3 ☆ Rademacher Complexity & ☆ VC-Dimension 复杂性

finite $|\mathcal{H}|$ $\begin{cases} \text{consistent: } R(h) \le \dfrac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{m} & \text{w.p.} \ge 1-\delta \\ \text{inconsistent: } R(h) \le \hat{R}_S(h) + \sqrt{\dfrac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}} & \text{w.p.} \ge 1-\delta \end{cases}$

infinite $|\mathcal{H}|$ $\rightarrow$ "reduction" 约简

Rademacher Complexity

$\mathcal{H}$: hypothesis set $\rightarrow h \in \mathcal{H} \begin{cases} \text{经验误差 } \hat{R}_S(h) \\ \text{泛化 } R(h) \end{cases}$ $\quad \mathbb{1}_{h(x) \neq y}$: 0-1 loss

$\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{[0,1]}$ $\quad G = \{ g : (x,y) \mapsto \mathcal{L}(h(x), y) ; h \in \mathcal{H} \}$

the family of loss functions associated with $\mathcal{H}$.

def (Empirical Rademacher Complexity) $S = (z_1, z_2, \cdots, z_m) \sim \mathcal{D}^m$, $z_i = (x_i, y_i)$

$$\hat{\mathcal{R}}_S(G) = \mathbb{E}_\sigma \left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right] \quad ☆ = \mathbb{E}_\sigma \left[\sup_{g \in G} \frac{\vec{\sigma} \cdot \vec{g}_S}{m} \right]$$

$\vec{\sigma} = (\sigma_1, \cdots, \sigma_m)^T$ $\quad \sigma_i \sim \{-1, +1\} \begin{cases} \sigma_i = -1, \text{ w.p. } \frac{1}{2} \\ \sigma_i = +1, \text{ w.p. } \frac{1}{2} \end{cases}$ $\quad \vec{g}_S = (g(z_1), \cdots, g(z_m))^T$

"随机噪声"

内积是相似度衡量的指标

Rademacher 复杂度反映了函数集 G 对随机噪声 σ 的拟合能力
(后随机噪声下的表达能力)

def (Rademacher Complexity) $\mathcal{R}_m(G) = \mathbb{E}_{S \sim \mathcal{D}^m} [\hat{\mathcal{R}}_S(G)]$ ☆

Thm1 w.p. $\geq 1-\delta$, $\forall g \in G$ 损失函数

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathcal{R}}_S(G) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

Proof: $\mathbb{E}[g(z)]$

$$\frac{1}{m}\sum_{i=1}^{m} g(z_i) = \hat{\mathbb{E}}_S[g]$$

consider $\boxed{\Phi(S) = \sup_{g\in G}\left(\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]\right)}$

$$\geq \mathbb{E}[g] - \hat{\mathbb{E}}_S[g], \quad \forall g \in G$$

McDiarmid Inequality:

$X_1, \cdots, X_m$ i.i.d.r.v. $f: \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m \to \mathbb{R}$.

if $\forall i \in [m]$, $\sup_{\substack{x_1\in\mathcal{X}_1,\cdots,x_m\in\mathcal{X}_m \\ x_i' \in \mathcal{X}_i}} \left| f(x_1,\cdots,x_m) - f(x_1,\cdots,x_i',\cdots,x_m)\right| \leq c_i$

"Bounded Difference" (Lipschitz)

then $\forall t>0$. $\Pr\left(\left|f(X_1,\cdots,X_m) - \mathbb{E}[f(X_1,\cdots,X_m)]\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{m} c_i^2}\right)$

two-sided

考虑Φ在S变化为S'时的差,

$$\Phi(S') - \Phi(S) \leq \sup_{g\in G}\left(\hat{\mathbb{E}}_S[g] - \hat{\mathbb{E}}_{S'}[g]\right)$$

$$= \sup_{g\in G} \frac{1}{m}\left(g(z_i) - g(z_i')\right) = \frac{1}{m}$$

$\forall x, y$ $|f(x) - f(y)| \leq L|x-y|$

$\sup f' - \sup f \leq \sup(f' - f) = \sup f' - \inf f$

$\Rightarrow$ by McDiarmid, $\Phi(S) \leq \mathbb{E}_S[\Phi(S)] + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$ w.p. $\geq 1-\delta$

one-sided

只需证: $\mathbb{E}_S[\Phi(S)] \leq 2\mathcal{R}_m(G)$

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_{g\in G}\left(\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]\right)\right]$$

$\mathbb{E}[g] = \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g]\right]$

$$= \mathbb{E}_{S}\left[\sup_{g\in G} \mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_{S}[g]\right]\right].$$

$\sup f + \sup g \geq \sup(f+g)$

(sub-additivity)

$$\leq \mathbb{E}_{S,S'}\left[\sup_{g\in G}\left(\hat{\mathbb{E}}_{S'}[g] - \hat{\mathbb{E}}_{S}[g]\right)\right].$$

$$= \mathbb{E}_{S,S'}\left[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}\left(g(z_i') - g(z_i)\right)\right]$$

$z_i, z_i' \sim \mathcal{D}.$

$$= \mathbb{E}_{\sigma,S,S'}\left[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(g(z_i') - g(z_i)\right)\right],$$

$\sigma_i \sim \{-1, 1\}$

$$\leq \mathbb{E}_{\sigma,S}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] + \mathbb{E}_{\sigma,S'}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i')\right]$$

$$= 2\,\mathbb{E}_{\sigma,S}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)\right] = 2\,\mathbb{E}_{S}\left[\hat{\mathcal{R}}_S(G)\right] = 2\,\mathcal{R}_m(G)$$

$S, S' \sim \mathcal{D}^m$

$$\Rightarrow \mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \quad ①$$

w.p. $\geq 1-\delta/2$

$\mathbb{E}_S[\hat{\mathcal{R}}_S(G)] = \mathcal{R}_m(G)$ $\Rightarrow$ McDiarmid 不等式!

w.p. $\geq 1-\delta/2$ $\quad \hat{\mathcal{R}}_S(G) \leq \mathcal{R}_m(G) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}} \quad ②$

$$\Rightarrow \mathbb{E}[g(z)] \leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(G) + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m} g(z_i) + 2\hat{\mathcal{R}}_S(G) + 3\cdot\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

w.p. $\geq 1-\delta$ $\quad \square.$

Lemma $G = \{(x,y) \mapsto \mathbb{1}_{h(x)\neq y} : h \in H\}$

$$\boxed{\hat{\mathcal{R}}_S(G) = \frac{1}{2}\hat{\mathcal{R}}_S(H)} \quad ☆$$

Proof: sample set $S = \{(x_1,y_1), \cdots, (x_m,y_m)\}$ $\quad \mathcal{X}\times\{-1,+1\}.$

$h(x_i), y_i \in \{-1,+1\}$

$$\hat{\mathcal{R}}_S(G) = \mathbb{E}_{\sigma}\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i \mathbb{1}_{h(x_i)\neq y_i}\right].$$

$h(x_i) = y_i \Rightarrow 0 \qquad h(x_i)\cdot y_i = 1$

$h(x_i) \neq y_i \Rightarrow 1 \qquad h(x_i)\cdot y_i = -1.$

$$= \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \cdot \frac{1-y_i h(x_i)}{2}\right]$$

$\mathbb{E}[\sigma_i]=0$

$$= \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} (-\sigma_i y_i) h(x_i)\right]$$

$\sigma_i \sim -\sigma_i y_i$

$$= \frac{1}{2}\mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] = \frac{1}{2}\hat{\mathfrak{R}}_S(\mathcal{H}) \qquad \square.$$

Thm2

注比赛 ☆☆

$\mathcal{H}$ 为假设空间 $h\in\mathcal{H}$ 满足 $h(x)\in\{-1,+1\}$ $\forall \delta>0$.

w.p. $\geq 1-\delta$

$$R(h) \leq \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$R(h) \leq \hat{R}_S(h) + \hat{\mathfrak{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

注比误差 经验误差 Rademacher 复杂度

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)\right] \qquad \sigma \sim -\sigma$$

fix σ, $\Rightarrow$ $\inf_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} -\sigma_i h(x_i)$ $\Leftrightarrow$ $\boxed{\text{empirical risk minimization}}$ (ERM)

minimize

NP-hard 计算 $\hat{\mathfrak{R}}_S(\mathcal{H})$ 可能过于复杂 (in Chapter 4 模型选择)

$\Rightarrow$ Growth Function. VC-Dimension

---

def (Growth Function) $\Pi_{\mathcal{H}}: \mathbb{N}\to\mathbb{N}$

增长函数

$$\forall m\in\mathbb{N}.\ \Pi_{\mathcal{H}}(m) = \max_{\{x_1,\cdots,x_m\}\subseteq\mathcal{X}} \left|\{(h(x_1),\cdots,h(x_m)) : h\in\mathcal{H}\}\right| \leq 2^m$$

☆

= 分类: $h(x_i)\in\{-1,+1\}$ $\Rightarrow$ 对分 (dichotomy)

$\Pi_{\mathcal{H}}(m)$ is the maximum number of dichotomy that are realized by the hypothesis

① $\mathcal{H}$: richer $\Rightarrow \Pi_{\mathcal{H}}(m)$

"m个样本最多可以产生多少种不同m对分"

② not depend on distribution (purely combinatorial)

Lemma 1 (Maximal Inequality) $X_1, \cdots, X_n$ r.v. $\forall j \in [n]$ $t>0$ $\mathbb{E}[e^{tX_j}] \le e^{\frac{t^2 r^2}{2}}$. $\exists r>0$

(MGF)

$$\mathbb{E}[\max_j X_j] \le r\sqrt{2\log n}$$

Proof: Jenson's inequality (convex function f: $\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$)

$$e^{t\,\mathbb{E}[\max X_j]} \le \mathbb{E}[e^{t\cdot\max X_j}] = \mathbb{E}[\max_j e^{tX_j}] \le \mathbb{E}[\sum_j e^{tX_j}] \le n\cdot e^{\frac{t^2r^2}{2}}$$

$\Rightarrow$ take logarithm: $\mathbb{E}[\max_j X_j] \le \frac{\log n}{t} + \frac{tr^2}{2}$. $\forall t>0$.

Let $t = \frac{\sqrt{2\log n}}{r}$ $\Rightarrow \mathbb{E}[\max_j X_j] \le r\sqrt{2\log n}$ $\square$

Lemma 1' (Maximal Inequality) $X_1, \cdots, X_n$. r.v. $\forall j \in [n]$. $X_j = \sum_{i=1}^m Y_{ij}$

fix j. Y_{ij} independent. zero-mean. taking values $[-r_i, +r_i]$.

$\Rightarrow \mathbb{E}[\max_j X_j] \le r\sqrt{2\log n}$. where $r = \sqrt{\sum_{i=1}^m r_i^2}$

Proof: $\mathbb{E}[e^{tX_j}] = \mathbb{E}[\prod_{i=1}^m e^{tY_{ij}}] \underset{\text{independent}}{=} \prod_{i=1}^m \mathbb{E}[e^{tY_{ij}}] \underset{\text{Hoeffding's Lemma}}{\le} \prod_{i=1}^m e^{\frac{t^2 r_i^2}{2}} = e^{\frac{t^2r^2}{2}}$ $\square$

(r.v. Y 满足 $\mathbb{E}[Y]=0$, $Y \in [a,b]$, $\mathbb{E}[e^{tY}] \le e^{\frac{t^2}{8}(b-a)^2}$)

Thm (Massart's Lemma). $A \in \mathbb{R}^m$ finite set.

$r = \max_{x \in A} \|x\|_2$.

$$\Rightarrow \mathbb{E}_\sigma\left[\frac{1}{m}\sup_{x\in A}\sum_{i=1}^m \sigma_i x_i\right] \le \frac{r\cdot\sqrt{2\log|A|}}{m}.$$

Proof: By Maximal Inequality, $\underbrace{\sigma_i x_i}_{Y_{ij}} \in [-|x_i|, +|x_i|]$. $\sqrt{\sum_{i=1}^m x_i^2} \le r$. $\square$

$\Downarrow$

Rademacher 复杂度 & 增长函数

Cor G 为函数集 $g \in G$ $g(x) \in \{-1, +1\}$

则 $\boxed{R_m(G) \le \sqrt{\frac{2\log \Pi_G(m)}{m}}}$ ☆

Proof: fixed sample set $S=(x_1,\cdots,x_m)$

$$G|_S \leftarrow \boxed{(g(x_1),\cdots,g(x_m))},\ g\in G$$

$$R_m(G)=\mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{u\in G|_S}\frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right]$$

$$\le \mathbb{E}_S\left[\frac{\sqrt{m}\cdot\sqrt{2\log|G|_S|}}{m}\right]=\mathbb{E}_S\left[\sqrt{\frac{2\log|G|_S|}{m}}\right]\le \mathbb{E}_S\left[\sqrt{\frac{2\log\Pi_G(m)}{m}}\right] \quad \square$$

$\Delta\ r=\sqrt{m}$

$|G|_S|\le \Pi_G(m)$ by definition of Growth Function

统计量空间对随机噪声的拟合能力

统计量"对分"程度

σ_i: Rademacher Variables,

Thm3

泛化界(by Growth Function)

$$\boxed{R(h)\le \hat{R}_S(h)+\sqrt{\frac{2\log\Pi_{\mathcal{H}}(m)}{m}}+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}} \quad \text{w.p.} \ge 1-\delta$$

$\Pi_{\mathcal{H}}(m)$ $m\ge 1$

*BTW: directly- $\Pr\left(|R(h)-\hat{R}_S(h)|>\varepsilon\right)\le 4\Pi_{\mathcal{H}}(2m)\cdot\exp\left(-\frac{m\varepsilon^2}{8}\right)=\delta$

VC维 (VC-dimension)

Vapnik Chervonenkis

$$\Pi_{\mathcal{H}}(m)=\max_{\{x_1,\cdots,x_m\}\subseteq X}\left|\{(h(x_1),\cdots,h(x_m)):h\in\mathcal{H}\}\right|\le 2^m$$

产生对分

若对某样本集 $S=(x_1,\cdots,x_m)$ 而言 每一种对分情况都存在 $h\in\mathcal{H}$ 使之实现(realize). 则称S能被$\mathcal{H}$"打散"

Shattering

$\mathcal{H}$ realize all possible dichotomies. $\Pi_{\mathcal{H}}(m)=2^m$

3个样本:

$$\begin{cases} +++ & h_1\\ ++- & h_2\\ +-+ & \vdots\\ +-- & \\ -++ & \\ -+- & \\ --+ & \\ --- & h_8\end{cases}$$

def (VC-Dimension)

$$\boxed{VC\dim(\mathcal{H})=\max\{m:\Pi_{\mathcal{H}}(m)=2^m\}}$$

"the size of the largest set that can be shattered by $\mathcal{H}$".

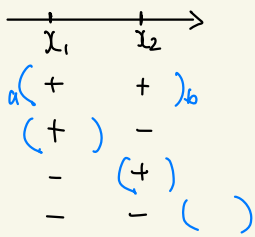
$VC\dim(\mathcal{H})=d \Rightarrow$

① $\exists\ (x_1,\cdots,x_d)$ 能被$\mathcal{H}$打散

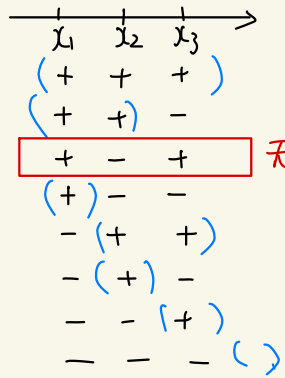
② $\forall\ (x_1,\cdots,x_d,x_{d+1})$ 不能被$\mathcal{H}$打散

examples: ① Intervals on the real line

$\mathcal{H}$ 为数轴上所有区间 $\begin{cases} \text{内}: + \\ \text{外}: - \end{cases}$



$\exists$ 2个点能被打散



无法满足

$\forall$ 3个点不能被打散.

$\Downarrow$

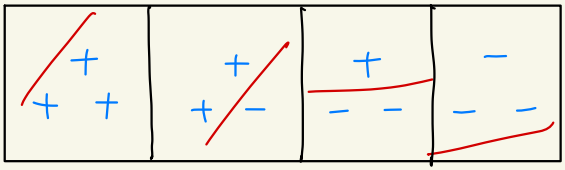
$\mathrm{VCdim}(\text{intervals in } \mathbb{R}) = 2$

② Hyperplanes

$\mathcal{H}$ 为所有而超平面 $w \cdot x = 0$ $\begin{cases} >0: + \\ <0: - \end{cases}$

先考虑 $\mathbb{R}^2$

$\exists$ 3个点:

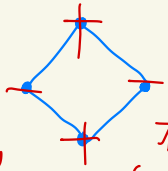


$\checkmark$

$\forall$ 4个点: $\mathrm{Convex\text{-}Hull}(x_1, \cdots, x_4) = \left\{ x = \sum_{i=1}^{4} \alpha_i x_i \;\middle|\; \alpha_i \in [0,1], \sum_{i=1}^{4} \alpha_i = 1 \right\}$

不能被打散

Case 1. 四个点都在凸包而边上.

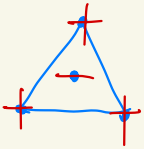


dichotomy

不可分!

(XOR问题)

Case 2: 三个点在凸包而边上



不可分!

$\Rightarrow \mathrm{VCdim}(\text{hyperplanes in } \mathbb{R}^2) = 3$

再考虑 $\mathbb{R}^d$ $\mathrm{VCdim}(\text{hyperplanes in } \mathbb{R}^d) = d+1$

$\exists$ $d+1$ points can be shattered.

labels: $y_0, y_1, \cdots, y_d \in \{-1, +1\}$

$d+1$ 点 $\begin{cases} \text{零向量 } x_0 = (0, \cdots, 0, \cdots, 0) \\ \text{单位向量 } x_i = (0, \cdots, \underset{\uparrow}{1}, \cdots, 0) \end{cases}$

i^{th} entry

$\checkmark$

$\vec{w} = (y_1, \cdots, y_d)$ $b = \frac{y_0}{2}$ $\vec{w} \cdot x + b = 0$

$i = 0$ $\mathrm{sgn}(w x_0 + b) = \mathrm{sgn}\left(\frac{y_0}{2}\right) = y_0$

$i \in [1 \cdots d]$ $\mathrm{sgn}(w x_i + b) = \mathrm{sgn}\left(\underset{\pm 1}{y_i} + \underset{\pm\frac{1}{2}}{\frac{y_0}{2}}\right) = y_i$

$\forall$ $d+2$ points cannot be shattered.

Thm (Radon's Theorem) Any set $\mathcal{X}$ of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ s.t. the convex hulls of $\mathcal{X}_1$ and $\mathcal{X}_2$ intersect.

Proof: $\mathcal{X} = \{\vec{x}_1, \cdots, \vec{x}_{d+2}\} \subseteq \mathbb{R}^d$.

Consider $\begin{cases} \sum_{i=1}^{d+2} \alpha_i \vec{x}_i = 0 & (\to d \text{ 个方程}) \\ \sum_{i=1}^{d+2} \alpha_i = 0 & (\to 1 \text{ 个方程}) \end{cases}$ } linear system. $\begin{cases} d+1 \text{ equations} \\ d+2 \text{ unknown variables} \end{cases}$

$\Downarrow$

underdeterminate.

$\alpha = (\alpha_1, \cdots, \alpha_{d+2}) \in \mathbb{R}^{d+2}$.

$\exists$ non-zero solution $\beta_1, \cdots, \beta_{d+2}$ s.t. $\begin{cases} \sum_{i=1}^{d+2} \beta_i \vec{x}_i = 0 \\ \sum_{i=1}^{d+2} \beta_i = 0 \end{cases}$ $\to$ $\begin{cases} I_1 = \{i \in [d+2] : \beta_i > 0\} \\ I_2 = \{i \in [d+2] : \beta_i \le 0\} \end{cases}$

$\Downarrow$

form a partition of $\mathcal{X}$.

$\sum_{i \in I_1} \beta_i = -\sum_{i \in I_2} \beta_i \triangleq \beta$

$$\sum_{i \in I_1} \frac{\beta_i}{\beta} \vec{x}_i = \sum_{i \in I_2} \frac{-\beta_i}{\beta} \vec{x}_i$$

Observe that $\sum_{i \in I_1} \frac{\beta_i}{\beta} = 1$, $\sum_{i \in I_2} \frac{-\beta_i}{\beta} = 1$, $\Rightarrow$ $\sum_{i \in I_1} \frac{\beta_i}{\beta} x_i \in \text{Convex-Hull}(x_i, i \in I_1)$
$\in \text{Convex-Hull}(x_i, i \in I_2)$

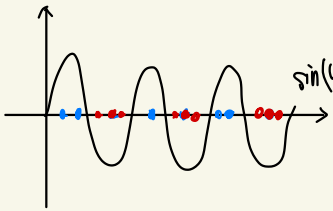
End of Proof $\square$.

$\mathcal{H}$ $\mathbb{R}^d$ $d+1$. $\vec{w}\vec{x} + b = 0$ 需要 $d+1$ 个未知量 (自由度是 $d+1$)

Not generally holds!

Consider Sine Functions. $\mathcal{H} = \{h : h(t) = \sin(wt), w \in \mathbb{R}\}$

$\text{VCdim}(\mathcal{H}) = +\infty$



$\begin{cases} y_i > \sin(wx_i), & + \\ y_i \le \sin(wx_i), & - \end{cases}$

$\forall m$ 个点的任意对分, 都可以实现

VC维泛化界

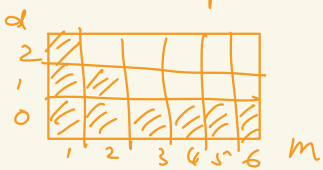
Thm (Sauer's Lemma) hypothesis set $\mathcal{H}$. $VCdim(\mathcal{H}) = d$.

$$\forall m \in \mathbb{N}, \quad \Pi_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$$

finite — polynomial

$$\begin{cases} VCdim(\mathcal{H}) = d < +\infty \Rightarrow \Pi_{\mathcal{H}}(m) = O(m^d) \\ VCdim(\mathcal{H}) = +\infty \Rightarrow \Pi_{\mathcal{H}}(m) = 2^m \end{cases}$$

infinite — exponential



Proof: by Induction.

Base case: $m=1,\ d=0 \Rightarrow \Pi_{\mathcal{H}}(m) = 1,\ \sum_{i=0}^{0}\binom{m}{i} = 1,$

$m=1,\ d=1 \Rightarrow \Pi_{\mathcal{H}}(m) = 2,\ \sum_{i=0}^{1}\binom{m}{i} = 1+m = 2$

Assume that ① $(m-1, d-1)$ & ② $(m-1, d)$ holds. ($\Leftrightarrow$ prove (m, d) holds)

Consider $S = \{x_1, \cdots, x_m\}$ $\quad S' = \{x_1, \cdots, x_{m-1}\}$

$G_{|S} = \{(h(x_1), \cdots, h(x_m)),\ h \in \mathcal{H}\} \subseteq \mathbb{R}^m$

$G_{|S'} = \{(h(x_1), \cdots, h(x_{m-1})),\ h \in \mathcal{H}\} \subseteq \mathbb{R}^{m-1}$

$\boxed{(h(x_1), \cdots, h(x_{m-1}),\ h(x_m))}$

± 1 & -1

$h \in \mathcal{H}$ 对 x_m 预测结果 $+1/-1$.

$G_{|S'}$ 中的元素(串)会在 $G_{|S}$ 中出现一次或两次

令 $G_{S'|S}$ 表示在 $G_{|S}$ 中出现两次的 $G_{|S'}$ 中的串组成的集合.

$$G_{S'|S} = \{(y_1, \cdots, y_{m-1}) \in G_{|S'} \mid \exists h, h' \in \mathcal{H},\ h(x_i) = h'(x_i) = y_i,\ i \in [m-1],\ h(x_m) \ne h'(x_m)\}$$

$G_{|S'}$ < 一次 $|G_{|S'}| - |G_{S'|S}|$; 两次 $|G_{S'|S}|$

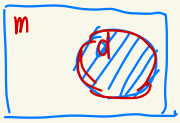
$$\Rightarrow |G_{|S}| = (|G_{|S'}| - |G_{S'|S}|) + 2 \cdot |G_{S'|S}|$$

$$\boxed{|G_{|S}| = \underbrace{|G_{|S'}|}_{②} + \underbrace{|G_{S'|S}|}_{①}}$$

for $G_{|S'}$: $|S'| = m-1$,

$$|G_{|S'}| \le \Pi_{\mathcal{H}}(m-1) \overset{②}{\le} \sum_{i=0}^{d} \binom{m-1}{i}$$

不知 x_m 是否能被打散



for $G_{S'|S}$: 设 $G_{S'|S}$ 的子集 Q 能被 H 打散

则 $Q \cup \{x_m\}$ 一定能被 H 打散.

即 $|Q \cup \{x_m\}| \le d$ (by definition of VC-dimension)

$\Rightarrow |Q| \le d-1. \Rightarrow \mathrm{VCdim}(G_{S'|S}) \le d-1.$

$$\Rightarrow |G_{S'|S}| \le \Pi_H(m-1) \overset{①}{\le} \sum_{i=0}^{d-1} \binom{m-1}{i}$$

$$|G_{|S}| = |G_{|S'}| + |G_{S'|S}| \le \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \qquad \binom{m-1}{-1} = 0$$

$$= \sum_{i=0}^{d} \left(\binom{m-1}{i} + \binom{m-1}{i-1} \right) = \sum_{i=0}^{d} \binom{m}{i}.$$

by arbitrary $S = \{x_1, \cdots, x_m\}$, $\Pi_H(m) = \max_S |G_{|S}| \le \sum_{i=0}^{d} \binom{m}{i}$. □

Cor $\mathrm{VCdim}(H) = d$. $\forall m \ge d$. $\Pi_H(m) \le \left(\frac{em}{d}\right)^d = O(m^d)$

growth function — VC-dimension

Proof: by Sauer's Lemma $\Pi_H(m) \le \sum_{i=0}^{d} \binom{m}{i}$

$$\le \sum_{i=0}^{d} \binom{m}{i} \cdot \left(\frac{m}{d}\right)^{d-i} \quad (i \le d, m \ge d)$$

$$\le \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \quad (m \ge d)$$

$$= \left(\frac{m}{d}\right)^d \cdot \sum_{i=0}^{m} \binom{m}{i} \cdot \left(\frac{d}{m}\right)^i \cdot 1^{m-i}$$

$$= \left(\frac{m}{d}\right)^d \cdot \left(1 + \frac{d}{m}\right)^m \le \left(\frac{m}{d}\right)^d \cdot e^d \quad □$$

$(1+x \le e^x)$

Bound by Growth Function:

$$\boxed{R(h) \le \hat{R}_S(h) + \sqrt{\frac{2 \log \Pi_H(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}}$$

Thm 4

$$R(h) \le \hat{R}_S(h) + \sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \quad \text{w.p.} \ge 1-\delta$$

$$O\left(\sqrt{\frac{\log m/d}{m/d}}\right)$$

$\frac{m}{d}\downarrow$ $\downarrow$

注比界 (by VC-dimension)

upper bound.

VC-dimension < 大：H rich ; 小：不容易overfit！ } trade-off.

"适中的VC dimension"

简单为王

Occam's Razor

BTW: $$R(h) \le \hat{R}_S(h) + \sqrt{\frac{8d\log\frac{2em}{d} + 8\log\frac{4}{\delta}}{m}} \quad \text{w.p.} \ge 1-\delta$$

复杂度 {
- Rademacher Complexity: $\hat{\mathcal{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_h \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)\right]$. $R_m(\mathcal{H})$ 抛硬币 对称 打散
- Growth Function: $\Pi_{\mathcal{H}}(m) = \max_S |\{(h(x_1),\cdots,h(x_m)), h\in\mathcal{H}\}|$
- VC-Dimension: $VCdim(\mathcal{H}) = \max\{m: \Pi_{\mathcal{H}}(m) = 2^m\}$

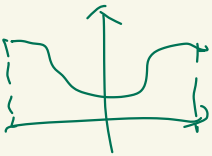
distribution-free

data-independent

Lower Bound

$R(h)\to 0$? 肯定不行！分布有关！ "bad distribution"

adverserial (enemy)



(FML Chapter 3.4)

Thm (lower bound) hypothesis set $\mathcal{H}$

① realizable case: $\forall$ algorithm $\mathcal{A}$. $\exists$ distribution $\mathcal{D}$. target function f

$$\Pr_{S\sim\mathcal{D}^m}\left(R_{\mathcal{D}}(h_S, f) > \frac{d-1}{32m}\right) \ge \frac{1}{100}.$$ "constant probability"

$\Rightarrow$ PAC-learning in the realizable case is impossible when the VC-dimension is infinite.

② non-realizable case: ....

$$\Pr_{S\sim\mathcal{D}^m}\left(R_{\mathcal{D}}(h_S) - \inf_{h\in\mathcal{H}} R_{\mathcal{D}}(h) > \sqrt{\frac{d}{320m}}\right) \ge \frac{1}{64}$$

Cor $\boxed{\text{With infinite VC-dimension, agonstic PAC-learning is impossible!}}$ $\Leftrightarrow$ sample complexity $m \ge \frac{d}{320\varepsilon^2}$.

4. Model Selection 模型选择

$\mathcal{H}$: $\begin{cases} \text{VC dimension } d \Rightarrow \times \text{ PAC learnable} \\ \text{VC dimension } +\infty \Rightarrow \text{ PAC learnable} \end{cases}$

How to choose the hypothesis set $\mathcal{H}$? → Rademacher 复杂度, Growth Function, VC dimension

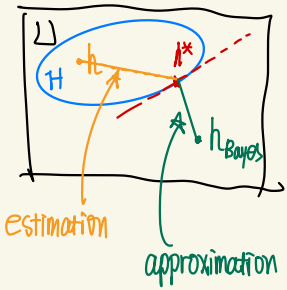
How to choose the hypothesis $h \in \mathcal{H}$?

Binary Classification: let $\mathcal{H}$ be a family of functions mapping $\mathcal{X} \to \{-1, +1\}$

$\forall h \in \mathcal{H}$ excess error = estimation error + approximation error ☆

$$R(h) - R^* = \Big(R(h) - \inf_{h \in \mathcal{H}} R(h)\Big) + \Big(\inf_{h \in \mathcal{H}} R(h) - R^*\Big)$$

h ; $\mathcal{H}$ (richness)



Agnostic: $\Pr\big(R(h_S) - \inf_h R(h) \le \varepsilon\big) \ge 1 - \delta$

def (Bayes error) $h: \mathcal{X} \to \mathcal{Y}$ h measurable

Bayes error $R^* = \inf_h R(h)$ → h: Bayes hypothesis (Bayes classifier) 贝叶斯最优分类器

$\mathbb{E}_S[\hat{R}_S(h)] = \mathbb{E}_S\Big[\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x_i) \ne y_i}\Big]$

deterministic & stochastic

$R^* = 0$ $R^* \ne 0$

Intuition: Binary Classification $\forall x \in \mathcal{X}$

☆ $$\boxed{h_{Bayes}(x) = \operatorname*{argmax}_{y \in \{\pm 1\}} \Pr(y|x)}$$

average error of h_{Bayes}: $\min\{\Pr(-1|x), \Pr(1|x)\}$

⇒ minimum possible error

$\begin{cases} \Pr(1|x) > \Pr(-1|x) \Rightarrow h_{Bayes}(x) = 1. \\ \Pr(1|x) \le \Pr(-1|x) \Rightarrow h_{Bayes}(x) = -1 \end{cases}$

def (Noise) $x \in \mathcal{X}$.

噪声

$\text{noise}(x) = \min\{\Pr(-1|x), \Pr(1|x)\}$

$\text{noise} = \mathbb{E}_{x \sim D}[\text{noise}(x)]$

learn: $\Pr(y|x)$ 学习是概率分布

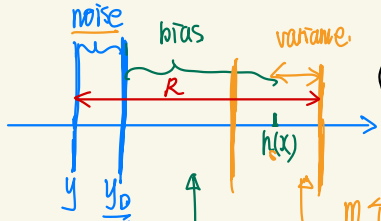
Bayesian Decision Theory 贝叶斯决策论

e.g. Naive Bayes (频率 ⇒ 概率)

$\text{noise} = R^*$

Bayes error

$* \begin{cases} \text{noise} \\ \text{bias} \\ \text{variance} \end{cases}$



(Chapter 2.5 [illegible])

分布本身特征

算法期望输出和真实label差值

m个样本下

$S \sim \mathcal{D}^m$

不同训练集而输出差

(underfitting) learner (overfit)

trade-off ☆

(bias-variance dilemma)

bias-variance decomposition:

for Regression task $R(h) = \text{bias}^2(x) + \text{Var}(x) + \text{noise}^2$

ERM (Empirical Risk Minimization)

经验风险最小化

☆ $\boxed{h_S^{ERM} = \underset{h \in H}{\arg\min}\ \hat{R}_S(h)}$ (by Optimization Methods)

Proposition: $\forall$ sample set S, $\Pr\left(R(h_S^{ERM}) - \inf_{h \in H} R(h) > \varepsilon\right) \le \Pr\left(\sup_{h \in H} |R(h) - \hat{R}_S(h)| > \frac{\varepsilon}{2}\right)$

estimation error

$= e^{-2m\left(\frac{\varepsilon}{2} - \mathcal{R}_m(\mathcal{H})\right)^2}$

if large sample set, favorable $\mathcal{R}_m(\mathcal{H})$

Generalization Bound:

$$R(h) \le \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} = \frac{\varepsilon}{2}$$

Proof: by definition of inferium:

$\forall \varepsilon > 0.\ \exists h_\varepsilon$ s.t.

$R(h_\varepsilon) \le \inf_{h \in H} R(h) + \varepsilon$

$\Rightarrow R(h_S^{ERM}) - \inf_{h \in H} R(h)$

$= R(h_\varepsilon^{ERM}) - R(h_\varepsilon) + R(h_\varepsilon) - \inf_{h \in H} R(h) \le R(h_\varepsilon^{ERM}) - R(h_\varepsilon) + \varepsilon.$

$= R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_S^{ERM}) - R(h_\varepsilon) + \varepsilon.$

$\le \left(R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM})\right) + \left(\hat{R}_S(h_\varepsilon) - R(h_\varepsilon)\right) + \varepsilon \le 2 \sup_{h \in H} |R(h) - \hat{R}_S(h)| + \varepsilon.$

by arbitrary $\varepsilon > 0$. $R(h_S^{ERM}) - \inf_{h \in \mathcal{H}} R(h) \le 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)|$ $\square$

In practice, performance of ERM is so poor.

- complexity of hypothesis space
- computationally intractable (NP-hard)

SRM (Structural Risk Minimization 结构风险最小化)

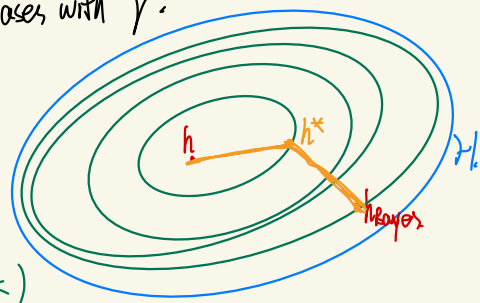
a very complex family $\mathcal{H}$ with no/small approximation error.

$\mathcal{H} = \bigcup_{\gamma \in \Gamma} \mathcal{H}_\gamma$ with complexity of $\mathcal{H}_\gamma$ increases with γ.

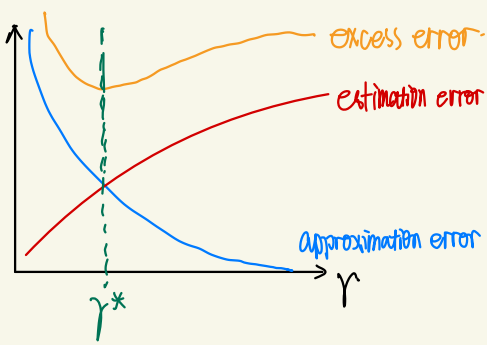
↳ select $\gamma^* \in \Gamma$

$\mathcal{H}_{\gamma^*}$ balances estimation error & approximation error

↳ excess error (upper bounded by excess risk)



"decomposition"



assume $\mathcal{H} = \bigcup_{k \ge 1} \mathcal{H}_k$. $\mathcal{H}_k \subsetneq \mathcal{H}_{k+1}$ $(k \ge 1)$.

SRM: ① selecting the index $k^* \ge 1$

☆ ② finding the ERM hypothesis in $\mathcal{H}_{k^*}$.

$$R(h) \le \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log k}{m}} + \sqrt{\frac{\log 2/\delta}{2m}} \qquad (*)$$

w.p. $\ge 1-\delta$

minimize. $F_k(h) = \hat{R}_S(h) + \mathfrak{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}$

$$\Rightarrow h_S^{SRM} = \underset{k \ge 1,\, h \in \mathcal{H}_k}{\operatorname{argmin}} F_k(h) = \underset{k \ge 1,\, h \in \mathcal{H}_k}{\operatorname{argmin}} \underbrace{\hat{R}_S(h)}_{\in h_S^{ERM}} + \mathfrak{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}} \in k^*$$

Thm (SRM learning guarantee)

$$R(h_S^{SRM}) \le \inf_{h\in\mathcal{H}}\left(R(h) + 2\mathcal{R}_m(\mathcal{H}_{k(h)}) + \sqrt{\frac{\log k(h)}{m}}\right) + \sqrt{\frac{2\log\frac{3}{\delta}}{m}}.$$

$$R(h^*) = \inf_{h\in\mathcal{H}} R(h)$$

Proof: 先证明(*)式成立

$$\Pr\left(\sup_{h\in\mathcal{H}}\left(R(h) - F_{k(h)}(h)\right) > \varepsilon\right) = \Pr\left(\sup_{k\ge1}\sup_{h\in\mathcal{H}_k}\left(R(h) - F_{k(h)}(h)\right) > \varepsilon\right)$$

$$\le \sum_{k=1}^{\infty}\Pr\left(\sup_{h\in\mathcal{H}_k} R(h) - F_{k(h)}(h) > \varepsilon\right) \quad \text{(Union Bound)}$$

$$= \sum_{k=1}^{\infty}\Pr\left(\sup_{h\in\mathcal{H}_k} R(h) - \hat{R}_S(h) - \mathcal{R}_m(\mathcal{H}_k) > \varepsilon + \sqrt{\frac{\log k}{m}}\right)$$

$$\le \sum_{k=1}^{\infty}\exp\left(-2m\left(\varepsilon + \sqrt{\frac{\log k}{m}}\right)^2\right)$$

$$\le \sum_{k=1}^{\infty} e^{-2m\varepsilon^2}\cdot e^{-2\log k} = e^{-2m\varepsilon^2}\cdot\sum_{k=1}^{\infty}\frac{1}{k^2} = \frac{\pi^2}{6}e^{-2m\varepsilon^2} \le \underbrace{2e^{-2m\varepsilon^2}}_{\delta}$$

☆ $R(h) \le \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} = \varepsilon + \sqrt{\frac{\log k}{m}}$

Rademacher complexity generalization bound.

$$\Rightarrow \Pr\left(\sup_{h\in\mathcal{H}}\left(R(h) - F_{k(h)}(h)\right) \le \varepsilon\right) \ge 1-\delta$$

w.p. $\ge 1-\delta$ $\quad R(h) - F_{k(h)}(h) \le \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$ $\quad$ (*)式即证.

下证 learning guarantee. $\quad F_k(h) = \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}_k) + \sqrt{\frac{\log k}{m}}$

r.v. X_1, X_2,

$$\Pr(X_1 + X_2 > \varepsilon) \le \Pr(X_1 > \tfrac{\varepsilon}{2}) + \Pr(X_2 > \tfrac{\varepsilon}{2})$$

$$\Pr\left(R(h_S^{SRM}) - R(h) - 2\mathcal{R}_m(\mathcal{H}_k) - \sqrt{\frac{\log k}{m}} > \varepsilon\right)$$

$$\le \Pr\left(R(h_S^{SRM}) - F_{k(h_S^{SRM})}(h_S^{SRM}) > \frac{\varepsilon}{2}\right) + \Pr\left(F_{k(h_S^{SRM})}(h_S^{SRM}) - R(h) - 2\mathcal{R}_m(\mathcal{H}_k) - \sqrt{\frac{\log k}{m}} > \frac{\varepsilon}{2}\right)$$

$$\overset{(*)}{\le} 2e^{-m\cdot\frac{\varepsilon^2}{2}} + \Pr\left(\hat{R}_S(h) - R(h) - \mathcal{R}_m(\mathcal{H}_k) > \frac{\varepsilon}{2}\right)$$

$$\overset{GB}{\le} 2e^{-m\cdot\frac{\varepsilon^2}{2}} + e^{-m\cdot\frac{\varepsilon^2}{2}} = 3\cdot e^{-m\cdot\frac{\varepsilon^2}{2}} = \delta \Rightarrow \varepsilon = \sqrt{\frac{2\log\frac{3}{\delta}}{m}} \quad \square$$

if $R(h^*) = \inf_{h \in H} R(h) \Rightarrow R(h_S^{SRM}) \le R(h^*) + 2R_m(\mathcal{H}_{k^*}) + \sqrt{\frac{\log k(h^*)}{m}} + \sqrt{\frac{2\log\frac{3}{\delta}}{m}}$

Choosing k^*

$k^* \to F_k(h)$

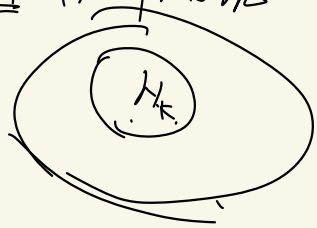
in practice, Bi-Partition ($=\xi$)

$[1, k_{max}]$

k^*

$\min_{h \in H_k} F_k(h) \le \min_{h \in H_k} F_{k+1}(h)$. 对 $k' > k+1$ 的情况都不用考虑

用ERM (k定后只需考虑 $\hat{R}_S(h)$)

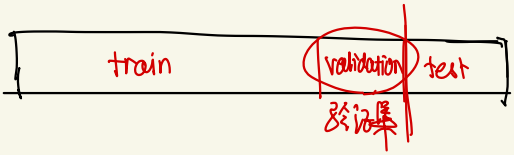


$\mathcal{H}$

But: $\mathcal{H} = \bigcup_{k \ge 1} \mathcal{H}_k$ (countable many) $\to$ hard.

SRM中需用多次ERM. ($O(\log k_{max})$ 次)

Cross-Validation 交叉验证



$(\mathcal{H}_k)_{k \ge 1}$

S : sample set

S_1 : $(1-\alpha)m$ $\to$ training

S_2 : αm $\to$ validation.

$$h_S^{CV} = \underset{h \in \{h_{S_1,k}^{ERM} : k \ge 1\}}{\operatorname{argmin}} \hat{R}_{S_2}(h)$$

Proposition.

$$\Pr\left(\sup_{k \ge 1} \left| R(h_{S_1,k}^{ERM}) - \hat{R}_{S_2}(h_{S_1,k}^{ERM}) \right| > \varepsilon + \sqrt{\frac{\log k}{\alpha m}}\right) \le 4e^{-2\alpha m \varepsilon^2}$$

注比 经验

(S_1, S_2 partition $\Rightarrow$ guarantee)

$\mathcal{H}_k$: h_k^{ERM} (based on S_1)

$\downarrow$

h^{CV} (based on S_2)

Proof: $\Pr\left(\sup_{k\geq 1}\left|R(h_{S_1,k}^{ERM})-\hat{R}_{S_2}(h_{S_1,k}^{ERM})\right|>\varepsilon+\sqrt{\frac{\log k}{\alpha m}}\right)$

$$\leq \sum_{k=1}^{\infty}\Pr\left(\left|R(h_{S_1,k}^{ERM})-\hat{R}_{S_2}(h_{S_1,k}^{ERM})\right|>\varepsilon+\sqrt{\frac{\log k}{\alpha m}}\right) \quad \text{(Union Bound)}$$

fix-S_1

$$=\sum_{k=1}^{\infty}\mathbb{E}_{S_1}\left[\Pr\left(\left|R(h_{S_1,k}^{ERM})-\hat{R}_{S_2}(h_{S_1,k}^{ERM})\right|>\varepsilon+\sqrt{\frac{\log k}{\alpha m}}\;\middle|\;S_1\right)\right]$$

(Law of Total Expectation:

$$\mathbb{E}_{S_1}[\Pr(A|S_1)] = \sum_x \Pr(S_1)\cdot\Pr(A|S_1) = \Pr(A)\,)$$

$$R(h_{S_1,k}^{ERM})=\mathbb{E}_{S_2}[\hat{R}_{S_2}(h_{S_1,k}^{ERM})].$$

by Hoeffding's inequality,

$$\Pr\left(\left|R(h_{S_1,k}^{ERM})-\hat{R}_{S_2}(h_{S_1,k}^{ERM})\right|>\varepsilon+\sqrt{\frac{\log k}{\alpha m}}\;\middle|\;S_1\right)\leq 2e^{-2\boxed{\alpha m}\left(\varepsilon+\sqrt{\frac{\log k}{\alpha m}}\right)^2}$$

$$\leq 2e^{-2\alpha m\varepsilon^2-2\log k}=\frac{2}{k^2}e^{-2\alpha m\varepsilon^2}$$

$$\leq\sum_{k=1}^{\infty}\frac{2}{k^2}e^{-2\alpha m\varepsilon^2}=\frac{\pi^2}{3}e^{-2\alpha m\varepsilon^2}<4e^{-2\alpha m\varepsilon^2} \quad \square$$

Thm (CV. versus SRM)

$$R(h_S^{CV})-R(h_{S_1}^{SRM})\leq 2\sqrt{\frac{\log\max\{K(h_S^{CV}),K(h_S^{SRM})\}}{\alpha m}}+2\sqrt{\frac{\log\frac{4}{\delta}}{2\alpha m}}$$

并不相差太大

$K(h)$: smallest index of k that $h\in\mathcal{H}_k$

Proof: $R(h_S^{CV})\overset{(*)}{\leq}\hat{R}_{S_2}(h_S^{CV})+\sqrt{\frac{\log K(h_S^{CV})}{\alpha m}}+\sqrt{\frac{\log 4/\delta}{2\alpha m}}$

$(*)$ $\Pr\left(\sup_{k\geq 1}\left|R(h_{S_1,k}^{ERM})-\hat{R}_{S_2}(h_{S_1,k}^{ERM})\right|>\varepsilon+\sqrt{\frac{\log k}{\alpha m}}\right)\leq 4e^{-2\alpha m\varepsilon^2}=\delta\Rightarrow\varepsilon=\sqrt{\frac{\log 4/\delta}{2\alpha m}}$

definition of cross-validation

$$\leq\hat{R}_{S_2}(h_{S_1}^{SRM})+\sqrt{\frac{\log K(h_S^{CV})}{\alpha m}}+\sqrt{\frac{\log 4/\delta}{2\alpha m}}$$

$$\overset{(*)}{\leq}\left(R(h_{S_1}^{SRM})+\sqrt{\frac{\log K(h_{S_1}^{SRM})}{\alpha m}}+\sqrt{\frac{\log 4/\delta}{2\alpha m}}\right)+\sqrt{\frac{\log K(h_S^{CV})}{\alpha m}}+\sqrt{\frac{\log 4/\delta}{2\alpha m}}$$

$$h_S^{CV},\ h_{S_1}^{SRM} \in \{h_{S,k}^{ERM} \mid k \geq 1\}$$

$$\leq R(h_{S_1}^{SRM}) + \sqrt{\frac{2\max(K(h_S^{CV}), K(h_{S_1}^{SRM}))}{\alpha m}} + 2\sqrt{\frac{\log 4/\delta}{2\alpha m}}$$

Q.E.D. □

trade-off : $\begin{cases} \alpha: \text{small} & \text{to learn \& avoid unfavorable regimes.} \\ \alpha: \text{large} & \text{CV} \sim \text{SRM bound "informative."} \end{cases}$

In practice, the amount of labeled data available is often limited.

n-fold cross validation

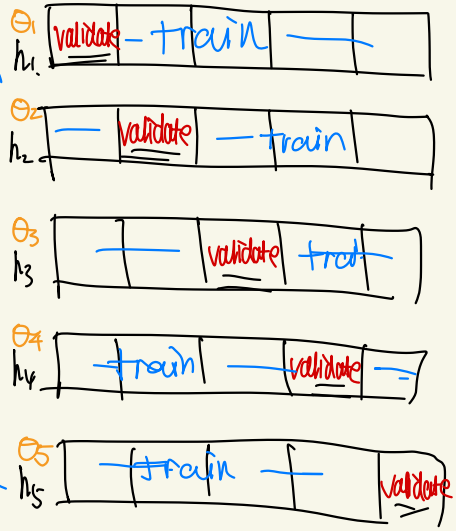
"n折"

$$\hat{R}_{CV}(\theta) = \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m/n} \mathcal{L}(h_i(x_{ij}), y_{ij})$$

h^{CV}

$\theta^* = \underset{i}{\operatorname{argmin}}\, \mathcal{L}(\theta_i)$

(hyper-parameter)



trade-off
- n大: bias 小, var 大
- n小: bias 大, var 小

n is typically 5 or 10

Special case: $n=m$. 留一法 (Leave-One-Out)

unbiased estimate

efficiently computation

costly computation

> trade-off

* performance evaluation

ε : error

σ : standard deviation

SRM : $\mathcal{H} = \bigcup_{\gamma \in \Gamma} \mathcal{H}_\gamma$

uncountable

e.g. $\mathcal{H}$ 所有线性函数

$\mathcal{H}_\gamma = \{x \mapsto w \cdot \Phi(x) : \|w\| \leq \gamma\}$

"dense"

e $\|w\|_p = \left(\sum_{i=1}^{n} |w_i|^p\right)^{1/p}$

$$\underset{\gamma>0,\ h\in\mathcal{H}_\gamma}{\operatorname{argmin}} F_\gamma(h) = \underset{\gamma>0,\ h\in\mathcal{H}_\gamma}{\operatorname{argmin}} \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}_\gamma) + \sqrt{\frac{\log\gamma}{m}}$$

penalty(γ, m) 惩罚

Often $\Omega: \mathcal{H} \to \mathbb{R}$

$$\boxed{\underset{h\in\mathcal{H}}{\operatorname{argmin}} \hat{R}_S(h) + \lambda \cdot \Omega(h)}$$

☆☆☆

带结构风险 (Structral Risk)

经验风险 (empirical risk)

Regularization term

正则化

$\lambda > 0$: regularization parameter

"结构风险"

反映了模型的结构

Occam Razor's principle

in most algorithm,

$\boxed{\Omega(h): \|\cdot\| \text{ (norm)}}$

完备 "complete"

$\mathcal{H}$: Hilbert Space 内积空间

Banach Space 赋范线性空间

e.g
- $\|\cdot\|_1$: lasso regression
- $\|\cdot\|_2$: ridge regression.

λ choose ? cross-validation
- $\lambda \to 0$: ERM
- $\lambda \to \infty$: null.

} trade-off.

$\hat{R}_S(h) + \lambda \cdot \|h\|$

convex

"Convex Optimization" 凸优化

$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{f_h(x_i)\neq y_i}$ (zero/one loss) $\Longrightarrow$ ☆ Surrogate loss function

替代损失函数

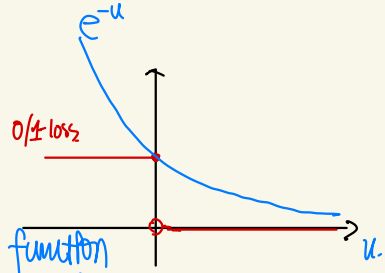
$\leq \mathbb{1}_{y_i h(x_i) \leq 0}$: non-convex

(Heaviside step function)

derivative: Dirac Delta function

$h: \mathcal{X} \to \mathbb{R}$.

$$f_h(x) = \begin{cases} +1, & h(x) \geq 0 \\ -1, & h(x) < 0 \end{cases}$$



凸 convex

upper-bound of 0/1-loss

0/1-loss: $R(h) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}_{f_h(x)\neq y}\right]$

令 $\eta(x) = \Pr(y=+1 \mid x)$ 真实样例标记概率

后验概率

$$= \mathbb{E}_{x\sim D_x}\left[\eta(x)\cdot \mathbb{1}_{h(x)<0} + (1-\eta(x))\cdot \mathbb{1}_{h(x)\geq 0}\right]$$

true label ⇔ prediction

Bayes Classifier: $h^*(x) = \eta(x) - \frac{1}{2}$ → Bayes error: $R^* = R(h^*)$

Lemma $\forall h: \mathcal{X}\to\mathbb{R}$, $R(h) - R^* = 2\,\mathbb{E}_{x\sim D_x}\left[|h^*(x)|\cdot \mathbb{1}_{h(x)h^*(x)\leq 0}\right]$

excess error < estimation, approximation

Proof: $R(h) = \mathbb{E}_{x\sim D_x}\left[\eta(x)\cdot \mathbb{1}_{h(x)<0} + (1-\eta(x))\cdot \mathbb{1}_{h(x)\geq 0}\right]$

$$= \mathbb{E}_{x\sim D_x}\left[\eta(x)\cdot \mathbb{1}_{h(x)<0} + (1-\eta(x))\cdot\left(1 - \mathbb{1}_{h(x)<0}\right)\right]$$

$$= \mathbb{E}_{x\sim D_x}\left[(2\eta(x)-1)\,\mathbb{1}_{h(x)<0} + (1-\eta(x))\right]$$

$$= \mathbb{E}_{x\sim D_x}\left[2h^*(x)\,\mathbb{1}_{h(x)<0} + (1-\eta(x))\right]$$

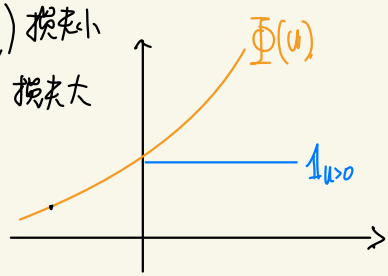
$$\Rightarrow R(h) - R^* = \mathbb{E}_{x\sim D_x}\left[2h^*(x)\cdot\left(\mathbb{1}_{h(x)<0} - \mathbb{1}_{h^*(x)<0}\right)\right] = \text{RHS} \quad \square$$

$h(x)$	$h^*(x)$		
+	+	0	0
+	−	$2h^*(x)\cdot(-1)$	$-2h^*(x)$
−	+	$2h^*(x)\cdot 1$	$2h^*(x)$
−	−	0	0

Surrogate < convex, non-decreasing

$\Phi: \mathbb{R}\to\mathbb{R}$, s.t. $\mathbb{1}_{u\leq 0} \leq \Phi(-u)$

Φ-loss : $\Phi(-yh(x))$ $\begin{cases} yh(x)>0 \text{ 正确} & \Phi(\text{negative}) \text{ 损失小} \\ yh(x)<0 \text{ 错误} & \Phi(\text{positive}) \text{ 损失大} \end{cases}$



$$\mathcal{L}_\Phi(h) = \mathbb{E}_{(x,y)\sim D}\left[\Phi(-yh(x))\right]$$

$\eta(x) = \Pr(y=+1|x)$

$$= \mathbb{E}_{x\sim D_x}\left[\underbrace{\eta(x)\cdot\Phi(-h(x))}_{\text{正例}} + \underbrace{(1-\eta(x))\cdot\Phi(h(x))}_{\text{负例}}\right]$$

$$\mathbb{1}_{yh(x)\le 0} \le \Phi(-yh(x))$$: upper bound.

$$\Rightarrow R(h) \le \mathcal{L}_\Phi(h)$$

0/1-loss　　Φ-loss　　generalization.

def $\mathcal{L}_\Phi(x,u) = \eta(x)\cdot\Phi(-u) + (1-\eta(x))\Phi(u)$

每个 u 是一个假设 h 的输出 $h(x)$

$$\Rightarrow \mathcal{L}_\Phi(h) = \mathbb{E}_{x\sim D_x}\left[\mathcal{L}_\Phi(x,h(x))\right]$$

Bayes solution for loss function $\mathcal{L}_\Phi$: $h^*_\Phi(x) = \underset{u\in[-\infty,+\infty]}{\operatorname{argmin}}\ \mathcal{L}_\Phi(x,u)$

$$= \underset{u\in[-\infty,+\infty]}{\operatorname{argmin}}\ \eta(x)\cdot\Phi(-u) + (1-\eta(x))\cdot\Phi(u)$$

$$\Rightarrow \mathcal{L}^*_\Phi = \mathbb{E}_{(x,y)\sim D}\left[\Phi(-y\,h^*_\Phi(x))\right]$$

Special case: $\eta(x)=0$. $u=-\infty$ i.e. $h^*_\Phi(x)=-\infty$

$\eta(x)=1$. $u=+\infty$ i.e. $h^*_\Phi(x)=+\infty$

$\eta(x)=\frac{1}{2}$. $\mathcal{L}_\Phi(x,u) = \frac{1}{2}\left(\Phi(-u)+\Phi(u)\right)$

$\overset{\text{convexity}}{\ge} \Phi\left(-\frac{u}{2}+\frac{u}{2}\right) = \Phi(0)$

$\Rightarrow u=0$ i.e. $h^*_\Phi(x)=0$

Proposition Φ $\begin{cases} \text{convex} \\ \text{non-decreasing} \\ \text{differentiable at } 0.\ \Phi'(0)>0 \end{cases}$

$\boxed{\text{the minimizer of } \Phi \text{ defines the Bayes classifier}}$.

$$\forall x\in\mathcal{X}.\begin{cases} h^*_\Phi(x)>0 \Leftrightarrow h^*(x)>0 \\ h^*_\Phi(x)=0 \Leftrightarrow h^*(x)=0 \\ h^*_\Phi(x)<0 \Leftrightarrow h^*(x)<0 \end{cases}$$

☆ $h^*_\Phi(x)$ 的符号和 $h^*(x)$ 的符号一一对应!

$\Rightarrow$ Surrogation Φ 不影响 Bayes Classifier 的拟合

Proof: fixed $x \in \mathcal{X}$. $\eta(x)=0 \Rightarrow h^*_\Phi(x) = -\infty$. $h^*(x) = -\frac{1}{2}$

$\eta(x) = 1 \Rightarrow h^*_\Phi(x) = +\infty$. $h^*(x) = +\frac{1}{2}$. Same sign.

$$L_\Phi(x,u) = \eta(x)\cdot\Phi(-u) + (1-\eta(x))\cdot\Phi(u)$$

$$\left.\frac{\partial L_\Phi(x,u)}{\partial u}\right|_{u=u^*} = -\eta(x)\cdot\underbrace{\Phi'(-u^*)}_{V_1} + (1-\eta(x))\cdot\underbrace{\Phi'(u^*)}_{V_2} = 0$$

u^* is the minimizer of $L_\Phi(x,u)$, (by convexity)

$$\Rightarrow \eta(x)\cdot V_1 = (1-\eta(x))\cdot V_2.$$

if $u^*=0. \Rightarrow V_1 = V_2 = \Phi'(0) > 0 \Rightarrow \eta(x) = 1-\eta(x) \Rightarrow \eta(x) = \frac{1}{2} \Rightarrow h^*(x) = 0$

if $h^*(x)=0 \Rightarrow \eta(x) = \frac{1}{2} \Rightarrow V_1 = V_2 \Rightarrow u^* = 0$

$\Rightarrow h^*(x) = 0 \Leftrightarrow \eta(x) = \frac{1}{2} \Leftrightarrow \underbrace{h^*_\Phi(x)}_{u^*} = 0$

Convexity: $u_1 > u_2 \Rightarrow \Phi'(u_1) \geq \Phi'(u_2) \Rightarrow V_1 \geq V_2$ $\quad \eta(x) \notin \{0, 1, \frac{1}{2}\}$

if $\underbrace{u^*>0}_{h^*_\Phi(x)>0} \Rightarrow u^* > -u^* \Rightarrow V_1 \leq V_2 \xrightarrow[V_1 \neq V_2]{} V_1 < V_2 \Rightarrow \eta(x) > 1-\eta(x) \Rightarrow \eta(x) > \frac{1}{2} \Rightarrow h^*(x) > 0$

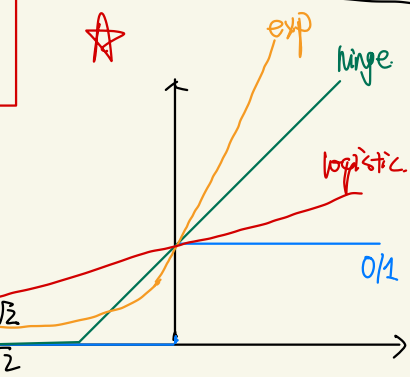
if $h^*(x) > 0 \Rightarrow \eta(x) > \frac{1}{2} \Rightarrow V_1 < V_2 \Rightarrow u^* > -u^* \Rightarrow u^* > 0$

$h^*_\Phi(x)$ ($u^*(x)$) 和 $h^*(x)$ 的符号一一对应 □.

Thm assume $\exists\, s \geq 1.\ c > 0.$ s.t. $\forall x \in \mathcal{X}$. $|h^*(x)|^s = |\eta(x) - \frac{1}{2}|^s \leq c^s \left(\overset{\Phi(0)}{L_\Phi(x,0)} - L_\Phi(x, h^*_\Phi(x)) \right)$

then $\forall h \in \mathcal{H}$. $\boxed{R(h) - R^* \leq 2c \cdot \left(L_\Phi(h) - L^*_\Phi \right)^{\frac{1}{s}}}$

excess error (by 0/1 loss) $\qquad$ excess-error (by Φ-loss)



Surrogate:

- Hinge loss: $\Phi(u) = \max(0, 1+u)$ → Soft-SVM $\quad s=1,\ c=\frac{1}{2}$
- Exponential loss: $\Phi(u) = \exp(u)$ → AdaBoost $\quad s=2.\ c=1/\sqrt{2}$
- Logistic loss: $\Phi(u) = \log_2(1+e^u)$ → Logistic Regression $\quad s=2.\ c=1/\sqrt{2}$

Proof: $\Phi(-2h^*(x)h(x)) = \Phi((1-2\eta(x))\cdot h(x))$

$$= \Phi(\eta(x)\cdot(-h(x)) + (1-\eta(x))\cdot h(x))$$

$$\overset{\text{Convexity}}{\leq} \eta(x)\cdot\Phi(-h(x)) + (1-\eta(x))\cdot\Phi(h(x)) = L_\Phi(x, h(x))$$

$$R(h) - R(h^*) = 2\mathbb{E}_{x\sim D_x}\left[|h^*(x)|\cdot \mathbb{1}_{h^*(x)h(x)\leq 0}\right]$$

$$= \mathbb{E}_{x\sim D_x}\left[(2\eta(x)-1)\cdot \mathbb{1}_{h^*(x)h(x)\leq 0}\right]$$

$$\leq \mathbb{E}_{x\sim D_x}\left[|2\eta(x)-1|^s \mathbb{1}_{h^*(x)h(x)\leq 0}\right]^{\frac{1}{s}}$$

Jenson's ineq. $|\mathbb{E}[X]|^s \leq |\mathbb{E}[X^s]|$

$$= 2c\cdot \mathbb{E}_{x\sim D_x}\left[\left(\Phi(0) - L_\Phi(x, h^*_\Phi(x))\right)\cdot \mathbb{1}_{h^*(x)h(x)\leq 0}\right]^{1/s}$$

Φ-nondecreasing

$$\leq 2c\cdot \mathbb{E}_{x\sim D_x}\left[\left(\Phi(-2h^*(x)h(x)) - L_\Phi(x, h^*_\Phi(x))\right)\cdot \mathbb{1}_{h^*(x)h(x)\leq 0}\right]^{1/s}$$

$$\leq 2c\cdot \mathbb{E}_{x\sim D_x}\left[\left(L_\Phi(x, h(x)) - L_\Phi(x, h^*_\Phi(x))\right)\cdot \mathbb{1}_{h^*(x)h(x)\leq 0}\right]^{1/s}$$

$$\leq 2c\, \mathbb{E}_{x\sim D_x}\left[L_\Phi(x, h(x)) - L_\Phi(x, h^*_\Phi(x))\right] = 2c\left(\mathcal{L}_\Phi(h) - \mathcal{L}^*_\Phi\right)^{1/s} \quad \square$$

Summary: Model Selection

$R(h) - R^* \to$ ERM $\to$ SRM, CV

$$\min \hat{R}_S(h) + \lambda\cdot\Omega(h)$$

"convex optimization"

non-convex

surrogate