Empirical Risk Minimization (ERM)

Recap: $(X_1, Y_1), \dots, (X_n, Y_n)$ – data

$f \in \mathcal{F}$ – candidate hypothesis

$P_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$ – empirical risk of f

---

ERM: choose $\hat{f}_n \in \mathcal{F}$ s.t. $P_n(\hat{f}_n) \leftarrow \min$

$$P_n(\hat{f}_n) = \min_{f \in \mathcal{F}} P_n(f)$$

PAC: $$\lim_{n \to \infty} \sup_P P^n \left\{ P(\hat{f}_n) \geq \inf_{f \in \mathcal{F}} P(f) + \varepsilon \right\} = 0 \quad \forall \varepsilon > 0$$

where $\hat{f}_n$ is returned by ERM algo.

Sufficient condition: Uniform Convergence of Empirical Means (UCEM)

$$\lim_{n \to \infty} \sup_P P^n \left\{ \sup_{f \in \mathcal{F}} |P_n(f) - P(f)| \geq \varepsilon \right\} = 0 \quad \forall \varepsilon > 0$$

(property of ℓ and $\mathcal{F}$ only)

UCEM $\Longrightarrow$ ERM is PAC

Notation:

(X, Y) $\qquad$ Z

$\ell(Y, f(X))$ $\qquad$ $f(Z)$

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

$P(f) = \mathbb{E}_P[\ell(Y, f(X))]$ $\qquad$ $P(f) := \mathbb{E}_P[f(Z)]$

ERM: $\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} P_n(f)$

$P_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \in A\}}$ $\qquad$ $Z^n = (Z_1, \ldots, Z_n)$

— empirical distribution of Z^n

$P(A) = P[Z \in A]$

$\mathbb{E} P_n(A) = P(A)$ $\quad \forall$ event A

$\mathbb{E} P_n(f) = P(f)$

$0 \le f \le 1$: $\mathbb{P}\{|P_n(f) - P(f)| \ge \varepsilon\} \le 2e^{-2n\varepsilon^2}$ (Hoeffding)

P, P' : two possible prob. dist. for Z

$$\left.\begin{aligned} P(f) &:= \mathbb{E}_P[f(Z)] \\ P'(f) &:= \mathbb{E}_{P'}[f(Z)] \end{aligned}\right\} \forall f \in \mathcal{F}$$

$\sup_{f \in \mathcal{F}} |P(f) - P'(f)| =: \|P - P'\|_{\mathcal{F}}$

$\ell^\infty(\mathcal{F})$ seminorm: $\|P - P'\|_{\mathcal{F}} \le \|P - P''\|_{\mathcal{F}} + \|P'' - P'\|_{\mathcal{F}}$

$\|P - P'\|_{\mathcal{F}} = 0 \not\Rightarrow P = P'$

Key quantity: $\|P_n - P\|_{\mathcal{F}}$

$$P_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \in \cdot\}}$$

$$P(\cdot) = P[Z \in \cdot]$$

UCEM: $\|P_n - P\|_{\mathcal{F}} \to 0$ as $n \to \infty$
in probability, uniformly in P

Thm (from last lecture – new notation)
Let $\hat{f}_n$ be a random element of $\mathcal{F}$ computed from data $Z_1, \dots, Z_n \overset{iid}{\sim} P$.

1) $P(\hat{f}_n) \le P_n(\hat{f}_n) + \|P_n - P\|_{\mathcal{F}}$

2) if $\hat{f}_n$ minimizes $P_n(\cdot)$ on $\mathcal{F}$ [ERM], then

$$P(\hat{f}_n) \le \inf_{f \in \mathcal{F}} P(f) + 2\|P_n - P\|_{\mathcal{F}}.$$

Goal: get a handle on $\|P_n - P\|_{\mathcal{F}}$.

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(Z_i) - P(f) \right|$$

$P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(Z_i)$
$Z_1, \dots, Z_n \overset{iid}{\sim} P$

– f_{cn} of $Z_1, \dots, Z_n$,
has bdd differences $c_1 = \dots = c_n = \frac{1}{n}$

$$\Rightarrow P^n\left\{ \|P_n - P\|_{\mathcal{F}} \ge \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \varepsilon \right\} \le e^{-2n\varepsilon^2}$$

by McDiarmid

or: $\|P_n - P\|_{\mathcal{F}} \le \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}$ w.p. $\ge 1-\delta$

How big is $\mathbb{E}\|P_n - P\|_{\mathcal{F}}$?

- symmetrization (Vapnik-Chervonenskis, 1970s; Giné-Zinn, 1980s)

$\mathbb{E}P_n(f) = P(f) \quad \forall f$

$Z_1, \dots, Z_n \overset{iid}{\sim} P$

$\bar{Z}_1, \dots, \bar{Z}_n \overset{iid}{\sim} P$, $\perp\!\!\!\perp Z_1, \dots, Z_n$ (ghost sample)

$\bar{P}_n(f) := \frac{1}{n}\sum_{i=1}^n f(\bar{Z}_i)$

$P_n(f) \approx P(f)$, $\bar{P}_n(f) \approx P(f)$ w.h.p.

$\Longrightarrow$ can expect $P_n(f) - \bar{P}_n(f) \approx 0$ w.h.p.

$\frac{1}{n}\sum_{i=1}^n \{f(Z_i) - f(\bar{Z}_i)\}$ should have many "cancellations"

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |P_n(f) - \mathbb{E}\bar{P}_n(f)|$$

$$\le \sup_{f\in\mathcal{F}} \bar{\mathbb{E}} |P_n(f) - \bar{P}_n(f)|$$

$\bar{\mathbb{E}}[\cdot]$ - w.r.t. $\bar{Z}_1, \dots, \bar{Z}_n$

$$\le \bar{\mathbb{E}}\left\{\sup_{f\in\mathcal{F}} |P_n(f) - \bar{P}_n(f)|\right\}$$

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} \le \mathbb{E}\bar{\mathbb{E}}\|P_n - \bar{P}_n\|_{\mathcal{F}} \qquad Z^n, \bar{Z}^n \text{ iid}$$

$$\mathbb{E}\,\|P_n - P\|_{\mathcal{F}} \le \mathbb{E}\,\|P_n - \bar{P}_n\|_{\mathcal{F}}$$

$$= \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} [f(Z_i) - f(\bar{Z}_i)] \right| \right\}$$

$$\forall_{f,i}: \quad f(Z_i) - f(\bar{Z}_i) \overset{d}{=} f(\bar{Z}_i) - f(Z_i)$$

a r.v. U is symmetric if $U \overset{d}{=} -U$

Fact: if U is symmetric, then $U \overset{d}{=} \varepsilon U$ where $\varepsilon \perp\!\!\!\perp U$ and takes values ± 1 w. equal prob.

[ε is a Rademacher r.v.]

$\Rightarrow$ if $\varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} \mathrm{Rad} \perp\!\!\!\perp Z^n, \bar{Z}^n$

then $\left(f(Z_i) - f(\bar{Z}_i)\right)_{i=1}^{n} \overset{d}{=} \left(\varepsilon_i (f(Z_i) - f(\bar{Z}_i))\right)_{i=1}^{n}$

$$\therefore \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} [f(Z_i) - f(\bar{Z}_i)] \right| \right\}$$

$$= \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (f(Z_i) - f(\bar{Z}_i)) \right| \right\} \to \mathbb{E}_{\varepsilon^n, Z^n, \bar{Z}^n}$$

$$\le \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) \right| \right\} + \mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(\bar{Z}_i) \right| \right\}$$

$$= 2\,\mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) \right| \right\} \to \mathbb{E}_{\varepsilon^n, Z^n}$$

Thm (Giné-Zinn) $\mathbb{E}\,\|P_n - P\|_{\mathcal{F}} \le 2\,\mathbb{E}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f(Z_i) \right| \right\}$

Rademacher averages (or complexities)

Def A – bdd subset of $\mathbb{R}^n$

Rademacher average (or complexity) of A:

$$R_n(A) := \mathbb{E}\left\{ \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i a_i \right| \right\}$$

where $a = (a_1, \dots, a_n)^T$ are elements of A, and $\varepsilon_1, \dots, \varepsilon_n$ are iid Rad r.v.'s

Learning setting:

$z_1, \dots, z_n$; $\mathcal{F}$ – fcns, $f(z_i) \in [0,1]$, $\forall i$

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

– bdd subset of $\mathbb{R}^n$

$[\mathcal{F}(z^n) \subseteq [0,1]^n]$

$$R_n(\mathcal{F}(z^n)) = \mathbb{E}_\varepsilon \left\{ \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(z_i) \right| \right\}$$

$R_n(\mathcal{F}(z^n))$ ↓ fixed

Giné-Zinn: $\mathbb{E}\|P_n - P\|_{\mathcal{F}} \le 2\,\mathbb{E} R_n(\mathcal{F}(z^n))$