Chapter 26. Rademacher Complexity.

- we learned in lecture 3 that

Uniform Convergence $\Longrightarrow$ PAC learnable with ERM

$\triangleq$ $\mathcal{H}$ is uniformly convergent

if $|L_S(h) - L_D(h)| \le \varepsilon$ w.p $1-\delta$ for all $h \in \mathcal{H}$

when $m \ge m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ for some $m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$.

Recall:

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell_h(z_i)$$

$$L_D(h) = \mathbb{E}_{z \sim D}[\ell_h(z)]$$

$\uparrow$ generalize

VC-dim: 0-1 loss

- Definition [Representativeness]

Representativeness of S w.r.t $\mathcal{H}$ is defined as

$$Rep_D(\mathcal{H}, S) \triangleq \sup_{h \in \mathcal{H}} L_D(h) - L_S(h)$$

Since this cannot be measured from sample S, consider a surrogate as follows.

*Another way to look at symmetrization.

- Partition S into $S = S_1 \cup S_2$ & $|S_1| = |S_2|$, and consider

$$\sup_{h \in \mathcal{H}} L_{S_1}(h) - L_{S_2}(h)$$

- write it as

$$\frac{2}{m} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \ell_h(z_i) \quad , \text{ where}$$

$\sigma_i = \begin{cases} +1 & \text{if } i \in S_1 \\ -1 & \text{if } i \in S_2 \end{cases}$ and $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_m)$

"Rademacher Complexity" uses this surrogate with random σ

$\sigma_i = \begin{cases} +1 & \text{w.p } 1/2 \\ -1 & \text{w.p } 1/2. \end{cases}$

<u>Rademacher Complexity</u> of $\mathcal{H}$ w.r.t. S is defined as

$$R(\ell \circ \mathcal{H} \circ S) \triangleq \frac{1}{m} \mathbb{E}_{\sigma \sim U[\{\pm 1\}^m]} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \cdot \ell_h(z_i) \right]$$

Example> bounded linear class

$$\mathcal{H}_2 \triangleq \{ \underbrace{\langle w, x \rangle}_{h_w(x)} : \|w\|_2 \le 1 \}$$

Claim> for $S = \{x_1, \dots, x_m\}$, $x_i \in \mathbb{R}^n$,

$$R(\mathcal{H}_2 \circ S) \le \frac{\max_i \|x_i\|_2}{\sqrt{m}} \qquad (**)$$

Proof> Cauchy-Schwarz: $\langle w, v \rangle \le \|w\| \cdot \|v\|$

$$m \cdot R(\mathcal{H}_2 \circ S) = \mathbb{E}_\sigma \left[\sup_{w : \|w\| \le 1} \sum_{i=1}^{m} \sigma_i \cdot \langle w, x_i \rangle \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{w:\|w\|\le 1} \langle w, \sum_{i=1}^m \sigma_i x_i \rangle \right]$$
$$\le \mathbb{E}_\sigma \left[\sup_{w:\|w\|\le 1} \|w\| \cdot \|\sum_{i=1}^m \sigma_i x_i\| \right]$$
$$= \mathbb{E}_\sigma \left[\|\sum_{i=1}^m \sigma_i x_i\| \right]$$

Using Jensen's ineq., $\sqrt{a}$ is concave in a

$$\mathbb{E}_\sigma\left[\|\sum_{i=1}^m \sigma_i x_i\|\right] = \mathbb{E}_\sigma\left[\left(\|\sum_{i=1}^m \sigma_i x_i\|_2^2\right)^{1/2}\right] \le \left(\mathbb{E}_\sigma\left[\|\sum_{i=1}^m \sigma_i x_i\|_2^2\right]\right)^{1/2}$$

Finally

$$\mathbb{E}_\sigma\left[\|\sum_{i=1}^m \sigma_i x_i\|^2\right] = \mathbb{E}_\sigma\left[\langle \sum_{i=1}^m \sigma_i x_i, \sum_{j=1}^m \sigma_j x_j \rangle\right]$$
$$= \mathbb{E}_\sigma\left[\sum_{i=1}^m \sigma_i^2 \langle x_i, x_i\rangle + \underbrace{\sum_{i\ne j} \sigma_i\sigma_j \langle x_i, x_j\rangle}_{\mathbb{E}[\ ]=0} \right]$$
$$= \sum_{i=1}^m \langle x_i, x_i \rangle \le m \max_i \|x_i\|^2$$

[Theorem 26.5] Suppose $|l_h(z)| \le C$ for all $h \in \mathcal{H}$ and z,

(data independent bound)

① with probability at least $1-\delta$, for all $h \in \mathcal{H}$

$$\underbrace{L_D(h) - L_S(h)}_{R.V} \le \underbrace{2 \cdot \mathbb{E}_{S'}\left[R(l \circ \mathcal{H} \circ S')\right] + c\sqrt{\frac{2\ln(2/\delta)}{m}}}_{\text{deterministic}}$$

S ‖

② (data dependent bound)

$$L_D(h) - L_S(h) \le 2 \cdot \overset{R.V}{\overset{\downarrow}{R(l \circ \mathcal{H} \circ S)}} + 4c \cdot \sqrt{\frac{2\ln(4/\delta)}{m}}$$

③ (data dependent bound) $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} L_D(h)$

$$L_D(h_{ERM(S)}) - L_D(h^*) \le 2 \cdot R(l \circ \mathcal{H} \circ S) + 5c\sqrt{\frac{2\ln(8/\delta)}{m}}$$

e.g. from (**) for Lipschitz & bounded l: $|l(z)| \le C$, $|l(x) - l(y)| \le \rho\|x-y\|$

$$R(\mathcal{H}_2 \circ S) \le \frac{\max_i \|x_i\|_2}{\sqrt{m}}$$

← Data dependent. also n-independent.

$$R(l \circ \mathcal{H}_2 \circ S) \le \rho \cdot \frac{\max_i \|x_i\|_2}{\sqrt{m}}$$ if $l(\cdot, y)$ is ρ-Lipschitz ← Skip Proof

$$\to L_D(h) - L_S(h) \le 2 \cdot \rho \cdot \frac{\max_i \|x_i\|_2}{\sqrt{m}} + 4c \cdot \sqrt{\frac{2 \cdot \ln(4/\delta)}{m}}$$

[Remark 26.2]
$$\mathbb{E}_S[\mathrm{Rep}_D(\mathcal{H},S)] \le 2\cdot \mathbb{E}_S[R(\ell\circ\mathcal{H}\circ S)] \qquad (*)$$

[McDiarmid's inequality]

Let V be a set and $f : V^m \to \mathbb{R}$, $(x_1,\dots,x_m) \mapsto f(x_1,\dots x_m)$, such that for some $c>0$, for all $i\in[m]$,

and for all $(x_1,\dots x_m)\in V^m$, and $x_i'\in V$, we have

$$|f(x_1,\dots x_{i-1},x_i,x_{i+1},\dots,x_m) - f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_m)| \le c. \leftarrow \text{bounded difference condition } (***)$$

Let $X_1,\dots,X_m$ be m independent random variables taking values in V.

Then,

$$\mathbb{P}(|f(X^m) - \mathbb{E}[f(X^m)]|>\varepsilon) \le 2e^{-\frac{2\varepsilon^2}{mc^2}}$$

$\iff$ with probability $1-\delta$,

$$|f(X^m) - \mathbb{E}[f(X^m)]| \le c\sqrt{\ln\left(\tfrac{2}{\delta}\right)\cdot\tfrac{m}{2}}$$

e.g. $f(X^m)=\frac{1}{m}\sum_i X_i$, $a\le X_i\le b \to c=\frac{|b-a|}{m}$

$\to \mathbb{P}(|\quad|>\varepsilon)\le 2e^{-\frac{2\varepsilon^2 m}{(b-a)^2}}$

$\to$ Recovers Hoeffding's as a special case!

proof of [Theorem 26.5]

$\mathrm{Rep}_D(\mathcal{H},S) = \sup_{h\in\mathcal{H}} L_D(h) - L_S(h)$ satisfies $(***)$ with constant $\frac{2c}{m}$.

Applying McDiarmid's inequality, & $(*)$, with probability $1-\delta$

$$\mathrm{Rep}_D(\mathcal{H},S) \le \mathbb{E}_{S'}[\mathrm{Rep}_D(\mathcal{H},S')] + c\sqrt{\frac{2\ln(2/\delta)}{m}}$$
$$\le 2\,\mathbb{E}_{S'}[R(\ell\circ\mathcal{H}\circ S')] + c\sqrt{\frac{2\ln(2/\delta)}{m}}$$

This proves ①

proof of ② similarly follows from the fact that $R(\ell\circ\mathcal{H}\circ S)$ satisfies $(***)$ with $\frac{2c}{m}$ also.

$$\to 2\,\mathbb{E}_{S'}[R(\ell\circ\mathcal{H}\circ S')] \le 2R(\ell\circ\mathcal{H}\circ S) + 2c\sqrt{\frac{2\ln(2/\delta)}{m}} \quad \text{\& Union Bound for } \delta.$$

Similarly for ③

We are left to prove $(*)$ $\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} L_D(h)-L_S(h)\right] \le 2\cdot\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^m \ell_h(z_i)\sigma_i\right]\frac{1}{m}$

proof of $(*)$ > by symmetrization.

$$L_D(h) - L_S(h) = \mathbb{E}_{S'}[L_{S'}(h)] - L_S(h) = \mathbb{E}_{S'}[L_{S'}(h) - L_S(h)]$$

take sup

$$\to \sup_{h\in\mathcal{H}} L_D(h) - L_S(h) = \sup_{h\in\mathcal{H}} \mathbb{E}_{S'}[L_{S'}(h) - L_S(h)]$$
$$\le \mathbb{E}_{S'}\left[\sup_{h\in\mathcal{H}} L_{S'}(h) - L_S(h)\right]$$

take $\mathbb{E}_S$

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} L_D(h) - L_S(h)\right] \le \mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} L_{S'}(h) - L_S(h)\right]$$

$$= \frac{1}{m}\,\underbrace{\mathbb{E}_{S,S'}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \left(\ell_h(z_i') - \ell_h(z_i)\right)\right]}$$

Notice that $\ell_h(z_i') - \ell_h(z_i) \overset{dist}{=} \sigma_i\left(\ell_h(z_i') - \ell_h(z_i)\right)$ for $\sigma_i = \begin{cases} +1 & \text{w.p. } 1/2 \\ -1 & 1/2 \end{cases}$

then $\displaystyle \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \ell_h(z_i') - \ell_h(z_i) \overset{dist}{=} \sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i\left(\ell_h(z_i') - \ell_h(z_i)\right)$

and $\displaystyle \boxed{\mathbb{E}_{SS'}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \ell_h(z_i') - \ell_h(z_i)\right]} = \mathbb{E}_{S,S',\sigma}\left[\sup_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i\left(\ell_h(z_i') - \ell_h(z_i)\right)\right]$

Finally, $\displaystyle \boxed{\sup_{h\in\mathcal{H}} \sum_i \sigma_i\left(\ell_h(z_i') - \ell_h(z_i)\right)} \le \sup_{h\in\mathcal{H}} \sum_i \sigma_i\, \ell_h(z_i') + \sup_{h'\in\mathcal{H}} \sum_i \left(-\sigma_i\, \ell_{h'}(z_i)\right)$

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}} L_D(h) - L_S(h)\right] \le \frac{2}{m}\,\mathbb{E}_{S,\sigma}\left[\sup_{h\in\mathcal{H}} \sum_i \sigma_i\, \ell_h(z_i)\right] = 2\,\mathbb{E}_S\left[R(\ell\circ\mathcal{H}\circ S)\right]$$