mathcal{A}(x_1^n)| = 2^n$. The VC-dimension ν of $\mathcal{A}$ is the largest n such that some n-tuple x_1^n is shattered by $\mathcal{A}$.*

If $n > \nu$, then no n-tuple x_1^n is shattered by $\mathcal{A}$.

Example 16.8 *Here are two typcial examples.*

- *For $\mathcal{A} = \{(-\infty, x] : x \in \mathbb{R}\}$, the VC-dimension of $\mathcal{A}$ is $\nu(\mathcal{A}) = 1$.*
- *For $\mathcal{A} = \{(b, a] : b < a\}$, the VC-dimension of $\mathcal{A}$ is $\nu(\mathcal{A}) = 2$.*

If the VC dimension is finite, then the growth function cannot grow too quickly.

Lemma 16.9 (Sauer's lemma) *Suppose $\mathcal{A}$ has the finite VC-dimension $\nu_{\mathcal{A}}$. Then for $n \geq \nu_{\mathcal{A}}$,*

$$\max_{x_1^n} |\mathcal{A}(x_1^n)| \leq \sum_{i=0}^{\nu_{\mathcal{A}}} \binom{n}{i} \leq (n+1)^{\nu_{\mathcal{A}}}.$$

Let $\mathcal{S}_{\mathcal{A}}(n)$ be the shattering coefficient, $\max_{x_1^n} |\mathcal{A}(x_1^n)|$. Then, Lemma 16.9 provides the following inequality.

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \mathcal{S}_{\mathcal{A}}(2n)}{n}} \leq \sqrt{\frac{4\nu_{\mathcal{A}} \log n}{n}}$$

16.3 Controlling the VC-dimension

16.3.1 Basic operations

Let $\mathcal{A}$ and $\mathcal{B}$ be two collections of subsets in $\mathcal{X}(= \mathbb{R}^d)$. Then,

1. $\mathcal{S}_{\mathcal{A}}(n) = \mathcal{S}_{\mathcal{A}^C}(n)$.
2. If $A \cup B = \{A \cup B : A \in \mathcal{A}, B \in \mathcal{B}\}$, then

$$\mathcal{S}_{A \cup B}(n) \leq \mathcal{S}_{\mathcal{A}}(n)\mathcal{S}_{\mathcal{B}}(n).$$

3. The same bound holds for

$$A \cap B = \{A \cap B : A \in \mathcal{A}, B \in \mathcal{B}\}$$
$$A \times B = \{A \times B : A \in \mathcal{A}, B \in \mathcal{B}\}$$

4. $\mathcal{S}_{\mathcal{A}}(n+m) \leq \mathcal{S}_{\mathcal{A}}(n)\mathcal{S}_{\mathcal{A}}(m)$.
5. If $C = A \cup B$, then

$$\mathcal{S}_{\mathcal{C}}(n) \leq \mathcal{S}_{\mathcal{A}}(n) + \mathcal{S}_{\mathcal{B}}(n).$$

More examples are provided as

1. If $\mathcal{A} = \{(-\infty, x_1] \times \cdots \times (-\infty, x_n] : (x_1, \cdots, x_d) \in \mathbb{R}^d\}$, then $\nu_{\mathcal{A}} = d$.
2. Let $\mathcal{A}$ be the set of all rectangles in $\mathbb{R}^d$. Then, $\nu_{\mathcal{A}} = 2d$.

16.3.2 Vector space structure

Proposition 16.10 *Let $\mathcal{G}$ be a finite dimensional vector space of real-valued functions on $\mathbb{R}^d$. Then, the class*

$$\mathcal{A} = \{\{x : g(x) \leq 0\}, \forall g \in \mathcal{G}\}$$

has VC dimension at most dim($\mathcal{G}$).

Example 16.11 (Spheres in $\mathbb{R}^d$) *Consider the sphere*

$$\mathcal{A} = \{\mathbb{B}(x,r) : x \in \mathbb{R}^d, r > 0\} \quad \text{where} \quad \mathbb{B}(x,r) = \{y : ||x-y||^2 \leq r^2\}.$$

Then, $\nu_{\mathcal{A}} \leq d+2$.

Proof: *Notice that $\forall x \in \mathbb{R}^d$ and $r > 0$,*

$$\begin{aligned} f_{r,y}(x) &= \sum_{i=1}^{d} (x_i - y_i)^2 - r^2 \\ &= \sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d} y_i^2 - 2\sum_{i=1}^{d} x_i y_i - r^2 \leq 0. \end{aligned}$$

We first define a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^{d+2}$ via $\phi(x) = (1, x_1, \cdots, x_d, ||x||_2^2)$. Then, consider functions of the form

$$g_c(x) = c^T \phi(x) \quad \text{where} \quad c \in \mathbb{R}^{d+2}.$$

The family of functions $\{g_c, c \in \mathbb{R}^{d+2}\}$ is a vector space of dimension $d+2$, and it contains the function class $f_{r,y}(x)$. Then, the result follows by Proposition 16.10. ■