Ch4 Model Selection

Sep / 26 / 2023

§1 Motivation / Overview

1. $S=((x_1,y_1),\cdots,(x_m,y_m))\sim D^m$, $y_i\in\{\pm1\}$.
   $\mathcal{H}$. Find good $h\in\mathcal{H}$ to reduce $R(h)=\mathop{\mathbb{E}}_{(x,y)\sim D}[\mathbb{1}_{h(x)\neq y}]$.

2. Reduce to ERM. $\operatorname{argmin}_{h\in\mathcal{H}} \hat{R}(h)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{\{h(x_i)\neq y_i\}}$.

3. Impose a structure on $\mathcal{H}$.

   $\{\mathcal{H}_k\}_{k\in\mathbb{N}}$ be $\mathcal{H}_k\subseteq\mathcal{H}_{k+1}$, and $\mathcal{H}=\bigcup_{k\in\mathbb{N}}\mathcal{H}_k$.

   Search for a good hypothesis set $\mathcal{H}_k$ and also for a good hypothesis $h\in\mathcal{H}_k$.

   - Some algorithm do this simultaneously:
     - Structural Risk Minimization (SRM)
     - Regularization
   - Some does one after another
     - Cross Validation (CV)
     - n-fold CV

   $\mathcal{H}_k$: model.

4. Convex Surrogate Loss.

   $h\in\mathcal{H}_k$, $f\in[\mathcal{X}\to\mathbb{R}]$, $h_f(x)=\operatorname{sgn}(f(x))$.

   $\hat{R}_S(h_f)\overset{?}{\longleftrightarrow}\hat{R}_S^{\Phi}(f)$

§2 Analysis of Error

1. $D \in Pr(\mathcal{X} \times \mathcal{Y})$, $S \in D^m$, $\mathcal{H}$.

2. $h_{Bayes} \in [\mathcal{X} \to \mathcal{Y}]$, which has the minimum $R^* = R(h_{Bayes})$ — Bayes Error

$$h_{Bayes}(x) = \begin{cases} +1 & \text{if } \mathbb{P}[y=+1|x] \geq \mathbb{P}[y=-1|x] \\ -1 & \text{otherwise.} \end{cases}$$

3. $$R(h_{Bayes}) = \mathop{\mathbb{E}}_{x \sim D_{\mathcal{X}}} [\min\{\mathbb{P}[y=+1|x], \mathbb{P}[y=-1|x]\}].$$

4. $h \in \mathcal{H}$ ··· Result of algorithm

Excess error : $R(h) - R^* = \underbrace{R(h) - \inf_{h \in \mathcal{H}} R(h')}_{\geq 0 \text{ (estimation error)}} + \underbrace{\inf_{h' \in \mathcal{H}} R(h') - R^*}_{\geq 0 \text{ (approximation error)}}$

§3 Empirical Risk Minimization Formally

1. $D \in Pr(\mathcal{X} \times \mathcal{Y})$, $S \sim D^m$, $\mathcal{H}$

2. $$h_S^{ERM} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{R}_S(h).$$

3. What can we say about $R(h_S^{ERM})$?

Theorem

$$\mathop{\mathbb{P}}_{S \sim D^m} \left[R(h_S^{ERM}) \leq \inf_{h \in \mathcal{H}} R(h) + \underbrace{2R_m(\mathcal{H}) + 2\sqrt{\frac{\log 2/\delta}{2m}}}_{\varepsilon} \right] \geq 1 - \delta.$$

$\Rightarrow \delta = 2e^{-2m(\frac{\varepsilon}{2} - R_m(\mathcal{H}))^2}$

$\Rightarrow R(h_S^{ERM}) - R^* \leq \underbrace{\inf_{h \in \mathcal{H}} R(h)}_{\text{approx. error}} - R^* + (\quad)$.

$\Rightarrow$ estimation $\leq 2R_m(\mathcal{H}) + 2\sqrt{\frac{\log 2/\delta}{2m}}$ with prob $\geq 1 - \delta$.

Equivalently, $\mathbb{P}_{S\sim D^m}\left[R(h_S^{ERM}) - \inf_{h\in H} R(h) > \varepsilon\right] \le 2e^{-2m\left(\frac{\varepsilon}{2} - R_m(H)\right)^2}$.

Proposition 4.1

$$\mathbb{P}\left[R(h_S^{ERM}) - \inf_{h\in H} R(h) > \varepsilon\right] \le \mathbb{P}\left[\sup_{h\in H} |R(h) - \hat{R}_S(h)| > \frac{\varepsilon}{2}\right]$$

$\exists h\in H$ s.t. $R(h) - \hat{R}_S(h) > \frac{\varepsilon}{2}$ or $R(h) - \hat{R}_S(h) < -\frac{\varepsilon}{2}$.

(Recall) (Theorem 3.5)

$$\mathbb{P}_{S\sim D^m}\left[\forall h\in H \quad R(h) \le \hat{R}_S(h) + R_m(H) + \sqrt{\frac{\log(1/\delta)}{2m}}\right] \ge 1-\delta$$

$$\mathbb{P}_{S\sim D^m}\left[\forall h\in H \quad R(h) \ge \hat{R}_S(h) - R_m(H) - \sqrt{\frac{\log(1/\delta)}{2m}}\right] \ge 1-\delta$$

(pf of thm)

$$\mathbb{P}\left[R(h_S^{ERM}) - \inf_{h\in H} R(h) > \varepsilon\right] \le \mathbb{P}\left[\exists h\in H \text{ s.t. } R(h) - \hat{R}_S(h) > \tfrac{\varepsilon}{2}\right] + \mathbb{P}\left[\exists h\in H \text{ s.t. } R(h) - \hat{R}_S(h) < -\tfrac{\varepsilon}{2}\right]$$

(By prop. 4.1)

$$= \mathbb{P}\left[\exists h\in H \text{ s.t. } R(h) > \hat{R}_S(h) + R_m(H) + \sqrt{\tfrac{\log(2/\delta)}{2m}}\right] + \mathbb{P}\left[\text{"} \quad R(h) < \hat{R}_S(h) - R_m(H) - \sqrt{\tfrac{\log(2/\delta)}{2m}}\right]$$

$$\le \frac{\delta}{2} + \frac{\delta}{2} = \delta \quad \text{(By thm 3.5)}.$$

Thus, $\mathbb{P}\left[R(h_S^{ERM}) - \inf_{h\in H} R(h) \le \varepsilon\right] \ge 1-\delta$.

(pf of prop 4.1)

$$R(h_S^{ERM}) - \inf_{h\in H} R(h) \le 2\sup_{h\in H} |R(h) - \hat{R}_S(h)|$$

Take $K>0$. $\exists h_K\in H$ s.t. $R(h_K) < \inf_{h\in H} R(h) + K$

$$R(h_S^{ERM}) - \inf_{h\in H} R(h) < R(h_S^{ERM}) - R(h_K) + K$$

$$= R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_S^{ERM}) - R(h_K) + K$$

$$\le R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_K) - R(h_K) + K$$

$$\le 2\sup_{h\in H} |R(h) - \hat{R}_S(h)| + K$$

As $K>0$ is arbitary, done.

(another pf)

$$R(h_S^{ERM}) - R(h) = R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_S^{ERM}) - R(h)$$
$$\le R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h) - R(h)$$
$$\le 2 \sup_{h \in H} |R(h) - \hat{R}_S(h)|$$

Take sup over $h \in H$ on both side

$$\sup_{h \in H} \left(R(h_S^{ERM}) - R(h) \right) = R(h_S^{ERM}) - \inf_{h \in H} R(h) \le 2 \sup_{h \in H} |R(h) - \hat{R}_S(h)|$$

§2. Structural Risk Minimization (SRM)

①

$H = \bigcup_{k=1}^{\infty} H_k$, $H_k \subseteq H_{k+1}$. Given $S \in (\mathcal{X} \times Y)^m$,

$$(k, h) \in \underset{k, h \in H_k}{\operatorname{argmin}} \left(\hat{R}_S(h) + R_m(H) + \sqrt{\frac{\log k}{m}} \right).$$

↳ Meaning of the term?

②Lemma (Uniform Generalization Bound).

$$\mathbb{P}_{S \sim D^m} \left[\forall k, \forall h \in H, \quad R(h) \le \hat{R}_S(h) + R_m(H) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}} \right] \ge 1 - \delta.$$

③Theorem 4.2 (Error Bound of SRM).

$$\mathbb{P}_{S \sim D^m} \left[R(h_S^{SRM}) \le \inf_{h \in H} \left(R(h) + 2R_m(H_{k(h)}) + 2\sqrt{\frac{\log k(h)}{m}} \right) + 2\sqrt{\frac{\log 4/\delta}{2m}} \right] \ge 1 - \delta,$$

where $k(h) = \min \{ k \mid h \in H_k \}$.

(pf of lemma)

$$\mathbb{P}_{S \sim D^m} \left[\exists k, \exists h \in H, \quad R(h) > \hat{R}_S(h) + R_m(H_k) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}} \right]$$
$$\le \sum_k \mathbb{P}_{S \sim D^m} \left[\exists h \in H, R(h) > \hat{R}_S(h) + R_m(H_k) + \varepsilon \right], \text{ where } \varepsilon = \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}$$
$$= \sum_k e^{-2m\varepsilon^2} = \sum_k e^{-2m\left(\sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}}\right)^2}$$
$$\le \sum_k e^{-2m\left(\frac{\log k}{m} + \frac{\log 2/\delta}{2m}\right)} = \sum_k \frac{1}{k^2}\left(\frac{\delta}{2}\right) = \frac{\pi^2}{6} \cdot \frac{\delta}{2} < \delta.$$

(pf of thm 4.2)

Let $\varepsilon = 2\sqrt{\frac{\log 4/\delta}{2m}}$.

$$\mathbb{P}\left[R(h_S^{SRM}) - \inf_{h \in H} \left(R(h) + 2R_m(H_{k(h)}) + 2\sqrt{\frac{\log k(h)}{m}} \right) > \varepsilon \right]$$

Suppose $\varphi_0 \Rightarrow \varphi_1 \Rightarrow \varphi_2 \Rightarrow \cdots$, $\varphi = \bigvee_{i=1}^{\infty} \varphi$. ↳ $F(h)$

Then, $\lim_{n\to\infty} \mathbb{P}[\varphi_n] = \mathbb{P}[\varphi]$.

$\exists h_1, h_2, h_3, \cdots \in \mathcal{H}$ s.t. $F(h_1) \geq F(h_2) \geq \cdots$ and $F(h_n) \to \inf_{h\in\mathcal{H}} F(h)$.
If $R(h_S^{SRM}) - F(h_n) > \varepsilon \ \forall n$, then $R(h_S^{SRM}) - \inf_{h\in\mathcal{H}} F(h) \geq \varepsilon$

Suffices to show $\forall h \in \mathcal{H}$, $\mathbb{P}[R(h_S^{SRM}) - F(h) > \varepsilon] \leq \delta$.

$$\mathbb{P}\Big[R(h_S^{SRM}) - \Big(R(h) + 2R_m(\mathcal{H}_{k(h)}) + 2\sqrt{\tfrac{\log k(h)}{m}}\Big) > \varepsilon\Big] \quad (*)$$

$$\leq \mathbb{P}\Big[R(h_S^{SRM}) - \Big(\hat{R}_S(h_S^{SRM}) + R_m(\mathcal{H}_{k_S^{SRM}}) + \sqrt{\tfrac{\log k_S^{SRM}}{m}}\Big) + \Big(\hat{R}_S(h) + R_m(\mathcal{H}_{k(h)}) + \sqrt{\tfrac{\log k(h)}{m}}\Big) - (*) > \varepsilon\Big]$$

$$= \mathbb{P}\Big[R(h_S^{SRM}) - \Big(\hat{R}_S(h_S^{SRM}) + R_m(\mathcal{H}_{k_S^{SRM}}) + \sqrt{\tfrac{\log k_S^{SRM}}{m}}\Big) - R(h) + \hat{R}_S(h) - R_m(\mathcal{H}_{k(h)}) - \sqrt{\tfrac{\log k(h)}{m}} > \varepsilon\Big]$$

$$\leq \mathbb{P}\Big[R(h_S^{SRM}) - \Big(\hat{R}_S(h_S^{SRM}) + R_m(\mathcal{H}_{k_S^{SRM}}) + \sqrt{\tfrac{\log k_S^{SRM}}{m}}\Big) > \tfrac{\varepsilon}{2}\Big] + \mathbb{P}\Big[-R(h) + \hat{R}_S(h) - R_m(\mathcal{H}_{k(h)}) - \sqrt{\tfrac{\log k(h)}{m}} > \tfrac{\varepsilon}{2}\Big]$$

$$\leq 2\cdot e^{-2m(\frac{\varepsilon}{2})^2} \times 2 \quad \text{(By lemma)}, \quad \text{using } \sqrt{\tfrac{\log 2/\delta}{2m}} = \tfrac{\varepsilon}{2}$$

$$= 4e^{-2m(\frac{\varepsilon}{2})^2} = \delta.$$

§ Cross Validation

① $\{\mathcal{H}_k\}_{k\in\mathbb{N}}$ with $\mathcal{H}_k \subseteq \mathcal{H}_{k+1}$.

SRM $\cdots \operatorname{argmin}_{k, h\in\mathcal{H}_k} \Big(\hat{R}_S(h) + R_m(\mathcal{H}_k) + \sqrt{\tfrac{\log k}{m}}\Big)$.

② (Cross Validation)

Given $S \in (\mathcal{X}\times\mathcal{Y})^m$, $S = S_1 \dot{\cup} S_2$, $|S_1| = (1-\alpha)m$, $|S_2| = \alpha m$

$k \in \operatorname{argmin}_k \hat{R}_{S_2}(\underbrace{h_{S_1,k}^{ERM}}_{\in \mathcal{H}_k})$.

Return $h_{S_1,k}^{ERM}$.

Theorem 4.4 $\mathbb{P}_{S\sim D^m}\left[R(h_S^{CV}) \le \inf_k \left(R(h_{S_1,k}^{ERM}) + 2\sqrt{\frac{\log \max(k, k(h_S^{CV}))}{\alpha m}}\right) + 2\sqrt{\frac{\log 4/\delta}{\alpha m}}\right]$

$\ge 1-\delta.$

Proposition 4.3 $\mathbb{P}_{S\sim D^m}\left[\sup_k \left|R(h_{S_1,k}^{ERM}) - \hat{R}(h_{S_1,k}^{ERM})\right| - \sqrt{\frac{\log k}{\alpha m}} > \varepsilon\right] \le 4e^{-2\alpha m \varepsilon^2}.$

(pf of thm) ① $\mathbb{P}\left[\bigcup_k \left|R(h_{S_1,k}^{ERM}) - \hat{R}(h_{S_1,k}^{ERM})\right| \ge \varepsilon + \sqrt{\frac{\log k}{\alpha m}}\right] \ge 1 - 4e^{-2\alpha m \varepsilon^2}.$

(By proposition 4.3). $\varepsilon = \sqrt{\frac{\log 4/\delta}{\alpha m}}$ $= 1-\delta.$

$\not\Leftarrow$ (

$\mathbb{P}\left[R(h_S^{CV}) \le R(h_{S_1,k}^{ERM}) + 2\sqrt{\frac{\log \max(k, k(h_S^{CV}))}{\alpha m}} + 2\sqrt{\frac{\log 4/\delta}{\alpha m}}\right]$

$\ge \mathbb{P}\left[R(h_S^{CV}) - \hat{R}_{S_1}(h_S^{CV}) + \hat{R}_{S_1}(h_{S_1}^{ERM}) - R(h_{S_1,k}^{ERM})\right.$

≤ 0

$\left. \le 2\sqrt{\frac{\log(\max(k, k(h_S^{CV}))}{\alpha m}} + 2\sqrt{\frac{\log 4/\delta}{\alpha m}}\right]$

$\ge \mathbb{P}\left[R(h_S^{CV}) - \hat{R}_{S_1}(h_S^{CV}) \ge \sqrt{\log}\right.$

thm 4.4 $\Leftrightarrow$ ② $\mathbb{P}\left[\bigcup_k R(h_S^{CV}) \le R(h_{S_1,k}^{ERM}) + 2\sqrt{\frac{\log \max(k, k(h_S^{CV}))}{\alpha m}} + 2\sqrt{\frac{\log 4/\delta}{\alpha m}}\right] \ge 1-\delta.$

if ① $\Rightarrow$ ②, done.

$R(h_S^{CV}) - R(h_{S_1,k}^{ERM}) \le R(h_S^{CV}) - \hat{R}_{S_2}(h_S^{CV}) + \hat{R}_{S_2}(h_{S_1,k}^{ERM}) - R(h_{S_1,k}^{ERM})$

$\le 2\sqrt{\frac{\log \max(k, k(h_S^{CV}))}{\alpha m}} + 2\sqrt{\frac{\log 4/\delta}{\alpha m}}$ (By ②)

(pf of prop. 4.3) $\Leftrightarrow \mathbb{P}_{S\sim D^m}\left[\exists k \;\; \left|R(h_{S_1,k}^{ERM}) - \hat{R}_{S_2}(h_{S_1,k}^{ERM})\right| > \sqrt{\frac{\log k}{\alpha m}} + \varepsilon\right]$

$\le \sum_k \mathbb{P}_{S\sim D^m}\left[\left|R(h_{S_1,k}^{ERM}) - \hat{R}_{S_2}(h_{S_1,k}^{ERM})\right| > \sqrt{\frac{\log k}{\alpha m}} + \varepsilon\right]$

$= \sum_k \mathbb{E}_{S_1}\left[\mathbb{P}_{S_2}\left|R(h_{S_1,k}^{ERM}) - \hat{R}_{S_2}(h_{S_1,k}^{ERM})\right| > \sqrt{\frac{\log k}{\alpha m}} + \varepsilon \;\middle|\; S_1\right]$

$$= \sum_k \mathop{\mathbb{E}}_{S_1} 2e^{-2\alpha m\left(\sqrt{\frac{\log k}{\alpha m}}+\varepsilon\right)^2}$$

$$\le \sum_k \mathop{\mathbb{E}}_{S_1} \frac{2}{k^2} e^{-2\alpha m\varepsilon^2} = \frac{\pi^2}{3} e^{-2\alpha m\varepsilon^2} \le 4e^{-2\alpha m\varepsilon^2}.$$

§ n-fold CV.

① Given $S \in (\mathcal{X}\times\mathcal{Y})^m$, $\{\mathcal{H}_k\}_{k\in\mathbb{N}}$.

$S = S_1 \dot\cup S_2 \dot\cup \cdots \dot\cup S_n$, $|S_i| = m_i$.

$$\hat R_{cv}(\mathcal{H}_k) = \frac{1}{n}\sum_{i=1}^n \hat R_{S_i}(h^{ERM}_{S\setminus S_i,k})$$

$k \in \operatorname{argmin}_k \hat R_{cv}(\mathcal{H}_k)$. Return $h^{ERM}_{S,k}$.

② Consider case $m_1 = m_2 = \cdots = m_n = a$.

$$\mathop{\mathbb{E}}_S[\hat R_{cv}(\mathcal{H}_k)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\hat R_{S_i}(h^{ERM}_{S\setminus S_i,k})]$$

$$= \frac{1}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{S\setminus S_i} \mathop{\mathbb{E}}_{S_i}[\hat R_{S_i}(h^{ERM}_{S\setminus S_i,k}) \mid S\setminus S_i]$$

$$= \frac{1}{n}\sum_{i=1}^n \mathop{\mathbb{E}}_{S\setminus S_i} R(h^{ERM}_{S\setminus S_i,k})$$

$$= \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^{m-a}} R(h^{ERM}_{S',k}).$$