36-755: Advanced Statistics **Fall 2017**

Lecture 20: VC Dimension and VC theory

Lecturer: Alessandro Rinaldo *Scribes: Shengming Luo*

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

This lecture's notes illustrate some uses of various LaTeX macros. Take a look at this and imitate.

20.1 Recap

Let $\mathcal{A}$ be a collection of subset of $\mathcal{X} = \mathbb{R}^d$. For $x_1^n = (x_1, \cdots, x_n) \subseteq \mathcal{X}$, let

$$\mathcal{A}(x_1^n) = \Big\{ A \cap x_1^n, A \in \mathcal{A} \Big\}.$$

If $|\mathcal{A}(x_1^n)| = 2^n$ then x_1^n is *shattered* by $\mathcal{A}$. The VC-dimension of $\mathcal{A}$ is largest integer n such that

$$\text{there exists } x_1^n \in \mathcal{X}, \text{ s.t. } |\mathcal{A}(x_1^n)| = 2^n.$$

Hence by definition, if $n > V_A$, no n-tuple x_1^n is shattered by $\mathcal{A}$, where V_A is VC-dim of $\mathcal{A}$.

Warning: sometimes VC-dimension is defined as the smallest integer n s.t. no x_1^n is shattered on $\mathcal{A}$.

Let $S_{\mathcal{A}}(n) = \max_{x_1^n \subseteq \mathcal{X}} |\mathcal{A}(x_1^n)|$ be the shattered coefficient. The Sauer's Lemma states

Lemma 20.1 (Sauer's) *If $\mathcal{A}$ has VC dimension v, then*

$$S_{\mathcal{A}} \leq \begin{cases} \sum_{i=0}^{v} \binom{n}{i} \leq (n+1)^v & \forall n \geq 1, \\ (\frac{en}{v})^v, & \text{if } n \geq v, \end{cases}$$

If $\mathcal{A}$ has VC-dimension v, then $\mathcal{A}$ has polynomial discemination with parameter v.

20.1.1 Examples of class of finite VC-dim:

(1) $\mathcal{A} = \big\{(-\infty, x], x \in \mathbb{R}\big\}$, the VC-dimension = 1.

(2) $\mathcal{A} = \big\{(-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_d], x \in \mathbb{R}^d\big\}$: VC-dimension = d.

(3) If $|\mathcal{A}| = N$, then $V_A \leq \log_2(N)$.

(4) $\mathcal{A} = \Big\{\text{Rectanlges in } \mathbb{R}^d\Big\}$, then VC-dimension is $2d$.

20.2 Properties of VC Dimension

Let $\mathcal{A}, \mathcal{B}$ be classes of subsets of $\mathcal{X}$. Then

- Let $\mathcal{A}^c = \left\{A^c, A \in \mathcal{A}\right\}$ and $S_A(n) = S_{A^c}(n)$, for all $n \geq 1$.
- Let $\mathcal{A} \cup \mathcal{B} = \left\{A \cup B, A \in \mathcal{A}, B \in \mathcal{B}\right\}$, $\mathcal{A} \cap \mathcal{B} = \left\{A \cap B, A \in \mathcal{A}, B \in \mathcal{B}\right\}$, $\mathcal{A} \times \mathcal{B} = \left\{A \times B, A \in \mathcal{A}, B \in \mathcal{B}\right\}$. Their shatter coefficient $\leq S_A(n) S_B(n)$.
- $S_A(n+m) \leq S_A(n) S_B(m)$, where $n, m \in \mathbb{N}$.
- $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, then $S_C(n) \leq S_A(n) + S_B(n)$.

20.3 Vector space structure

Let $\mathcal{G}$ be a vector space of functions on $\mathcal{R}^d$. Then, the class

$$\mathcal{A} = \left\{\left\{x \in \mathcal{R}^d, g(x) \geq 0\right\}, g \in C_g\right\}$$

has VC-dimension $\leq dim(\mathcal{G})$

Proof:

Let $r = dim(\mathcal{G})$, need to show that no collection of $r+1$ points in $\mathbb{R}^d$ is shattered on $\mathcal{A}$. Fix $r+1$ points $(x_1, \cdots, x_{v+1}) \in \mathbb{R}^d$. Define mapping $L : \mathcal{G} \to \mathbb{R}^{v+1}$, where $L(g) = (g(x_1), \cdots, g(x_{v+1})) \subseteq \mathbb{R}^{r+1}$. The image of $\mathcal{G}$ under mapping L is a linear subspace of finite dimension $\leq r$. Then there exists $r = (r_1, \cdots, r_{r+1}) \in \mathbb{R}^{r+1}$, s.t.

$$\sum_{i=1}^{r+1} r_i g(x_i) = 0, \forall g \in \mathcal{G}.$$

W.L.O.G, we need assume that r has at least one negtive coordinate Then,

$$\sum_{r_i \geq 0} r_i g(x_i) = \sum_{r_i \leq 0} -r_i g(x_i).$$

Suppose that there exists $g \in \mathcal{G}$ s.t. the set $\{x \in \mathbb{R}^d, g(x) \geq 0\}$ picks out the x_i's on the LHS. But then LHS ≥ 0 and RHS ≤ 0, which is a contradiction. As a result, $V_A \leq v$.

Example: Consider the sphere

$$\mathcal{A} = \left\{B(x, r), x \in \mathbb{R}^d, r > 0\right\},$$

where $B(x, r) = \left\{y \in \mathbb{R}^d, \|x - y\| \leq 1\right\}$. Then $V_A \leq d + 2$.

To see this, write

$$\sum_{i=1}^{d} (x_i - y_i)^2 - r^2 = \sum_{i=1}^{d} x_i^2 + \sum_{i=1}^{d} y_i^2 - 2 \sum_{i} x_i y_i - r^2.$$

We 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

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

The family of functions $\{g_c, c \in \mathbb{R}^{d+2}\}$ is a vector space of dimension $d+2$. Then, the result follows by previous example. Actually $V_A = d+1$.

20.4 VC dimension continued

20.4.1 Traditional version

Recall: The definition in Martin's book is non-classical. More traditional version

$$\mathbb{P}\big(\sup_{A\in\mathcal{A}} |P_n(A) - P_n(B)| \geq \lambda\big) \leq c_1 S_A(n) e^{-c_2 n\lambda^2} \leq (n+1)^{V_A}$$

Proof:

- (Symmetrization by Ghost Sample) $\mathbb{P}\big(\sup_{A\in\mathcal{A}} |P_n(A) - P_n(B)| \geq \lambda\big) \leq 2\mathbb{P}\big(\sup_{A\in\mathcal{A}} |P_n(A) - P_n'(A)| \geq \lambda/2\big)$, where P_n' is empirical probability measure associated to the ghost samples $(y_1, \cdots, y_n) \overset{i.i.d.}{\sim} P$.
- (Symmetrization of random signs (Radmacher))

$$\leq 2\mathbb{P}(\sup_{A\in\mathcal{A}} \frac{1}{n}\big(\sum_{i=1}^n \epsilon_i 1_{\{x_i\in A\}}\big| \geq \lambda/4\big)$$

$$\leq 2\mathbb{E}_x\mathbb{P}_{\epsilon,x}(\sup_{A\in\mathcal{A}} \frac{1}{n}\big(\sum_{i=1}^n \epsilon_i 1_{\{x_i\in A\}}\big| \geq \lambda/4\big|x_1, \cdots, x_n\big).$$

  By VC theory,

$$\text{RHS } \leq S_A(n) \sup_{A\in\mathcal{A}} \mathbb{P}(\frac{1}{n}\big|\sum_{i=1}^n \epsilon_i 1_{\{x_i\in A\}} - \mathbb{E}_\epsilon[\epsilon_i f(x_i)]\big| \geq \lambda/4\big)$$

- (Hoeffding) By Hoeffding, we can bound RHS $\leq C S_A(n) e^{-n\epsilon^2/32}$.

20.4.2 Relative VC-inequality

As we just saw, the VC inequality depends on the hoeffding inequality. We know hoeffding isnt always the sharpest. So, going from hoeffding to Bernstein is a way to improve the inequality when the variances are small.

The relative VC-inequality is useful if $\mathbb{P}(\mathcal{A})$ as a ranges in $\mathcal{A}$ is not bounded away from 0.

$$\mathbb{P}\big(\sup_{A\in\mathcal{A}} \big|\frac{P_(A) - P_n(A)}{\sqrt{P(A)}} \geq \lambda\big|\big) \leq 4 S_A(2n) e^{-n\lambda^2/4}.$$

As a result, with probability at least $1-\delta$,

$$\mathbb{P}(A) \leq \mathbb{P}_n(A) + 2\sqrt{\frac{\mathbb{P}_n(A)\log(S_A(2n)) + \log(4/\delta)}{n}} + \frac{\log(4S_A(2n)) + \log(4/\delta)}{n}.$$

20.4.3 Extension to functions: Combinatorial dimension and uniform covering

In the previous section we developed, in the special case of classes of sets, a combinatorial method to control uniformly the random covering numbers that appear in symmetrization bounds. Whether such ideas are still useful in the general setting of classes of functions is far from clear at this point: even when restricted to a nite set, a class of functions is still a continuous object (with a potentially nontrivial geometric structure) and is not, a priori, combinatorial in nature. Nonetheless, the theory of previous section admits a very natural generalization to classes of functions, which we develop presently.

We begin with a proposition.

Proposition 20.2

$$\sup_{f\in\mathcal{F}} \|P_n - P\|_{\mathcal{F}} \leq \sup_{f\in\mathcal{F}, t\in[0,1]} \frac{1}{n}\sum_{i=1}^{n} [1_{\{f(x_i)>t\}} - \mathbb{P}(f(x_i) > t)].$$

On the space $\mathcal{F}$ use this metric: $d_{1,p_n}(f,g) = \frac{1}{n}\sum_i |f(xi)g(xi)|$ where $x_1, \cdots, x_n \overset{i.i.d.}{\sim} P$. This is a random metric because it depends on our sample. Let $N(\mathcal{F}, \epsilon, P_{2n}) \to \frac{\epsilon}{\delta}$ covering number of $\mathcal{F}$ with respect to s_1, P_{2n}. Here, N is a random covering.

Then,

$$\mathbb{P}(\|P_n - P\|_{\mathcal{F}} > \epsilon) \leq E[N(\mathcal{F}, \epsilon, d_1, 2n)]e^{n\epsilon^2/32}$$

Instead of using d_{1,p_n} we can use $d_\infty(f,g) = sup_x|f(x) - g(x)|$ because the latter is often easier to compute.