Section 5.4 Empirical Risk Minimization (ERM)

Consider $(X, Y, U, P, \mathcal{F}, \ell)$ on $Z = X \times Y$

$f : Z \to U$

$\ell(y, u) = $ loss $\in [0,1]$

Empirical risk: for some f Z^n

$Z_i := (X_i, Y_i)$

$$\frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) = P_n(\ell_f)$$

Want to find f to minimize

$$L_P(f) = E_P[\ell(Y, f(X))]$$

ERM- $\hat{f} = \arg\min \{\text{Empirical risk}\}$.

Let $\ell_f(x, y) \triangleq \ell(y, f(x))$ $L_P(f) = E_P[\ell_f]$

$= P(\ell_f)$

P_n = empirical dist$^{\underline{n}}$ of Z^n $Z_i = (X_i, Y_i)$

$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{[Z_i \in A]}$ $P_n(\ell_f) = \frac{1}{n} \sum_{i=1}^{n} \ell_f(Z_i)$

e.g. if $\ell(y,u) = (y-u)^2$ $\quad \ell_f(x,y) = (y-f(x))^2$

$\mathcal{F}$ $\quad \{\ell_f : f \in \mathcal{F}\}$ $\quad \underbrace{[a_i, b_i]}_{n} = [0,1]_{n}$

Hoeffding inequality

$$P\{|P_n(\ell_f) - P(\ell_f)| \geq \varepsilon\} \leq 2e^{-2n\varepsilon^2}$$

for any f.

$$g(n,\varepsilon) \equiv \sup_{P} P^n\{Z^n : \sup_{f\in\mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \geq \varepsilon\}$$

<u>Theorem 5.3</u> If $g(n,\varepsilon) \to 0$ as $n\to\infty$ for every $\varepsilon > 0$, then ERM is PAC.

(Short proof : Mismatch min lemma.)

(W.h. prob. $1-\overset{\delta}{g(n,\varepsilon)}$

$\sup_{f\in\mathcal{F}} |P_n(\ell_f) - P(\ell_f)| \leq \varepsilon$ for all f)

$$L_P(\hat{f}_n) - L_P^*(\mathcal{F}) = P(\ell_{\hat{f}_n}) - P(\ell_{f^*})$$

$$= P(\ell_{\hat{f}_n}) - P_n(\ell_{\hat{f}_n}) + \underbrace{P_n(\ell_{\hat{f}_n}) - P_n(\ell_{f^*})}_{\leq 0} + P_n(\ell_{f^*}) - P(\ell_{f^*})$$

$$L_P(\hat{f}_n) - L_P^*(\mathcal{F})$$

$$\leq \left| P(l_{\hat{f}_n}) - P_n(l_{\hat{f}_n}) \right| + \underbrace{\cancel{P_n(l_{\hat{f}_n}) - P_n(l_{f^*})}}_{\leq 0} + \left| P_n(l_{f^*}) - P(l_{f^*}) \right|$$

$$\leq \varepsilon + 0 + \varepsilon \leq 2\varepsilon$$

Chapter 6 ... Rademacher averages

6.1 A new abstract framework for ERM.

Given $(Z, \mathcal{F}, \mathcal{P})$ (old l_f's)

$\mathcal{F}$ = a set of functions $f: Z \to [0,1]$

$\mathcal{P}$ is a set of probability distr^n^s on Z

Given $P \in \mathcal{P}$, $Z^n = (Z_1, \dots, Z_n)$ is a vector of n independent draws from P.

Goal: find $f^* \in \mathcal{F}$ to minimize $P(f^*)$

ERM: $\hat{f} \in \arg\min_{f \in \mathcal{F}} P_n(f)$ $\left(= \frac{1}{n}\sum_{i=1}^{n} f(Z_i) \right)$

$$\Delta_n(Z^n) = \|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |P_n(f) - P(f)|$$

Proposition 6.1

$$P(\hat{f}_n) \le \underbrace{L^*(\mathcal{F})}_{= P(f^*)} + 2\Delta_n(Z^n)$$

6.2 Definition Let $A \subset \mathbb{R}^n$, A bounded

The Rademacher average of A is

$$R_n(A) = E_\varepsilon\left[\sup_{a \in A} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i a_i \right|\right] \quad \langle \varepsilon, a \rangle$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent

$$P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$$

$$R_n^\circ(A) = E\left[\sup_{a \in A} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i a_i\right]$$

$$A = \{ f(Z_1), \dots, f(Z_n) : f \in \mathcal{F} \} = \mathcal{F}(Z^n)$$

Motivation

Let $P \in \mathcal{P}$.

We want to show $\Delta_n(Z^n)$ is small with high probability.

If true, then for <u>all</u> f,

$$\frac{1}{n}\sum_{i=1}^{n} f(Z_i) \approx P(f) = E[f(Z_i)]$$

and

$$\frac{1}{n}\sum_{i=n+1}^{2n} f(Z_i) \approx P(f)$$

$Z_{n+1}, \cdots, Z_{2n}$ same dist as $Z_1, \cdots Z_n$, all indep.

$$= 2 \cdot \frac{1}{2n}\left[\sum_{i=1}^{n} f(Z_i) - \sum_{i=n+1}^{2n} f(Z_i)\right] \approx 0 \quad \text{for all } f$$

$$= 2 \cdot \frac{1}{2n}\sum_{i=1}^{2n} \varepsilon_i f(Z_i)$$

$\varepsilon_i = +1$ if $1 \le i \le n$
$\quad\;\; -1$ if $n+1 \le i \le 2n$

expect $2E\left[\sup_f \left|\frac{1}{2n}\sum_{i=1}^{n} \varepsilon_i f(Z_i)\right|\right] \approx 0$ (i.e. be small

Theorem 6.1 $\mathcal{F}$ $f : Z \to [0,1]$.

Then for any $P \in \mathcal{P}(Z)$:

$$E[\Delta_n(Z^n)] \le 2 E \underline{R_n}(\mathcal{F}(Z^n))$$

$$\mathcal{F}(Z^n) = \{ (f(z_1), \dots f(z_n)) : f \in \mathcal{F} \}$$

Figure 1 $A \subset \mathbb{R}^2$