VC Classes and Uniform Metric Entropies

John Duchi

Stats 300b – Winter Quarter 2021

Outline

- VC Classes
- Sauer-Shelah lemma
- Uniform covering numbers

Reading:

- Wainwright, *High Dimensional Statistics*, Chapter 4.3
- van der Vaart, *Asymptotic Statistics*, Chapter 19.2

Motivation

- we have seen

$$\mathbb{E}[\|P_n - P\|_{\mathcal{F}} \mid X_1^n] \leq O(1) \int_0^{\infty} \sqrt{\frac{\sigma_{n,*}^2}{n} \log N(\mathcal{F}, L^2(P_n), \epsilon) d\epsilon}$$

  where

$$\sigma_{n,*}^2 := \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2 = \sup_{f \in \mathcal{F}} P_n f^2.$$

- today: develop some techniques for giving bounds on

$$\sup_Q \log N(\mathcal{F}, L^p(Q), \epsilon)$$

Complexities of finite sets

- let $\mathcal{F}(x_1^n) = \{(f(x_1), \ldots, f(x_n))\}_{f \in \mathcal{F}}$
- some classes take only finitely many values, i.e. $\mathrm{card}(\mathcal{F}(x_1^n)) < \infty$

Lemma

For Rademacher complexity $R_n(\mathcal{F} \mid x_1^n) = \mathbb{E}[\sup_{f \in \mathcal{F}} |\sum_{i=1}^n \varepsilon_i f(x_i)|]$,

$$R_n(\mathcal{F} \mid x_1^n) \le O(1)\sqrt{n\sigma_{n,*}^2 \log \mathrm{card}(\mathcal{F}(x_1^n))},$$

where $\sigma_{n,*}^2 = \sup_{f \in \mathcal{F}} P_n f^2$.

Polynomial discrimination

- class has *polynomial discrimination* of order d if

$$\mathrm{card}(\mathcal{F}(x_1^n)) \leq C(n+1)^d$$

  where $C < \infty$ is a constant
- some classes only grow polynomially as $n \to \infty$

Corollary

If $\mathcal{F}$ has order d polynomial discrimination and $\|f\|_\infty \leq b$ for $f \in \mathcal{F}$, then

$$\mathbb{E}[\|P_n - P\|_{\mathcal{F}}] \leq O(1) b \sqrt{\frac{d \log(Cn)}{n}}$$

Vapnik-Chervonenkis Classes: Shattering

- collection of classes that enjoy uniform laws, covering numbers, and polynomial discrimination

Definition (Shattering)

Let $\mathcal{C}$ be a collection of sets and $x_1^n = \{x_1, \ldots, x_n\}$ a collection of points. A *labeling* of x_1^n is a vector $y \in \{\pm 1\}^n$. The collection $\mathcal{C}$ *shatters* x_1^n if for all labelings y, there exists $A \in \mathcal{C}$ s.t.

$$\begin{cases} x_i \in A & \text{if } y_i = 1 \\ x_i \notin A & \text{if } y_i = -1. \end{cases}$$

Examples of shattering

- let $\mathcal{C}$ be half-spaces in $\mathbb{R}^2$
- $\mathcal{C}$ shatters any 3 non-collinear points $x_1^3 \subset \mathbb{R}^2$

Vapnik-Chervonenkis (VC) Dimension

Definition

For $\mathcal{C} \subset 2^{\mathcal{X}}$ the *shattering number* of $\mathcal{C}$ on x_1^n is

$$\Delta_n(\mathcal{C}, x_1^n) := \text{card}\left\{A \cap \{x_1, \ldots, x_n\} \text{ s.t. } A \in \mathcal{C}\right\}$$

i.e. the number of labelings $\mathcal{C}$ realizes on x_1^n

Definition (Vapnik-Chervonenkis (VC) Dimension)

The *VC-dimension* of $\mathcal{C}$ is

$$\text{VC}(\mathcal{C}) := \sup\left\{n \in \mathbb{N} : \sup_{x_1^n \in \mathcal{X}^n} \Delta_n(\mathcal{C}, x_1^n) = 2^n\right\}.$$

Sauer-Shelah Lemma

- amazing fact: VC classes have polynomial discrimination

Lemma (Sauer-Shelah)

For any collection of sets $\mathcal{C} \subset 2^{\mathcal{X}}$,

$$\sup_{x_1^n \in \mathcal{X}^n} \Delta_n(\mathcal{C}, x_1^n) \leq \sum_{j=0}^{\mathrm{VC}(\mathcal{C})} \binom{n}{j} = O(n^{\mathrm{VC}(\mathcal{C})}).$$

consequence: whenever $\max_{x_1^n} \Delta_n(\mathcal{C}, x_1^n) < 2^n$, then $\mathrm{VC}(\mathcal{C}) < n$ and

$$\Delta_n(\mathcal{C}, x_1^n) = O(n^{\mathrm{VC}(\mathcal{C})}).$$

(Proofs on course webpage)

Examples of VC classes

- For $\mathcal{C}$ = lower left boxes in $\mathbb{R}^d$,

$$\mathrm{VC}(\mathcal{C}) = O(d)$$

- For $\mathcal{C}$ = halfspaces in $\mathbb{R}^d$,

$$\mathrm{VC}(\mathcal{C}) = O(d)$$

Uniform covering numbers with VC-classes

- define $L^r(P)$ norm on sets $A \subset \mathcal{X}$ by

$$\|1_A - 1_B\|_{L^r(P)}^r := \int |1\{x \in A\} - 1\{x \in B\}|^r dP(x)$$

Theorem

There exists constant $K < \infty$ such that for any $\mathcal{C} \subset 2^{\mathcal{X}}$, for all $\epsilon > 0$

$$\sup_P N(\mathcal{C}, L^r(P), \epsilon) \leq K \cdot \mathsf{VC}(\mathcal{C})(4e)^{\mathsf{VC}(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{\mathsf{VC}(\mathcal{C}) \cdot r}.$$

intuition: only realizing polynomially many boxes allows us to cover with ϵ-separated "boxes" of probability

VC function classes

Definition

The *subgraph* of a function $f : \mathcal{X} \to \mathbb{R}$ at level t is the set $S_{f,t} := \{x \mid f(x) \leq t\}$. The *subgraph class* of a collection $\mathcal{F}$ is the collection

$$\mathcal{S}(\mathcal{F}) := \{S_{f,0}\}_{f \in \mathcal{F}}.$$

The collection $\mathcal{F} \subset \mathcal{X} \to \mathbb{R}$ has VC-dimension $\mathrm{VC}(\mathcal{S}(\mathcal{F}))$.

- linear discriminators $\mathcal{F} = \{f(x) = \mathrm{sign}(x^T\theta)\}$
- ellipsoidal discriminators
  $\mathcal{F} = \{f(x) = \mathrm{sign}((x - x_0)^T\Sigma^{-1}(x - x_0) - b)\}$

Preservation of VC-dimension

- often useful to build up VC classes from smaller ones

Proposition (van der Vaart and Wellner 1996, Lemma 2.6.17)

Let $\mathcal{C}, \mathcal{D}$ be VC-classes of sets. The following are VC-classes:

(i) $\mathcal{C}^c = \{C^c \mid C \in \mathcal{C}\}$, *and* $\mathrm{VC}(\mathcal{C}^c) = \mathrm{VC}(\mathcal{C})$

(ii) $\mathcal{C} \sqcap \mathcal{D} := \{C \cap D \mid C \in \mathcal{C}, D \in \mathcal{D}\}$, *and* $\mathrm{VC}(\mathcal{C} \sqcap \mathcal{D}) \lesssim \mathrm{VC}(\mathcal{C}) + \mathrm{VC}(\mathcal{D})$

(iii) $\mathcal{C} \sqcup \mathcal{D} := \{C \cup D \mid C \in \mathcal{C}, D \in \mathcal{D}\}$, *and* $\mathrm{VC}(\mathcal{C} \sqcup \mathcal{D}) \lesssim \mathrm{VC}(\mathcal{C}) + \mathrm{VC}(\mathcal{D})$

VC classes from vector spaces

Proposition (Finite-dimensional vector spaces)

Let $\mathcal{G}$ be a d-dimensional vector space of functions $\mathcal{X} \to \mathbb{R}$. Then the subgraph class $\mathcal{S}(\mathcal{G})$ has $\mathsf{VC}(\mathcal{S}(\mathcal{G})) \leq d$.