Lecture 6

* Please signup to scribe if you haven't yet!
* Signup for presentation topics tomorrow
* Last time
  - → VC Dimension
  - → VC Theorem

Today:

- → Finish proof of VC Theorem
- → Rademacher complexity

VC Theorem

Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) = d < \infty$.

Then there is a absolute constant $c > 0$ s.t.

$\mathcal{H}$ has uniform convergence property with,

$$n_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq c \cdot \frac{d \log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}$$

Corollary.

$\mathcal{H}$ is agnostically-PAC learnable with

$O\left(\frac{d \log(d/\epsilon) + \log(1/\delta)}{\epsilon^2}\right)$ samples.

Proof

Outline: 1) for any set $C \subseteq X$, $\mathcal{H}_C$: "Effective size" of $\mathcal{H}$

$|\mathcal{H}_C| \simeq |C|^d$.

2) small "effective size" good for union bound to get UC.

Def. (Restriction)

The restriction of class $\mathcal{H}$ to a set of examples $C = \{c_1, \ldots, c_n\} \in X$ is a subset of $\{0,1\}^{|C|}$ given by $\mathcal{H}_C = \{(h(c_1), \ldots, h(c_n)) : h \in \mathcal{H}\}$

Step 1 : Polynomial growth of $\mathcal{H}_C$.

Definition (Growth function)

The growth function of $\mathcal{H}$, $\tau_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, is defined as

$$\tau_{\mathcal{H}}(n) = \max_{C \subseteq X, |C| = n} |\mathcal{H}_C| .$$

If $VCdim(\mathcal{H}) = d$, then $\tau_{\mathcal{H}}(n) = 2^n$ $\forall$ $n \leq d$.

Sauer's lemma gives a good upper bound $\forall$ $n \geq d$.

Lemma : (Sauer's Lemma)

$\forall n$, $VCdim(\mathcal{H}) = d$, $\tau_{\mathcal{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$.

For $n > d+1$, this implies $\tau_{\mathcal{H}}(n) \leq \left(\frac{ne}{d}\right)^d$.

Step 2 : Symmetrization

Lemma : For a class $\mathcal{H}$ with growth function $\Pi_{\mathcal{H}}$,

$$\mathbb{E}_{S\sim D^n} \sup_{h\in\mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \leq \sqrt{\frac{2\log(2\Pi_{\mathcal{H}}(2n))}{n}}.$$

Proof

We will use the idea of symmetrization.

Let $S' = \{(x_i', y_i'), i \in [n]\}$ be identically distributed as S.

Note that $\mathbb{E}_{S'}[\hat{R}_{S'}(h)] = R(h)$.

Therefore

$$\mathbb{E}_S \sup_{h\in\mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| = \mathbb{E}_S \sup_{h\in\mathcal{H}} \left| \mathbb{E}_{S'}[\hat{R}_{S'}(h)] - \hat{R}_S(h) \right|$$

Fix S.

Claim. $\sup_{h\in\mathcal{H}} \left| \mathbb{E}_{S'}[\hat{R}_{S'}(h)] \right| \leq \mathbb{E}_{S'} \sup_{h\in\mathcal{H}} \left| \hat{R}_{S'}(h) \right|$.

Proof.

$|\cdot|$ is a convex function

& sup / max of convex functions is convex

$\therefore$ $\sup_{h\in\mathcal{H}} \left| \mathbb{E}_{S'}[\hat{R}_{S'}(h)] \right|$ is a convex function of $\hat{R}_{S'}(h)$

By Jensen's, claim follows. ∎

(Jensen's, $f(\mathbb{E}(x)) \le \mathbb{E}(f(x))$ if f convex)

$$\therefore \mathbb{E}_S \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \le \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left| \hat{R}_{S'}(h) - \hat{R}_S(h) \right|$$

$$= \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|.$$

Let $\sigma_{1:n} = \{\sigma_1, \ldots, \sigma_n\}$ be a Rademacher r.v. i.e. Unif($\{\pm 1\}$) (independent)

Since (x_i, y_i) & (x_i', y_i') are i.i.d.

$$\therefore \mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\}$$

$$\sim \mathbb{1}\{h(x_i) \neq y_i\} - \mathbb{1}\{h(x_i') \neq y_i'\}.$$

$$\therefore \mathbb{E}_S \sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right|$$

$$\le \mathbb{E}_{\sigma_{1:n}} \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|$$

$$= \mathbb{E}_{S,S'} \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|$$

Now fix s, s' & let C be the set of examples appearing in s, s' (the union). Note $|C| \le 2n$.

Key idea We can replace sup over $\mathcal{H}$ by max over $\mathcal{H}_C$.

$$\mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|$$

$$= \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right) \right|$$

Let $\theta_h = \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right)$

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| \le \mathbb{E}_{S,S'} \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h|$$

Want to bound $\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\theta_h|$

Lemma: (max of sub-Gaussians). If $(X_1, \dots X_m)$ are mean 0 σ-sub-Gaussian (not necessarily independent), then

$$\mathbb{E} \max_i X_i \leq \sigma \sqrt{2 \log(m)}$$

Proof

$$\mathbb{E} \max_i X_i = \frac{1}{\lambda} \log \exp\left(\lambda \mathbb{E}[\max_i X_i]\right) \quad \forall \lambda.$$

$$\leq \frac{1}{\lambda} \log \mathbb{E}[\exp(\lambda \max_i X_i)] \quad \text{(Jensen's)}$$

$$\leq \frac{1}{\lambda} \log \mathbb{E}\left[\sum_{i=1}^m \exp(\lambda X_i)\right]$$

$$= \frac{1}{\lambda} \log \left(\sum_{i=1}^m \mathbb{E}(\exp(\lambda X_i))\right)$$

$$\leq \frac{1}{\lambda} \log \left(\sum_{i=1}^m \exp\left(\frac{\lambda^2 \sigma^2}{2}\right)\right)$$

$$\lambda = \frac{\sqrt{2 \log(m)}}{\sigma}$$

$$\mathbb{E} \max_i X_i \leq \frac{\sigma}{\sqrt{2 \log(m)}} \log \left(\sum_{i=1}^m \exp(\log m)\right)$$

$$= \sigma \cdot \frac{2 \log m}{\sqrt{2 \log m}} = \sigma \sqrt{2 \log m}.$$

∎

Claim: θ_n is sub-Gaussian with parameter $\frac{1}{\sqrt{n}}$, $\mathbb{E}[\theta_n] = 0$.

Proof

$$\theta_n = \sum_{i=1}^{n} \frac{\sigma_i}{n} \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right)$$

$\mathbb{E}[\theta_n] = 0$. $(\sigma_i \sim \text{Unif}(\{\pm 1\}))$

$$\theta_n = \sum_{i=1}^{n} \underbrace{\frac{\sigma_i}{n} \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right)}$$

each term is sub-Gaussian with parameter $\frac{1}{n}$.

$\therefore$ θ_n is sub-Gaussian with parameter $\left(\sum_{i=1}^{n} \frac{1}{n^2} \right)^{1/2}$

$= \frac{1}{\sqrt{n}}$.

(θ_n is sum of independent sub-Gaussians with parameter $\frac{1}{n}$).

Claim: $\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_c} |\theta_n| \leq \frac{1}{\sqrt{n}} \sqrt{2 \log(2|\mathcal{H}_c|)}$

Proof

$$\mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_c} |\theta_n| = \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_c} \max\{\theta_n, -\theta_n\}$$

Recall that if θ_n is sub-Gaussian

$\Rightarrow$ $-\theta_n$ is sub-Gaussian (with same parameter)

$$\therefore \mathbb{E}_{\sigma_{1:n}} \max_{h \in \mathcal{H}_C} |\sigma_h| \leq \frac{1}{\sqrt{n}} \sqrt{2 \log(2|\mathcal{H}_C|)}$$

We have shown

$$\mathbb{E}_{S,S'}\left[\mathbb{E}_{\sigma_{1:n}} \max_{h} |\sigma_h|\right] \leq \sqrt{\frac{2\log(2|\mathcal{H}_C|)}{n}}$$

$$\therefore \mathbb{E}_S \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| \leq \sqrt{\frac{2\log(2|\mathcal{H}_C|)}{n}}$$

Note that $|\mathcal{H}_C| \leq \tau_{\mathcal{H}}(2n)$

since $|C| = 2n$.

Finishes proof by plugging in bound.

Step 3. McDiarmids.

Define $f(S) = \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)|$

observe that $f(S)$ satisfies the bounded different property with constant $\frac{1}{n}$.

(since changing (x_i, y_i) can only change $\hat{R}_S(h)$ by $\frac{1}{n}$ for any $h \in \mathcal{H}$ $\therefore$ the max also changes by atmost $\frac{1}{n}$)

Using McDiarmid's

$$\Pr\left[f(S) - \mathbb{E}[f(S)] > t \right] \leq 2\exp\left(-2nt^2\right)$$

Choose $t = \sqrt{\frac{\log(2/\delta)}{2n}}$ to get failure probability δ.

$\therefore$ with prob $1-\delta$

$$\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| < \sqrt{\frac{2\log\left(2\,\Gamma_{\mathcal{H}}(2n)\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

Step 4. Finish proof.

Using Sauer's lemma, for $n > d+1$, $\Gamma_{\mathcal{H}}(n) \leq \left(\frac{ne}{d}\right)^d$.

Plugging this in, with probability $(1-\delta)$,

$$\sup_{h \in \mathcal{H}} \left| R(h) - \hat{R}_S(h) \right| \leq \sqrt{\frac{2d\log(2ne/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

$\therefore$ for $n \geq O\left(\frac{d\log(d/\varepsilon) + \log(1/\delta)}{\varepsilon^2}\right)$ the RHS $\leq \varepsilon$.

[Exercise: finish this]

Rademacher complexity

Let us recall the proof of the VC theorem.

We wanted to bound

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)|.$$

This quantity is called an empirical process

Empirical process theory studies such quantities.

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} \left(R(h) - \hat{R}_S(h) \right) \leq \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \left(\hat{R}_{S'}(h) - \hat{R}_S(h) \right)$$

$$= \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right)$$

$$= \mathbb{E}_{\sigma_{1:n}} \mathbb{E}_{S,S'} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{1}\{h(x_i') \neq y_i'\} - \mathbb{1}\{h(x_i) \neq y_i\} \right)$$

$$\leq \mathbb{E}_{S'} \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}\{h(x_i') \neq y_i'\}$$

$$+ \mathbb{E}_S \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (-\sigma_i) \mathbb{1}\{h(x_i) \neq y_i\}$$

$$\mathbb{E}_S \sup_{h \in \mathcal{H}} R(h) - \hat{R}_S(h) \leq 2 \mathbb{E}_S \mathbb{E}_{\sigma_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}\{h(x_i) \neq y_i\}$$

Let $Z = X \times Y$

$\mathcal{F}$: function class $Z \to \mathbb{R}$

D : distribution over Z.

Definition (Rademacher Complexity).

Let $\mathcal{F}$ be a family of real-valued functions $f: Z \to \mathbb{R}$ where $Z = X \times Y$. Then the Rademacher complexity $R(\mathcal{F})$ is defined as

$$R(\mathcal{F}) = \frac{1}{n} \mathbb{E}_{\sigma \sim \{\pm 1\}^n} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(z_i) \right].$$

More generally, given a set of vectors $A \subset \mathbb{R}^n$, the Rademacher complexity $R(A)$ is defined as

$$R(A) = \frac{1}{n} \mathbb{E}_{\sigma \sim \{\pm 1\}^n} \sup_{a \in A} \sum_{i=1}^{n} \sigma_i a_i.$$

Intuition: $R(\mathcal{F})$ captures how well can $\mathcal{F}$ fit random noise.

If $\mathcal{F}$ can fit random noise, $\mathcal{F}$ will probably overfit.