Lecture 23

Last time

- Vapnik-Chervonenkis (VC) Theory.

Today

- Controlling the VC dimension
- Generalization for linear models.

Controlling the VC dimension

Let G be a class of real valued functions $\{g : \mathcal{X} \to \mathbb{R}\}$. Define

$$S_g = \{x \in \mathcal{X} \mid g(x) \leq 0\}$$

$$S(G) = \{S_g \mid g \in G\}.$$

Proposition: Let G be a vectors space of functions $g : \mathbb{R}^d \to \mathbb{R}$ with $\dim(G) < \infty$. Then,

$$vc(S(G)) \leq \dim(G).$$

Proof: Set $n = \dim(G) + 1$ and suppose $S(G)$ shatters some $X = \{x_1, \ldots, x_n\}$.

Define

$$L : G \to \mathbb{R}^n$$

$$L(g) \longmapsto \begin{pmatrix} g(x_1) \\ \vdots \\ g(x_n) \end{pmatrix}.$$

Since $n > \dim(G)$, there is (why?)

$0 \neq \gamma \in \mathbb{R}^n$ s.t. $\langle \gamma, L(g) \rangle = 0$

for all $g \in G$. Therefore,

$$\underbrace{-\sum_{\underbrace{\{i : \gamma_i \leq 0\}}_{\Delta_1}} \gamma_i g(x_i)}_{T_1} = \underbrace{\sum_{\underbrace{\{i : \gamma_i > 0\}}_{\Delta_2}} \gamma_i g(x_i)}_{T_2} \quad \forall g \in G.$$

WLOG assume $\gamma_i > 0$ for some i. Since $S(G)$ shatters X, there is a function g s.t.

$$g(x_i) \leq 0 \quad \forall i \in \Delta_1 \quad \text{and}$$
$$g(x_i) > 0 \quad \forall i \in \Delta_2$$

so

$$0 \geq T_1 = T_2 > 0. \qquad \lightning$$

$\square$

Example (Halfspaces)

Define

$$S_{a,b} = \{ x \in \mathbb{R}^d \mid \langle a, x \rangle + b \le 0 \}$$

$$S = \{ S_{a,b} \mid a \in \mathbb{R}^d,\ b \in \mathbb{R} \}$$

$$\mathcal{G} = \{ x \mapsto \langle a, x \rangle + b \}.$$

Then, ← HW5

$$\mathrm{VC}(S) \le \dim(\mathcal{G}) = d+1.$$

$\dashv$

Example (Balls):

Define

$$S_{a,b} = \{ x \in \mathbb{R}^d \mid \|x - a\|_2 \le b \}$$

$$S = \{ S_{a,b} \mid a \in \mathbb{R}^d,\ b \ge 0 \}.$$

Consider functions

$$g_{a,b}(x) = \|x - a\|^2 - b^2$$
$$= \|x\|^2 + 2\langle a, x \rangle + \|a\|^2 - b^2.$$

This is not a subspace. But it is contained in one.

The trick is to define

$$g_c(x) = \langle c, \theta(x) \rangle \quad \text{with } c \in \mathbb{R}^{d+2}$$

and

$$\theta(x) = (1, x_1, \dots, x_d, \|x\|_2^2).$$

Then, it is clear that

$$\left\{ g_{a,b} \mid \begin{array}{l} a \in \mathbb{R}^d \\ b \in \mathbb{R} \end{array} \right\} \subseteq \underbrace{\{ g_c \mid c \in \mathbb{R}^{d} \}}_{\text{subspace of dim } d+2}$$

So

$$VC(S) \le d+2.$$

The correct value is $d+1$ (much harder to prove)

As we have seen with these simple examples, we often get guarantees of the form

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum f(x_i) - \mathbb{E} f(x) \right| \right] = O\left(\sqrt{\frac{d \log(n)}{n}} \right)$$

where d is the input dimension. A natural question is

Can we get dimension independent

bounds for interesting $\mathcal{F}$?

Dimension free bounds

We will derive bounds for two types of losses: (1) linear models and (2) strongly convex functions.

Linear models

Consider the problem

ℓ is 1-Lipschitz in its first coordinate

Some set

$$\min_{\theta \in \Theta} \mathbb{E}_{(a,b)\sim \mathbb{P}} \ell(\langle \theta, a\rangle, b)$$

Thus, we aim to prove generalization bounds by controlling $R_n(\mathcal{F})$ of

$$\mathcal{F} = \{ (a,b) \mapsto \ell(\langle \theta, a\rangle, b) \mid \theta \in \Theta \}.$$

Recall that we had

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum \ell(\langle \theta, a_i\rangle, b_i) - \mathbb{E}\, \ell(\langle \theta, a\rangle, b) \right|$$

$$\leq 2 \mathbb{E}_{a_i, b_i \overset{iid}{\sim} \mathbb{P}} \sup \left| \sum_{i=1}^{n} \varepsilon_i \ell(\langle \theta, a_i\rangle, b_i) \right|.$$

In turn we can also prove the same thing without absolute values, which

Prove it

stills gives us interesting information. (why?)

Let $\bar{R}$ and $\bar{R}_n$ be Rademecher complexities without abosolute values.

Theorem (contraction principle): Consider a set $A \subseteq \mathbb{R}^n$ and let $\varphi_i : \mathbb{R} \to \mathbb{R}$ be ρ-Lispschitz functions and let

$$A' = \{ (\varphi_1(a_1), \dots, \varphi_n(a_n)) \mid a \in A \}.$$

Then,

$$\bar{R}(A') = \mathbb{E} \sup_{a \in A'} \sum \varepsilon_i a_i \le \underbrace{\mathbb{E} \sup_{a \in A} \sum \varepsilon_i a_i}_{= \bar{R}(A)} \quad \dashv$$

Notice it is A'.

Notice that this theorem allows us to take $\varphi_i(a) = \ell(\langle \theta, a \rangle, b_i)$ and

$$\bar{R}_n(\mathcal{F}) \le \bar{R}_n(\{ a \mapsto \langle \theta, a \rangle \mid \theta \in \Theta \}).$$

Proof of the contraction principle

WLOG $\rho = 1$. For any $\psi : \mathbb{R}^n \to \mathbb{R}^n$ with ψ_i 1-Lipschitz define

$$\psi(A) = \{ (\psi_1(a_1), \dots, \psi_n(a_n)) \mid a \in A \}.$$

Let $\varphi = (\varphi_1, \dots, \varphi_n)$, and $\tilde{\varphi} = (I, \varphi_2, \dots,$

Let $\varphi = (\varphi_1, \dots, \varphi_n)$, and $\tilde{\varphi} = (I, \varphi_2, \dots, \varphi_n)$. By the permutation invariance of $\bar{R}$, it suffices to prove

(★) $$\bar{R}(\varphi(A)) \leq \bar{R}(\tilde{\varphi}(A)).$$

Expanding

$$\bar{R}(\varphi(A)) = \mathbb{E}_{\varepsilon} \sup_{a \in A} \sum_{i=1}^{n} \varepsilon_i \varphi_i(a_i)$$

Def

$$= \frac{1}{2}\Big[\mathbb{E}_{\varepsilon_{2:n}} \sup_{a \in A} \varphi_1(a_1) + \sum_{i=2}^{n} \varepsilon_i \varphi_i(a_i) + \mathbb{E}_{\varepsilon_{2:n}} \sup_{a \in A} -\varphi_1(a_1) + \sum_{i=2}^{n} \varepsilon_i \varphi_i(a_i)\Big]$$

$$= \frac{1}{2}\Big[\mathbb{E}_{\varepsilon_{2:n}} \sup_{a, \tilde{a} \in A} \varphi_1(a_1) - \varphi(\tilde{a}_1) + \sum_{i=2}^{n} \varepsilon_i(\varphi_i(a_i) + \varphi_i(\tilde{a}_i))\Big]$$

1-Lipschitz

$$\leq \frac{1}{2}\Big[\mathbb{E}_{\varepsilon_{2:n}} \sup |a_1 - \tilde{a}_{a_2}| + \sum_{i=2}^{n} \varepsilon_i(\varphi_i(a_i) + \varphi_i(\tilde{a}_i))\Big]$$

Because of the sup we can ensure $a_1 - \tilde{a}_1 \geq 0$.

$$= \frac{1}{2}\Big[\mathbb{E}_{\varepsilon_{2:n}} \sup a_1 - \tilde{a}_{a_2} + \sum_{i=2}^{n} \varepsilon_i(\varphi_i(a_i) + \varphi_i(\tilde{a}_i))\Big]$$

$$= \frac{1}{2}\Big[\mathbb{E}_{\varepsilon_{2:n}} \sup_{a\in A} a_1 + \sum_{i=2}^{n} \varepsilon_i \varphi_i(a_i)$$

$$+ \mathbb{E}_{\varepsilon_{2:n}} \sup_{a\in A} -a_1 + \sum_{i=2}^{n} \varepsilon_i \varphi_i(a_i)\Big]$$

Def

$$= \bar{R}(\tilde{\varphi}(A)).$$

$\square$