Course Note for CPSC532D STAT Learn Theory

L2:

$$L_D(\hat{h}_S) - L_D(h^*) \le \underbrace{L_D(\hat{h}_S) - L_S(\hat{h}_S)}_{\text{uniform convergence}} + \underbrace{L_S(h^*) - L_D(h^*)}_{\text{Hoeffding}}$$

estimation error

$$\frac{1}{m}\sum_{i=1}^{m} \ell(h^*, z_i) - \mathbb{E}_z[\ell(h^*, z)]$$

$$\underbrace{L_D(\hat{h}_S) - L_{Bayes}}_{\text{excess error}} = \underbrace{L_D(\hat{h}_S) - \inf_{h\in\mathcal{H}} L_D(h)}_{\text{estimation error}} + \underbrace{\inf_{h\in\mathcal{H}} L_D(h) - L_{Bayes}}_{\text{approximation}}$$

$= \inf_h \mathbb{E}\,L(h(x), y)$ for measurable h

increase $\mathcal{H}$: estimation error ↑, approximation ↓

- central limit : ✗ $O(m)$
- concentration equalities : ✓ $O(m)$

Hoeffding: Let $x_1, \dots, x_m$ be iid with :

- $\mathbb{E}\,x_i = \mu$
- $\Pr(a \le x_i \le b) = 1$

Let $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$, Then

$$\Pr\left(\bar{x} \le \mu + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \ge 1-\delta \iff \Pr\left(\bar{x} > \mu + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) < \delta$$

$\therefore Y_i = -x_i$. Then $\Pr\left(\bar{Y} \le -\mu + (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \ge 1-\delta$

$$\hookrightarrow \Pr\left(\bar{x} \ge \mu - (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \ge 1-\delta \iff \Pr\left(\bar{x} < \mu - (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) < \delta$$

$$\Rightarrow \Pr\left(|x-\mu| > \mu - (b-a)\sqrt{\frac{\log 1/\delta}{2m}}\right) \le 2\delta$$

$L_S(h^*) - L_D(h^*)$

$$\Pr\left(L_S(h^*) - L_D(h^*) \ge (b-a)\sqrt{\frac{1/\delta}{2m}}\right) \le \delta$$

⚠ we can't use Hoeffding on $L_D(\hat{h}_S) - L_S(\hat{h}_S)$

$\ell(\hat{h}_S, z_i)$ are not independent of each other

change z_i, will change $\hat{h}_S$

↳ Uniform convergence !! : for every single $h \in \mathcal{H}$, gap is not too big

$\sup_{h \in H} L_D(h) - L_S(h) \leq \varepsilon \Rightarrow (\forall h \in H\ L_D(h) - L_S(h) \leq \varepsilon)$

① if $|H| < \infty$: union bound

add up!

$P(A \cup B) \leq P(A) + P(B)$

Not tight enough because of $P(AB)$

$q_S(h) = L_D(h) - L_S(h)$

$\sup_{h \in H} q_S(h) < \varepsilon$

$Pr(\exists h \in H,\ q_S(h) > \varepsilon) \leq \sum_{h \in H} Pr(q_S(h) > \varepsilon)$

now h can be fixed

Assume bounded $\begin{cases} \ell(h, z) \in [a, b]\ \forall z, h, \\ H \text{ finite} \\ \hat{h}_S \text{ is an ERM} \end{cases}$

$L_D(\hat{h}_S) - \min_{h \in H} L_D(h) \leq L_D(\hat{h}_S) - L_S(\hat{h}_S) + L_S(h^*) - L_D(h^*)$

$\leq \max_{h \in H} [L_D(h) - L_S(h)] + L_S(h^*) - L_D(h^*)$

by showing: $\begin{cases} \cdot\ \forall h \in H,\ Pr(L_D(h) - L_S(h) > \varepsilon) \leq \frac{\delta}{|H|+1} \\ \cdot\ Pr(L_S(h^*) - L_D(h^*) > \varepsilon) \leq \frac{\delta}{|H|+1} \end{cases}$

$\therefore Pr(L_D(\hat{h}_S) - \min_{h \in H} L_D(h) \leq 2\varepsilon) \geq 1 - \delta$

$\varepsilon = (b-a)\sqrt{\frac{1}{2m} \log \frac{|H|+1}{\delta}}$

$Pr(L_D(\hat{h}_S) - \min_{h \in H} L_D(h) \leq \sqrt{\frac{2}{m} \log \frac{|H|+1}{\delta}}) \geq 1 - \delta$

L3. MARKOV

x cannot be too big with high prob $\mathbb{P}(X \geq 0) = 1$

- Markov's inequality: X nonnegative-valued random variable
  $\Pr(X \geq t) \leq \frac{1}{t}\mathbb{E}X$ for all $t > 0$ — is weak because only assume nonneg

Pf $\begin{cases} X \geq 0 \\ X \geq t \text{ when } X \geq t \end{cases}$ $\therefore$ $X \geq t\mathbb{1}(X \geq t)$ (always holds) $\therefore \mathbb{E}X \geq t\mathbb{E}\mathbb{1}(X \geq t)$ (t is scalar) $= t\Pr(X \geq t)$

Let $\delta = \frac{\mathbb{E}X}{t}$ then $\Pr(X \geq \frac{\mathbb{E}X}{\delta}) \leq \delta \Rightarrow \Pr(X \leq \frac{\mathbb{E}X}{\delta}) \geq 1-\delta$

- Chebyshev's inequality: for any X (just random variable), $\Pr(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{\mathrm{Var}X}{\varepsilon^2}$

Pf $\Pr(|X-\mathbb{E}X| \geq \varepsilon) = \Pr((X-\mathbb{E}X)^2 \geq \varepsilon^2) \leq \frac{1}{\varepsilon^2}\mathbb{E}(X-\mathbb{E}X)^2 = \frac{\mathrm{Var}X}{\varepsilon^2}$ (Markov inequality)

with probability at least $1-\delta$, $|X - \mathbb{E}X| \leq \sqrt{\frac{\mathrm{Var}X}{\delta}}$ ($\frac{\mathrm{Var}X}{\varepsilon^2} = \delta$) $= \frac{\text{S+dix}}{\sqrt{\delta}}$

- Consider iid $X_1, X_2, \cdots, X_m$, $\mathbb{E}X_i = \mu$, $\mathrm{Var}(X_i) = \sigma^2$

$\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$ $\mathbb{E}\bar{X} = \frac{1}{m}\sum_{i=1}^m \mathbb{E}X_i = \mu$ $\mathrm{Var}(\bar{X}) = \frac{1}{m^2}\cdot m\sigma^2 = \frac{\sigma^2}{m}$

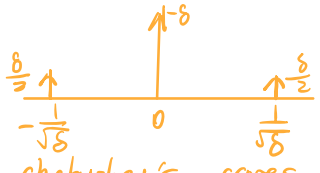
$\Rightarrow$ with $\Pr \geq 1-\delta$, $|\bar{X} - \mu| \leq \frac{\sigma}{\sqrt{m\delta}}$

chebyshev's inequality may not be good enough!

$\sqrt{m}$ is fine, but the dependence on δ is the problem $\sqrt{\log\frac{1}{\delta}}$ better than $\frac{1}{\sqrt{\delta}}$

$\Pr(X=0) = 1-\delta$, $\Pr(X = \frac{1}{\sqrt{\delta}}) = \Pr(X = -\frac{1}{\sqrt{\delta}}) = \frac{1}{2}\delta$ $\mathbb{E}X = 0$ $\mathrm{Var}X = 1$

$\Pr(|\bar{X}| \leq \frac{1}{\sqrt{m\delta}}) \geq 1-\delta$



chebyshev's cares about the worst case

- Chernoff bounds: construct a non-negative random variable

$\lambda > 0$ arbitrary $Y = e^{\lambda(X-\mu)}$ $\Pr(Y \geq t) \leq \frac{1}{t}\mathbb{E}Y$

$\Pr(e^{\lambda(X-\mu)} \geq e^{\lambda\varepsilon}) \leq e^{-\lambda\varepsilon}\mathbb{E}e^{\lambda(X-\mu)}$

$= \Pr(\lambda(X-\mu) \geq \lambda\varepsilon)$ centred moment-generating function $M_X(\lambda)$

$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots$

$M_X(\lambda) = \mathbb{E}\, e^{\lambda(X-\mu)} = 1 + \lambda \underbrace{\mathbb{E}[X-\mu]}_{0} + \frac{\lambda^2}{2!}\mathbb{E}[(X-\mu)^2] + \frac{\lambda^3}{3!}\mathbb{E}[(X-\mu)^3] + \cdots$

kth derivative of $M_X(\lambda)$ at $\lambda = 0$: $M_X^{(k)}(0) = \mathbb{E}[(X-\mu)^k]$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\, e^{\lambda(X-\mu)} = e^{\frac{1}{2}\lambda^2\sigma^2}$

for $X \sim \mathcal{N}(0,1)$

$$\mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,1)} e^{\lambda X} = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} e^{\lambda x} dx = e^{-\frac{1}{2}\lambda^2} \int \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2} dx}_{1} = e^{-\frac{1}{2}\lambda^2}$$

$\mathcal{N}(\lambda, 1)$

for $Y \sim \mathcal{N}(0, \sigma^2)$

$$\mathop{\mathbb{E}}_{Y\sim\mathcal{N}(0,\sigma^2)} e^{\lambda Y} = \mathop{\mathbb{E}}_{\sigma X\sim\mathcal{N}(0,\sigma^2)} e^{\lambda\sigma X} = \mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,1)} e^{(\lambda\sigma)X} = e^{\frac{1}{2}\lambda^2\sigma^2}$$

$$\Pr(X \geq \mu + \varepsilon) \leq e^{-\lambda\varepsilon} e^{\frac{1}{2}\lambda^2\sigma^2}$$

we can optimize λ to get tighter bound

$\sigma^2\lambda = \varepsilon \rightarrow \lambda = \frac{\varepsilon}{\sigma^2} \rightarrow e^{\frac{1}{2}\frac{\varepsilon^2}{\sigma^2} - \frac{\varepsilon^2}{\sigma^2}} = e^{-\frac{\varepsilon^2}{2\sigma^2}}$

$$\Pr(X \geq \mu + \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

$e^{-\frac{\varepsilon^2}{2\sigma^2}} = \delta \qquad \varepsilon = \sqrt{2\sigma^2 \log\frac{1}{\delta}}$

For $X \sim \mathcal{N}(\mu, \sigma^2)$:

- with probability at least $1-\delta$ $\quad X \leq \mu + \sigma\sqrt{2\log\frac{1}{\delta}}$

If X is s.t. $\mathbb{E}\, e^{\lambda(X-\mathbb{E}X)} \leq e^{\frac{1}{2}\lambda^2\sigma^2}$, then $\Pr(X \leq \mathbb{E}X + \sigma\sqrt{2\log\frac{1}{\delta}}) \geq 1-\delta$

$SG(\sigma)$ sub-gaussian with σ (no heavier tails than Gaussian)

σ does not imply the variance of the sub-gaussian X may have relation but not equal !

- sub-Gaussian · $X \sim SG(\sigma)$

random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian with σ

(i) if $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists and for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}\lambda^2\sigma^2}$

if $\sigma_1 \leq \sigma_2$ then $X \sim SG(\sigma_1) \Rightarrow X \sim SG(\sigma_2)$ $\quad SG(\sigma_1) \subseteq SG(\sigma_2)$

(ii) If $X \in SG(\sigma)$, $aX \in SG(|a|\sigma)$ for $\forall a \in \mathbb{R}$

$M_X(\lambda) \leq e^{\frac{1}{2}\lambda^2\sigma^2} \Rightarrow \mathbb{E}\, e^{\lambda(aX - a\mathbb{E}X)} = \mathbb{E}\, e^{a\lambda(X-\mathbb{E}X)} \leq e^{\frac{1}{2}(a\lambda)^2\sigma^2} = e^{\frac{1}{2}\lambda^2(|a|\sigma)^2}$

(iii) If $X_1 \in SG(\sigma_1)$, $X_2 \in SG(\sigma_2)$ are independent $\quad X_1 + X_2 \in SG(\sqrt{\sigma_1^2 + \sigma_2^2})$

$\mathbb{E}\, e^{\lambda(X_1+X_2-\mathbb{E}X_1-\mathbb{E}X_2)} = \mathbb{E}\, e^{\lambda(X_1-\mathbb{E}X_1)}\, \mathbb{E}\, e^{\lambda(X_2-\mathbb{E}X_2)} \leq e^{\frac{1}{2}\lambda^2\sigma_1^2} \cdot e^{\frac{1}{2}\lambda^2\sigma_2^2}$

we can use the above properties to construct a set of X

Hoeffding's lemma:

a real-valued random variable bounded in $[a,b]$ is $SG(\frac{b-a}{2})$

i.e. $Pr(a \le x \le b) = 1$, $X \in SG(\frac{b-a}{2})$ see pf in notes 3.

- Hoeffding: If $x_1, x_2, \dots x_m$ iid with $\mathbb{E}x_i = \mu$, $Pr(a \le x_i \le b) = 1$

  then $Pr(\frac{1}{m}\sum_{i=1}^{m} x_i > \mu + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}) \le \delta$

  $\frac{1}{m}\sum_{i=1}^{m} x_i \in SG(\frac{1}{m}\sqrt{m(\frac{b-a}{2})^2}) = SG(\frac{b-a}{2\sqrt{m}})$

  with prob at least $1-\delta$ $\frac{1}{m}\sum_{i=1}^{m} x_i \le \mu + \frac{b-a}{2\sqrt{m}}\sqrt{2\log\frac{1}{\delta}} = \mu + (b-a)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$

Lec 4. PAC Learning (infinite $\mathcal{H}$)

learning algorithm from sample to $\mathcal{H}$

Defn: an algorithm A is agnostically PAC learns $\mathcal{H}$ with loss ℓ

if there is a function $m(\varepsilon, \delta)$ sample complexity function

s.t. for any D, for any $\varepsilon, \delta \in (0,1)$

if $S \sim D^m$ with $m \ge m(\varepsilon, \delta)$

then $Pr(L_D(A(S)) \le \inf_{h \in \mathcal{H}} L_D(h) + \varepsilon) \ge 1-\delta$

solve for m will get sample complexity

efficient: polynomial run time

Def $\mathcal{H}$ is agnostically PAC learnable if $\exists A$ that agnostically PAC learn $\mathcal{H}$

agnostic PAC is the worst case, $\mathcal{H}$ has nothing to do with data distribution.

an $m(\varepsilon, \delta)$ should work for any D

PAC learnable does not show how learn quickly in terms of m.

- Agnostically PAC learning: work for any distribution ERM infinite $\mathcal{H}$
- PAC learning is weaker: work only for realizable distribution

Def (PAC-Learn) $\mathcal{A}$ PAC-learns $\mathcal{H}$ if

$\exists\, m : (0,1)^2 \to \mathbb{N}$ s.t.

for any (ε, δ), for any realizable $\mathcal{D}$, if $m \geq m(\varepsilon, \delta)$

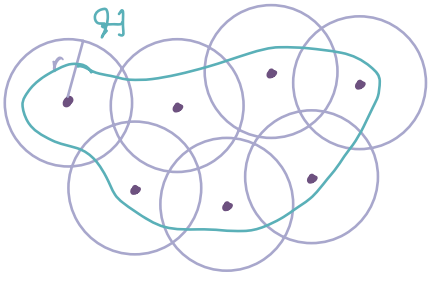
$\Pr_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(\mathcal{A}(S)) \leq \varepsilon) \geq 1-\delta$

Def (realizable). for a non-negative loss $\ell(h, z) \geq 0$.

$\mathcal{D}$ is realizable by $\mathcal{H}$ if there exists an $h^* \in \mathcal{H}$ s.t. $L_{\mathcal{D}}(h^*) = 0$

we already know finite case, what about infinite $\mathcal{H}$?

Idea of covering number



still : uniform convergence

$$\sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_S(h)$$

$$= \sup_{h\in\mathcal{H}} L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h_j) + L_{\mathcal{D}}(h_j) - L_S(h_j) + L_S(h_j) - L_S(h) \quad \forall\, h_j$$

$$\leq \sup_{h\in\mathcal{H}} [L_{\mathcal{D}}(h) - L_{\mathcal{D}}(h_j)] + \max_{h_j\in T} [L_{\mathcal{D}}(h_j) - L_S(h_j)] + \sup_{h\in\mathcal{H}} [L_S(h_j) - L_S(h)]$$

Lipschitz; union bound on size of T; finite; as ρ decrease, the set to cover points become tighter

T is a ρ-cover for $\mathcal{H}$ of size $N(\mathcal{H}, \rho)$

specific case : Logistic regression

$Z = X \times Y \quad X = \mathbb{R}^d \quad Y = \{-1, 1\}$

$\mathcal{H} = \{x \mapsto w \cdot x : w \in \mathbb{R}^d, \|w\| \leq B\}$ equivalent to L_2 reg

$\ell_{\log}(h, (x,y)) = \ell_y(h(x)) = \log(1 + \exp(-y\, h(x)))$

use absolute value to make analysis easier, but results looser

$$|L_{\mathcal{D}}(h_w) - L_{\mathcal{D}}(h_v)| = \left| \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell_{\log}(h_w, (x,y)) - \ell_{\log}(h_v, (x,y)) \right|$$

h_w: weight w; h_v: weight v

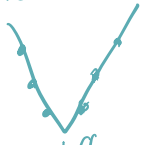
$\le \mathbb{E}\,|l_y(h_w(x)) - l_y(h_v(x))|$ l_y is G-Lipschitz

$\le G\,\mathbb{E}\,|h_w(x) - h_v(x)|$

$G=1$ for l_y

input close → output close

smoothness



G-Lipschitz: l_y is G-Lips if

$$|l_y(\hat{y}_1) - l_y(\hat{y}_2)| \le G\,|\hat{y}_1 - \hat{y}_2| \quad \forall \hat{y}_1, \hat{y}_2$$

Euclidean distance (can be other distance if particular specified)

if f is differetiable the Lips constance is the largest |derivatives| (upper bound of |derivative|)

$\Rightarrow$ if f' exist everywhere, and $\sup |f'(x)| \le G$, f is G-Lips

$$|f(x) - f(x')| = \left|\int_x^{x'} f'(t)\,dt\right| \le \int_x^{x'} |f'(t)|\,dt \le \int_x^{x'} G\,dt = G|x' - x|$$

for the LR: $|h_w(x) - h_v(x)| = |w\cdot x - v\cdot x| = |\langle w - v\rangle \cdot x| \le \|w - v\| \cdot \|x\|$

$l_y(\hat{y}) = \log(1 + \exp(-y\hat{y}))$ $\quad l_y'(\hat{y}) = \dfrac{-y\exp(-y\hat{y})}{1+\exp(-y\hat{y})} = \dfrac{-y}{\exp(y\hat{y})+1} = \begin{cases} -\frac{1}{e^{\hat{y}}+1}, & y=1 \\ \frac{1}{e^{-\hat{y}}+1}, & y=-1 \end{cases}$

$|l_y'(\hat{y})| \le 1$ $\quad l_y$ is 1-Lipschitz

$ax_1 + bx_2 \le \sqrt{a^2+b^2}\sqrt{x_1^2+x_2^2}$ $\quad a^2x_1^2 + b^2x_2^2 + a^2x_2^2 + b^2x_1^2 - 2abx_1x_2 = (ax_2 - bx_1)^2$

$$|h_w(x) - h_v(x)| = |w\cdot x - v\cdot x| = |\langle w - v\rangle \cdot x| \le \|w - v\| \cdot \|x\|$$

Then ① $|L_D(h_w) - L_D(h_v)| \le \underbrace{(\mathbb{E}\,\|x\|)}_{\text{constant of distribution}} \cdot \|w - v\|$ distance from h to $h_{\hat{y}}$ at most ρ

$\le C\rho$

Assume $\|x\| \le C$ on $\mathcal{D}$.

② $|L_S(h_w) - L_S(h_v)| \le \frac{1}{m}\sum_{i=1}^m \|x_i\|\,\|w - v\| \le C\rho$

③ $|L_D(h_{\hat{y}}) - L_S(h_{\hat{y}})| \le (b-a)\sqrt{\frac{1}{2m}\log\frac{N(B,\rho)}{\delta}}$

l_y is not bounded, but we assume $\|w\| \le B$ and $\|x\| \le C$ so we can use hoeffdings

$|h_w(x)| = |w\cdot x| \le \|w\|\cdot\|x\| \le BC$

$b - a \le \log(1 + \exp(BC)) \le BC + 1$ (just simpler)

So $\sup_{h\in H} L_D(h) - L_S(h) \le 2C\rho + (BC+1)\sqrt{\frac{1}{2m}\log\frac{N(B,\rho)}{\delta}}$

what is the covering number?

$N(B,\rho) \le \left(\frac{3B}{\rho}\right)^d$ (proof in notes Ψ)

$$\sup_{h\in H} L_D(h) - L_S(h) \le 2C\rho + (BC+1)\sqrt{\frac{1}{2m}\left(\log\frac{1}{\delta} + d\log\frac{3B}{\rho}\right)}$$

how to choose ρ to minimize the right term?

Let $\rho = \frac{B}{\sqrt{m}}$ (roughly optimal)

$$\le \frac{2CB}{\sqrt{m}} + (BC+1)\sqrt{\frac{\log\frac{1}{\delta} + \frac{d}{2}\log(9m)}{2m}}$$

"C"

dominate by $O_p\left(\sqrt{\frac{\log m}{m}}\right)$

Big Op notation means some random variable is stochastically bounded (Hoeffding)

$$\Pr\left(\sup_{h\in H} L_D(h) - L_S(h) > \frac{2BC}{\sqrt{m}} + (BC+1)\sqrt{\frac{\log 1/\delta + \frac{d}{2}\log 9m}{2m}}\right) < \delta$$

- the sample complexity depends on the input distribution $E\|X\| \le c$
  X not allowed in agnostically PAC learning
  so this bound doesn't show Linear Regression (ERM+) is agnostically PAC learnable
- we can never achieve 0 loss on any D, so it's not realizable

L5 Rademacher complexity

distribution-specific complexit

covering number approach : depend on dimension d.
need bounded norm. B. scale sensetive

Rademacher complexity : do not depend on dimension d.

here we are going to bound on average instead of high probability.

how much can I overfit to samples

$$\mathop{E}_{S\sim D^m}\sup_{h\in H} L_D(h) - L_S(h) = \mathop{E}_{S\sim D^m}\left(\sup_{h\in H} \mathop{E}_{S'\sim D^m}[L_{S'}(h) - L_S(h)]\right)$$

① $\sup_y \mathop{E}_x f_y(x) \le E_x \sup_y f_y(x)$

Pf: for any y, we have $f_y(x) \le \sup_{y'} f_{y'}(x)$

then $\mathop{E}_x f_y(x) \le \mathop{E}_x \sup_{y'} f_{y'}(x)$

and then $\sup_y \mathop{E}_x f_y(x) \le (\sup)\mathop{E}_x \sup_y f_y(x)$

$$\le \mathop{E}_{S,S'\sim D^m}\sup_{h\in H}[L_{S'}(h) - L_S(h)]$$

$$= \mathop{E}_{S,S'}\sup_h \frac{1}{m}\sum_{i=1}^m [\ell(h,z_i') - \ell(h,z_i)]$$

② Let $\sigma_i \in \{-1,1\}$ for $i\in[m]$,

$(u_i, u_i') = \begin{cases}(z_i, z_i') & \text{if } \sigma_i = 1\\ (z_i', z_i) & \text{if } \sigma_i = -1\end{cases}$

for any $\vec{\sigma} = (\sigma_1,\dots,\sigma_m)$

then $\ell(h,z_i') - \ell(h,z_i) = \sigma_i(\ell(h,z_i') - \ell(h,z_i))$

$\sigma_i \sim$ Unif ; $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$

$$= \mathop{E}_{S,S'}\mathop{E}_{\vec{\sigma}}\mathop{E}_{U,U'}\left[\sup_h \frac{1}{m}\sum_{i=1}^m \sigma_i(\ell(h,u_i') - \ell(h,u_i)) \mid S,S',\vec{\sigma}\right]$$

not doing anything (deterministic $UU'(SS'\vec{\sigma})$) condition

↓ switch order of expec $P(SS')\cdot P(\sigma)\, P(UU' \mid SS'\sigma)$

$$= \mathop{E}_{U,U'}\mathop{E}_{\vec{\sigma}}\mathop{E}_{S,S'}\left[\sup_h \frac{1}{m}\sum_{i=1}^m \sigma_i(\ell(h,u_i') - \ell(h,u_i)) \mid U,U',\vec{\sigma}\right]$$

$P(UU')\cdot P(\sigma)\, P(SS' \mid UU'\sigma)$

deterministic we can drop S

$U, U' | S, S', \vec{\sigma}$ is determinated
but $U, U' \sim D^m$, $U, U', \vec{\sigma}$ are independent
$S, S' | U, U', \vec{\sigma}$ is determinated

$$= \mathop{\mathbb{E}}_{U,U'} \mathop{\mathbb{E}}_{\sigma} \left[\sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i (\ell(h, u_i) - \ell(h, u_i')) \right]$$

$U, U' \sim D^m$ rename $U \to S$

$$= \mathop{\mathbb{E}}_{S,S' \sim D^m} \mathop{\mathbb{E}}_{\sigma \sim \mathrm{Rad}^m} \left[\sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i (\ell(h, z_i) - \ell(h, z_i')) \right]$$

$\sup [\sigma_i (\ell(h, z_i) + (-\sigma_i) \ell(h, z_i')]$

$$\le \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\sigma} \sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i \ell(h, z_i) + \mathop{\mathbb{E}}_{S'} \mathop{\mathbb{E}}_{\sigma} \sup_h \frac{1}{m} \sum_{i=1}^m (-\sigma_i) \ell(h, z_i')$$

↖ same ↗

$$= 2 \mathop{\mathbb{E}}_{S,\sigma} \sup_h \frac{1}{m} \sum_{i=1}^m \sigma_i \ell(h, z_i)$$

③

$$= 2 \mathop{\mathbb{E}}_{S} \mathrm{Rad}((\ell \circ \mathcal{H})|_S)$$

③

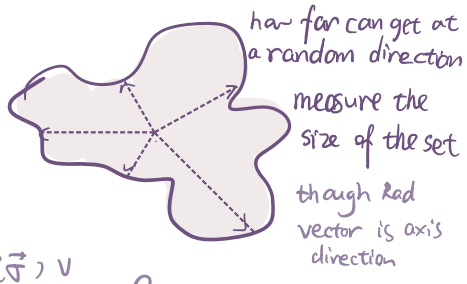
def $\ell \circ \mathcal{H} = \{ z \mapsto \ell(h, z) : h \in \mathcal{H} \}$

$\mathcal{F}|_S = \{ (f(z_1), f(z_2), \cdots, f(z_m)) : f \in \mathcal{F} \} \subseteq \mathbb{R}^m$ per-sample risk

$\mathrm{Rad}(V) = \mathop{\mathbb{E}}_{\vec{\sigma} \sim \mathrm{Rad}^m} \sup_{v \in V} \frac{1}{m} \sum_{i=1}^m \sigma_i v_i$

$\vec{\sigma} \cdot V$

the biggest product one can get given a Rademacher vector



(i) suppose one-element hypothesis :

$$\mathrm{Rad}(\{v\}) = \mathop{\mathbb{E}}_{\vec{\sigma}} \sup_{v' \in \{v\}} \frac{\vec{\sigma} \cdot v}{m} = \mathop{\mathbb{E}}_{\vec{\sigma}} \frac{\vec{\sigma} \cdot v}{m} = \frac{(\mathop{\mathbb{E}} \vec{\sigma}) \cdot v}{m} = 0$$

(ii) suppose $\mathcal{H}$ is bigger :

$$\mathrm{Rad}(\{-1, 1\}^m) = \mathop{\mathbb{E}}_{\sigma} \sup_{v \in \{-1,1\}^m} \frac{\sigma \cdot v}{m} = 1$$

$$\mathrm{Rad}([-1, 1]^m) = 1$$

$$\mathrm{Rad}(\{v : \|v\| \le \sqrt{m}\}) = 1$$

$\mathcal{H}$ bigger, but Rademacher is equal

because Rad vector only looks at $\{\pm 1\}^m$
Gaussian complexity would look at any direction

property :

(i) $\mathrm{Rad}(cV) = \mathrm{Rad}(\{cv : v \in V\}) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \sigma \cdot (cv) = \frac{|c|}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v \in V} \mathrm{sign}(c) \vec{\sigma} \cdot V$

$c \in \mathbb{R}$

$$= |c| \mathrm{Rad}(V)$$

(ii) $\mathrm{Rad}(V+W) = \mathrm{Rad}(\{v+w : v \in V, w \in W\}) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_{v,w} \sigma(v+w) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma} \sup_V \sigma v + \sup_W \sigma w$

$$= \mathrm{Rad}(V) + \mathrm{Rad}(W)$$

if w is constant vector $\Rightarrow \mathrm{Rad}(w) = 0$

How to compute $Rad(\ell \circ H|_S)$ practically?

- Telegrand's contraction lemma : (1-Lipschitz function)

Let $\phi : \mathbb{R}^m \to \mathbb{R}^m$ be given by $\phi(t) = (\varphi_1(t_1), \cdots, \varphi_m(t_m))$, where each φ_i is ρ-Lip

then

$$Rad(\phi \circ V) = Rad(\{\phi(v) : v \in V\}) \le \rho\, Rad(V)$$

(ρ-Lipschitz)

for linear regression, ℓ_Y is 1-Lipschitz

$$Rad(\ell_Y \circ H|_S) = Rad(\{\ell_{y_1}(h(x_1)), \ell_{y_2}(h(x_2)), \cdots, \ell_{y_m}(h(x_m))\})$$
$$\le Rad(H|_S)$$

$$Rad(H_B) = \mathbb{E}_\sigma \sup_{\|w\| \le B} \frac{1}{m} \sum_{i=1}^m \sigma_i \langle w, x_i \rangle$$

$$= \mathbb{E}_\sigma \sup_{\|w\| \le B} \frac{1}{m} \langle w, \sum_{i=1}^m \sigma_i x_i \rangle$$

$$\le \frac{1}{m} \mathbb{E}_\sigma B \left\| \sum_{i=1}^m \sigma_i x_i \right\|$$

$$\le \frac{1}{m} B \sqrt{\mathbb{E}_\sigma \left\| \sum_i \sigma_i x_i \right\|^2}$$ ← to get rid of σ

① $\|\sum_i a_i\|^2 = \langle \sum_i a_i, \sum_i a_i \rangle = \sum_{ij} \langle a_i, a_j \rangle$

$$\overset{①}{=} \frac{B}{m} \sqrt{\mathbb{E}_\sigma \sum_{ij} \langle \sigma_i x_i, \sigma_j x_j \rangle} = \frac{B}{m} \sqrt{\mathbb{E}_\sigma \sum_{ij} \sigma_i \sigma_j x_i x_j}$$

$$= \frac{B}{m} \sqrt{\sum_i \mathbb{E}_{\sigma_i} \sigma_i^2 \|x_i\|^2 + \sum_{i \ne j} \mathbb{E}_{\sigma_i \sigma_j} \langle x_i, x_j \rangle}$$

$\mathbb{E}_{\sigma_i} \sigma_i^2 = 1$; $\mathbb{E}_{\sigma_i \sigma_j} = 0$ (independent $= \mathbb{E}\sigma_i \mathbb{E}\sigma_j = 0$)

$$= \frac{B}{m} \sqrt{\sum_i \|x_i\|^2}$$

$$\mathbb{E}_S Rad(H_S) \le \frac{B}{\sqrt{m}} \mathbb{E}_S \sqrt{\frac{1}{m} \sum_i \|x_i\|^2} \le \frac{B}{\sqrt{m}} \sqrt{\mathbb{E} \|x\|^2}$$

Lec 6 Rademacher Complexity (2)

$Rad(V) = \mathop{\mathbb{E}}_{\sigma \sim unif} \frac{\sigma \cdot v}{m}$ for $V \subseteq \mathbb{R}^m$

Lemma if $\varphi : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz, $Rad(\{\varphi(v_1), v_2, \cdots, v_m\}) \le Rad(V)$

① use Lemma 2 to prove Telegrand's contraction Lemma

Pf: $Rad(\{(\frac{1}{\rho}\varphi_1(v_1), v_2, \cdots, v_m) : v \in V\}) \le Rad(V)$ ← ρ-Lipschitz

$=$ $Rad(\{(v_2, \frac{1}{\rho}\varphi_1(v_1), \cdots, v_m) : v \in V\})$ Rotate doesn't change Rad

$=$ $Rad(\{(v_2, v_3, \cdots, \frac{1}{\rho}\varphi_1(v_1)) : v \in V\})$

the same for $\varphi_1, \varphi_2 \cdots$

$Rad(\{(\frac{1}{\rho}\varphi_1(v_1), \frac{1}{\rho}\varphi_2(v_2), \cdots, v_m) : v \in V\}) \le Rad(V)$

$=$ $Rad(\{(\frac{1}{\rho}\varphi_2(v_2), \frac{1}{\rho}\varphi(v_1), \cdots, v_m) : v \in V\})$

$\therefore Rad(\frac{1}{\rho}\phi \circ V) \le Rad(V)$

$\therefore Rad(\varphi \circ V) \le \rho\, Rad(V)$

② prove this Lemma 2

$m \cdot Rad(\{(\varphi(v_1), v_2, \cdots, v_m) : v \in V\})$

$= \mathop{\mathbb{E}}_{\vec{\sigma} \sim Rad^m} \sup_{v \in V} \sigma_1 \varphi(v_1) + \underbrace{\sigma_2 v_2 + \cdots + \sigma_m \cdot v_m}_{\vec{\sigma}_2 \cdot v_2}$

$= \frac{1}{2} \mathop{\mathbb{E}}_{\vec{\sigma}_2 \sim Rad^m} \sup_{v \in V} [\varphi(v_1) + \vec{\sigma}_2 \cdot v_2] + \frac{1}{2} \mathop{\mathbb{E}}_{\vec{\sigma}_2 \sim Rad^m} \sup_{v \in V} [-\varphi(v_1') + \vec{\sigma}_2 \cdot v_2']$ ← two independent parts

switch v_1, v_1' won't change

$= \frac{1}{2} \mathop{\mathbb{E}}_{\vec{\sigma}_2} \sup_{v, v' \in V} [|\varphi(v_1) - \varphi(v_1')| + \vec{\sigma}_2 \cdot (v_2 + v_2')]$

switch v, v'

$\le \frac{1}{2} \mathop{\mathbb{E}}_{\vec{\sigma}_2} \sup_{v, v' \in V} [|v_1 - v_1'| + \vec{\sigma}_2 \cdot (v_2 + v_2')]$ 1-Lipschitz

$= \frac{1}{2} \mathop{\mathbb{E}}_{\vec{\sigma}_2} \sup_{v, v' \in V} [v_1 - v_1' + \vec{\sigma}_2 \cdot (v_2 + v_2')]$

$= \frac{1}{2} \mathop{\mathbb{E}}_{\vec{\sigma}_2} \sup_{v \in V} [v_1 + \vec{\sigma}_2 \cdot v_2] + \sup_{v' \in V} [v_1' + \vec{\sigma}_2 \cdot v_2'] = \mathop{\mathbb{E}}_{\vec{\sigma}} \sup_{v} \vec{\sigma} \cdot v = m\, Rad(V)$

Then for typical supervised learning losses

$$(\ell \circ \mathcal{H})|_S = \{(\ell(h,z_1), \dots, \ell(h,z_m)) : h \in \mathcal{H}\}$$
$$= \{(l_{y_1}(h(x_1)), \dots, l_{y_m}(h(x_m))) : h \in \mathcal{H}\}$$
$$= (l_{S_y} \circ \mathcal{H})|_{S_x}$$

might depend on S_y

if l_{y_i} (the loss func of a prediction for y_i) are all ρ-Lipschitz.

$S_x = (x_1, \dots, x_m)$

then Talagrand's lemma gives:

$$\mathrm{Rad}(\ell \circ \mathcal{H})|_S) \le \rho\, \mathrm{Rad}(\mathcal{H}|_{S_x})$$

only holds for real-valued hypothesis ($\mathcal{H}$ don't need to be function, can be param of Gaussian...)

- consider logistic regression $\mathcal{H}_B = \{x \mapsto w \cdot x : \|w\| \le B\}$.

$$\mathrm{Rad}((\ell_{\log} \circ \mathcal{H}_B)|_S) \le \mathrm{Rad}(\mathcal{H}_B|_{S_x})$$

Then $m\,\mathrm{Rad}(\mathcal{H}_B|_{S_x}) = \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\| \le B} \sum_i \sigma_i \cdot \langle w, x_i \rangle$

$$= \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\| \le B} \langle w, \textstyle\sum_i \sigma_i x_i \rangle$$

Cauchy-Shwartz

$$\le \mathop{\mathbb{E}}_{\sigma} \sup_{\|w\| \le B} \|w\| \cdot \|\textstyle\sum_i \sigma_i x_i\|$$
$$= B \mathop{\mathbb{E}}_{\sigma} \|\textstyle\sum_i \sigma_i x_i\|$$

$(\mathbb{E}T)^2 \le \mathbb{E}T^2$

$$\le B \sqrt{\mathop{\mathbb{E}}_{\sigma} \|\textstyle\sum_i \sigma_i x_i\|^2}$$
$$= B \sqrt{\mathop{\mathbb{E}}_{\sigma} \textstyle\sum_{ij} \sigma_i \sigma_j \langle x_i, x_j \rangle}$$
$$= B \sqrt{\textstyle\sum_i \mathop{\mathbb{E}}_{\sigma_i} \sigma_i^2 \|x_i\|^2 + \sum_{i \ne j} \mathop{\mathbb{E}}_{\sigma_i \sigma_j} [\sigma_i \sigma_j] \langle x_i, x_j \rangle}$$

($\mathbb{E}_{\sigma_i} \sigma_i^2 = 1$; $\mathbb{E}_{\sigma_i \sigma_j} [\sigma_i \sigma_j] = 0$)

$$= B \sqrt{\textstyle\sum_i \|x_i\|^2}$$

$$\mathrm{Rad}(\mathcal{H}_B|_{S_x}) \le \frac{B}{\sqrt{m}} \sqrt{\frac{1}{m} \textstyle\sum_i \|x_i\|^2}, \quad \mathop{\mathbb{E}}_{S_x} \mathrm{Rad}(\mathcal{H}|_{S_x}) \le \frac{B}{\sqrt{m}} \sqrt{\mathbb{E}\|x\|^2}$$

It is fine in some case to only look the average

If $\Pr(\|x\| \le C) = 1$ (hard bound), $\mathrm{Rad}(\mathcal{H}_B|_{S_x}) \le \frac{BC}{\sqrt{m}}$

use Randermacher to get a high probability bound

=> McDiarmid's inequality (concentration inequality)

McDiarmid's inequality: Let $X_1, \cdots, X_m$ be independent, let $f(X_1, \cdots, X_m)$ be a real-valued function satisfying bounded differences

$$\forall i \in [m] \quad \sup_{x_1, \cdots, x_m, x_i'} |f(x_1, \cdots, x_m) - f(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_m)| \le c_i$$

change any variable at any one index doesn't change result too much

Then with probability at least $1-\delta$,

$$f(X_1, \cdots, X_m) \le \mathbb{E} f(X_1, \cdots, X_m) + \sqrt{\frac{1}{2}(\sum_{i=1}^m c_i^2) \log \frac{1}{\delta}}$$

special case: $f(x) = \frac{1}{m}\sum_i x_i$, each $x_i \in [a,b]$

then $\forall i \in [m]$ $\frac{1}{m} \sup_{x_1, x_m, x_i'} |(x_1 + x_2 + \cdots + x_m) - (x_1 + x_2 + \cdots + x_k' + \cdots + x_m)| = \frac{1}{m} \sup_{x_k, x_k'} |x_k - x_k'|$

$= \frac{1}{m}(b-a)$

so $c_i = \frac{b-a}{m}$, $\frac{1}{m}\sum_i x_i - \mathbb{E} x_i \le \sqrt{\frac{m}{2}(\frac{b-a}{m})^2 \log \frac{1}{\delta}} = (b-a)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}$

(McDiarmid's inequality implies hoeffding)

in case of loss $\ell(h,z) \in [a,b]$ for all h, z, then with prob $\ge 1-\delta$

$$\sup_{h \in H} L_D(h) - L_S(h) \le \mathbb{E} \sup_{h \in H} L_D(h) - L_S(h) + (b-a)\sqrt{\frac{1}{2m} \log \frac{1}{\delta}} \quad (1)$$

if $\hat{h}_S$ is an ERM, with prob $\ge 1-\delta$ that

$$L_D(\hat{h}_S) - L_D(h^*) \le \mathbb{E} \sup_{h \in H} [L_D(h) - L_S(h)] + (b-a)\sqrt{\frac{2}{m} \log \frac{2}{\delta}} \quad (2)$$

Pf. prove (1): Let $S^{(i)} = (z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_m)$. we have

$$L_D(h) - L_S(h) = L_D(h) - L_{S^{(i)}}(h) + L_{S^{(i)}}(h) - L_S(h)$$

$$\sup_h L_D(h) - L_S(h) \le \sup_h L_D(h) - L_{S^{(i)}}(h) + \sup_h L_{S^{(i)}}(h) - L_S(h)$$

$$|\sup_h [L_D(h) - L_S(h)] - \sup_h [L_D(h) - L_{S^{(i)}}(h)]| \le |\sup_h L_{S^{(i)}}(h) - L_S(h)|$$ $S, S^{(i)}$ changable

$$|\sup_h [L_D(h) - L_S(h)] - \sup_h [L_D(h) - L_{S^{(i)}}(h)]| \le \sup_h |L_{S^{(i)}}(h) - L_S(h)| = \frac{b-a}{m}$$

apply McDiarmid's inequality ↓

with prob at least of $1-\delta$:

$$\sup_h L_D(h) - L_S(h) \le \mathbb{E}_S \sup_h L_D(h) - L_S(h) + (b-a)\sqrt{\frac{1}{2m} \log \frac{1}{\delta}}$$

$\neq \sup_h \mathbb{E}(L_D(h) - L_S(h))$ $\mathbb{E}$ and max does not commute!

Pf for McDiarmid — very general commonly used.

$X_{1:k-1} = (X_1, \dots, X_{k-1})$. Fix some $k \in [m]$, freeze some arbitrary values $x_{1:k-1}$

$\mathbb{E}_{X_{k+1:m}} f(x_{1:k-1}, X_k, X_{k+1:m})$ is random, depend on X_k

note that $\sup_{x_k} f(x_{1:m}) - \inf_{x_k} f(x_{1:m}) \le c_k$ by assumption

$$\Rightarrow \mathbb{E}_{X_{k+1:m}} \sup_{x_k} f(x_{1:k-1}, x_k, X_{k+1:m}) + \sup_{x_k}\left(-f(x_{1:k-1}, x_k, X_{k+1:m})\right) \le c_k \qquad \sup \mathbb{E} \le \mathbb{E} \sup$$

$$\sup_{x_k} \mathbb{E}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}) + \sup_{x_k} \mathbb{E}_{X_{k+1:m}} (-f(x_{1:k-1}, x_k, X_{k+1:m})) \le c_k$$

$$\sup_{x_k} \mathbb{E}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}) - \inf_{x_k} \mathbb{E}_{X_{k+1:m}} f(x_{1:k-1}, x_k, X_{k+1:m}) \le c_k$$

$X \sim SG(\sigma)$, $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\frac{1}{2}\lambda^2\sigma^2}$

by Hoeffding's Lemma $\mathbb{E}_{X_{k:m}}$ is $SG(c_k/2)$

then $\mathbb{E}_{X_k} \exp\left(\lambda \mathbb{E}_{X_{k+1:m}} f(x_{1:k-1}, X_k, X_{k+1:m})\right) \le \exp\left(\lambda \mathbb{E}_{X_k X_{k+1:m}} f(x_{1:k-1}, X_k, X_{k+1:m}) + \frac{1}{2}\lambda^2 \left(\frac{c_k}{2}\right)^2\right)$

holds for any $x_{1:k-1}$, then average on $x_{1:k-1}$

$$\mathbb{E}_{X_{1:k}} \exp\left(\lambda \mathbb{E}_{X_{k+1:m}} f(x_{1:m})\right) \le \mathbb{E}_{X_{1:k-1}} \exp\left(\lambda \mathbb{E}_{X_{k:m}} f(x_{1:m}) + \frac{1}{8}\lambda^2 c_k^2\right)$$

take log

$$\log \mathbb{E}_{X_{1:k}} \exp\left(\lambda \mathbb{E}_{X_{k+1:m}} f(x_{1:m})\right) \le \log \mathbb{E}_{X_{1:k-1}} \exp\left(\lambda \mathbb{E}_{X_{k:m}} f(x_{1:m})\right) + \frac{1}{8}\lambda^2 c_k^2$$

a_k $\qquad$ a_{k-1}

sum over $k = 1, \dots, m$ $\qquad a_m \le a_1 + \sum \frac{1}{8}\lambda^2 c_k^2$

$$\log \mathbb{E}_{X_{1:m}} \exp(\lambda f(x_{1:m})) \le \log \exp\left(\lambda \mathbb{E}_{X_{1:m}} f(x_{1:m})\right) + \sum_{k=1}^m \frac{1}{8}\lambda^2 c_k^2$$

$$\mathbb{E}_{X_{1:m}} \exp(\lambda f(x_{1:m})) \le \exp\left(\lambda \mathbb{E}_{X_{1:m}} f(x_{1:m})\right) \cdot \exp\left(\sum_{k=1}^m \frac{1}{8}\lambda^2 c_k^2\right)$$

$$\mathbb{E}_{X_{1:m}} \exp\left(\lambda f(x_{1:m}) - \lambda \mathbb{E}_{X_{1:m}} f(x_{1:m})\right) \le \exp\left(\frac{1}{2}\lambda^2 \sum_{k=1}^m \left(\frac{c_k}{2}\right)^2\right)$$

$f(x_{1:m}) \in SG\left(\frac{1}{2}\sqrt{\sum_{k=1}^m c_k^2}\right)$

with chernoff bound for sub-Gaussians

with prob at least $1-\delta$ $\quad f(x_{1:m}) \le \mathbb{E} f(x_{1:m}) + \frac{1}{2}\sqrt{\sum_{i=1}^m c_i^2} \cdot \sqrt{2 \log \frac{1}{\delta}}$

Rad is scale-sensitive: depend on the choice of function

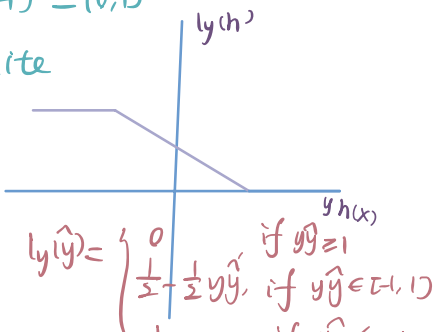
Lec 6 Growth Function and VC Dimension

How to bound Rad for binary classifier?

For binary $\mathcal{H}|_S = \{(h(x_1), h(x_2), \dots, h(x_m)) : h \in \mathcal{H}\} \subseteq \{0,1\}^m$

At most 2^m possible vectors even if $\mathcal{H}$ is finite

- $L_{0-1}(h,(x,y)) = \mathbb{1}(h(x) \neq y)$



$\|L_y\|_{Lip} = \frac{1}{2}$

$\mathrm{Rad}((l_{0-1} \circ \mathcal{H}_{\pm 1})|_S) \leq \frac{1}{2}\mathrm{Rad}(\mathcal{H}_{\pm 1}|_S)$

$\mathcal{H}|_S$ finite

- If $|V| < \infty$, $\|v\| \leq B$ for all $v \in V$

  Then $\mathrm{Rad}(V) \leq \frac{B}{m}\sqrt{2\log|V|}$

special case: if $V = \mathcal{H}_{\pm 1}|_S$, $\|V\| = \sqrt{1+1+\dots+1} = \sqrt{m}$

$\mathrm{Rad}(\mathcal{H}_{\pm 1}|_S) \leq \sqrt{\frac{2}{m}\log|\mathcal{H}_{\pm 1}|_S|}$

Pf $\mathrm{Rad}(V) = \mathbb{E}_\sigma \sup_{v \in V} \sum_{i=1}^m \frac{\sigma_i v_i}{m}$

$\sigma_i \in SG(1)$

$\frac{\sigma_i v_i}{m} \in SG(\frac{|v_i|}{m}) \Rightarrow \sum_i \frac{\sigma_i v_i}{m} \in SG(\sqrt{\sum_i \frac{v_i^2}{m^2}})$ independent

$= SG(\frac{1}{m}\|V\|) \leq SG(\frac{B}{m})$

- $T_i \in SG(\sigma)$, $\mathbb{E}T_i = 0$, T_i can be dependent

  then $\mathbb{E}\max_{i \in [m]} T_i \leq \sigma\sqrt{2\log m}$

  Pf. see in A2 Q(2,4) Jensen's Inequality: $\exp(\mathbb{E}Y) \leq \mathbb{E}\exp(Y)$

when taking sup $\vec{\sigma}v$ can be dependent, so $\mathbb{E}_\sigma \sup_{v \in V} \sum_{i=1}^m \frac{\sigma_i v_i}{m} \leq \frac{B}{m}\sqrt{2\log|V|}$ ($\sum_{i=1}^m \frac{\sigma_i v_i}{m}$ = T_i; $|V|$ = $|\mathcal{H}|_S|$)

$\mathrm{Rad}(\mathcal{H}|_S)$ depends on particular distribution

we can use $|\mathcal{H}|_S| \leq |\mathcal{H}|$ but it's very loose!

use growth function to drop dependence on particular S_x

- The growth function of $\mathcal{H}$ is

$\Pi_{\mathcal{H}}(m) = \sup_{x_1,\dots,x_m \in X} |\mathcal{H}|_{(x_1,\dots,x_m)}|$, $|\mathcal{H}_{S_x}| \leq \Pi_{\mathcal{H}}(m)$

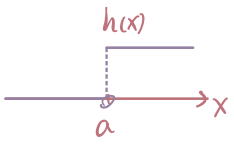
- (shatter): if $\mathcal{H}$ shatter a set $S_x \subseteq X$, means $\mathcal{H}$ can assign any possible label to the set i.e. $|\mathcal{H}|_{S_x}| = 2^m$

if the theory could explain any outcome then it's too general

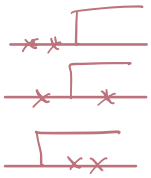
VC is defined on the size of set you can shatter.

• (VC dimension) $VCdim(\mathcal{H}) = \max\{m \geq 0 : \Gamma_{\mathcal{H}}(m) = 2^m\}$ worst case: there is A set that can be shattered

example: threshold $h_a(x) = \mathbb{1}(x \geq a)$ $VCdim = 1$



can't shatter two points



◦ $\mathcal{H} = \{x \mapsto sgn(w \cdot x) : w \in \mathbb{R}^d\}$ want to show $VCdim(\mathcal{H}) = d$

$$sgn(t) = \begin{cases} 1 & \text{if } t \geq 0 \\ -1 & \text{if } t < 0 \end{cases}$$

① $\mathcal{H}$ can shatter $\{e_1, e_2, \cdots, e_d\}$ to get $(y_1, \cdots, y_d)$

use $w = (y_1, \cdots, y_d)$, $w \cdot e_j = y_j$

(assume) ② Let $x_1, \cdots, x_{d+1}$ be shatterable by $\mathcal{H}$ (can't be linear dependent)

$\therefore \exists \alpha \in \mathbb{R}^{d+1} \setminus \{0\}$, s.t. $\sum_i^{d+1} \alpha_i x_i = 0$

Let $I_+ = \{i \in [d+1] : \alpha_i > 0\}$, $I_0 = \{i \in [d+1] : \alpha_i = 0\}$, $I_- = \{i \in [d+1] : \alpha_i < 0\}$

$\exists w$ s.t. $h_w(x_j) = \begin{cases} 1 & \text{if } j \in I_+ \\ -1 & \text{if } j \in I_- \end{cases}$

$$0 = w \cdot 0 = w \sum_{i=1}^{d+1} \alpha_i x_i = w \sum_{i \in I_+} \underbrace{\alpha_i}_{>0} x_i + w \sum_{i \in I_-} \underbrace{\alpha_i}_{<0} x_i + w \sum_{i \in I_0} \underbrace{\alpha_i}_{=0} x_i$$

$$= \sum_{i \in I_+} \underbrace{\alpha_i}_{>0} \underbrace{w x_i}_{\geq 0} + \sum_{i \in I_-} \underbrace{\alpha_i}_{<0} \underbrace{w x_i}_{<0} \geq 0$$ 取等 iff $\begin{cases} wx_i = 0 \text{ for } \forall i \in I_+ \\ I_- = \{\} \end{cases}$

we can also find some $\tilde{w}$ that $\tilde{w} \cdot x_i < 0$ for $\forall i \in I_+$

$$0 = \tilde{w} \cdot 0 = \tilde{w} \cdot \sum_{i \in I_+} \alpha_i x_i = \sum_{i \in I_+} \underbrace{\alpha_i}_{>0} \underbrace{\tilde{w} x_i}_{<0} < 0 \Rightarrow \text{contradiction!}$$

◦ $VCdim(\{x \mapsto w \cdot x + b : w \in \mathbb{R}^d, b \in \mathbb{R}\}) = d+1$

• $\Gamma_{\mathcal{H}}(m) = O(m^{VCdim(\mathcal{H})})$.

• Binary classifier with $VCdim(\mathcal{H}) = d$, zero-one loss for any $m > d$:

$$\mathbb{E}_{S_x} \sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \leq \mathbb{E}_{S_x} \sqrt{\frac{2}{m} \log|\mathcal{H}_{|S_x}|} \leq \sqrt{\frac{2}{m} \log \Gamma_{\mathcal{H}}(m)} \leq \sqrt{\frac{2d}{m}[\log m + 1 + \log d]}$$

Corollary, If $m \geq d := VCdim(\mathcal{H})$, then $\Gamma_{\mathcal{H}}(m) \leq (\frac{em}{d})^d$ $(\log \Gamma_{\mathcal{H}}(m) \leq d(\log \frac{m}{d} + 1))$

Sauer-Shelah Lemma ①: If $d = VCdim(\mathcal{H}) < \infty$, then $\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$

(we do not prove the lemma)

use S-S lemma to prove corollary, want to show $\sum_{i=0}^{d} \binom{m}{i} \le \left(\frac{em}{d}\right)^d$ for $m \ge d$

pf for corollary

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

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

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

$$= \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \quad (1+\tfrac{x}{n})^n \le e^x$$

$$\le \left(\frac{m}{d}\right)^d e^d$$

Lemma ②: For all finite $S_X \subseteq X$, $|\mathcal{H}|_{S_X}| \le |\{T \subseteq S_X : T \text{ is shattered by } \mathcal{H}\}|$

Pf for ① use ② $|\mathcal{H}|_{S_X}|$ is upper bounded by the number of subsets of S of size at most d.

Pf for ② inductive

These are equivlant for binary class, 0-1 loss:

- $Pr(\sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \le \varepsilon) \ge 1-\delta$ for $m \ge m(\varepsilon, \delta)$
- ERM agnostically PAC-learn $\mathcal{H}$
- $\mathcal{H}$ is ag PAC-learnable
- ERM PAC-learns $\mathcal{H}$
- $\mathcal{H}$ is PAC-learnable
- $VCdim(\mathcal{H}) < \infty$ implies uniform convergence.

Theorem. $\mathcal{H}$ is a binary classifer on X. $d = VCdim(\mathcal{H})$ $|X| = 2m$

Assume $m \le \frac{d}{2}$. then $\inf_A \sup_{\substack{D \\ \text{realizable} \\ \text{by } \mathcal{H}}} Pr_{S \sim D^m}\left(L_D(A(S)) \ge \frac{1}{8}\right) \ge \frac{1}{7}$

for a best-case learning algorithm there is a worst-case distribution

$$\mathop{\mathbb{E}}_{f \sim Unif(\hat{X} \to Y)} \mathop{\mathbb{E}}_{S \sim D^m(f)} L_{D(f)}(\hat{h}_S) = \mathop{\mathbb{E}}_{f, S_X} \mathop{\mathbb{E}}_{x \sim D_X} \mathbb{1}(\hat{h}_S(x) \ne f(x))$$

not in the training set

$$\mathop{\mathbb{E}}_{f, S_X, x} \mathbb{1}(\hat{h}_S(x) \ne f(x)) = \mathop{\mathbb{E}}_{f, S_X} [Pr(x \notin S_X) \mathop{\mathbb{E}}_{x \sim D_X} [\mathbb{1}(\hat{h}_S(x) \ne f(x)) \mid x \notin S_X]$$

$$+ [Pr(x \in S_X) \mathop{\mathbb{E}}_{x \sim D_X} [\mathbb{1}(\hat{h}_S(x) \ne f(x)) \mid x \in S_X]$$

$$P(x \notin S_x) = \frac{|\hat{X} \setminus S_x|}{|\hat{X}|} \geq \frac{m}{2m} = \frac{1}{2}$$ ↙ $S_x = ((x_1, f(x_1)), \cdots, (x_m, f(x_m)))$

$$\geq \frac{1}{2} \mathop{\mathbb{E}}_{f, S_x, x} [\mathbb{1}(\hat{h}_S(x) \neq h(x) \mid x \notin S_x]$$

$$= \frac{1}{4}$$

average over f $\geq \frac{1}{4}$, so $\exists g \in \mathcal{T}$. $\mathop{\mathbb{E}}_{S \sim D_g^m} L_{D_g}(\hat{h}_S) \geq \frac{1}{4}$

Want to show $\Pr_{S \sim D_g^m}(L_{D_g}(\hat{h}_S) < \frac{1}{8}) \leq \frac{6}{7}$

Markov : $\Pr(1 - L_{D(g)} \geq \frac{7}{8}) \leq \frac{\mathbb{E}(1 - L_{D(g)})}{\frac{7}{8}} \leq \frac{1 - \frac{1}{4}}{\frac{7}{8}} = \frac{6}{7}$

with no prior, all algorithms perform the same on average.

Let $\mathcal{H}$ be a set of binary classifer over $\mathcal{X}$. For any $m \geq 1$

$$\inf_A \sup_{D \text{ realizable by } \mathcal{H}} \Pr_{S \sim D^m} \left(L_D(A(S)) > \frac{VC\dim(\mathcal{H}) - 1}{32m} \right) \geq \frac{1}{100}$$

$\sim$ 0-1 loss

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

① $d = 1$, holds almost trivally

② $d \geq 2m$, this is a corollary of the above theorem

③ $2 \leq d < 2m$

What do we do for approximation error?

Lec Structural Risk Minimization — not uniform over all H_k

- $\mathcal{H} = \underset{w_1}{\mathcal{H}_1} \cup \underset{w_2}{\mathcal{H}_2} \cup \cdots$

weight: how much do I like $H_1, H_2 \ldots$ preference on each hypothesis

$\sum_{w_k > 0} w_k \le 1$

$\forall k \; \forall D. \; \Pr_{S\sim D^m}\left(\sup_{h\in\mathcal{H}_k} L_D(h) - L_S(h) \le \epsilon_k(m,\delta)\right) \ge 1-\delta$ with $\epsilon_k(m,\delta) \to 0$ as $m \to \infty$

$\therefore \Pr_{S\sim D^m}(\forall h\in\mathcal{H}.\; L_D(h) \le L_S(h) + \epsilon_{k_h}(m, \delta w_{k_h})) \ge 1-\delta$

$k_h := \underset{k\ge 1,\, h\in\mathcal{H}_k}{\operatorname{argmin}} \epsilon_k(m, \delta w_k)$ index that minimize ϵ_k corresponding to h

h lives in some of the hypo classes at least one H_k

① Def: $SRM_{\mathcal{H}}(S) \in \underset{h\in\mathcal{H}}{\operatorname{argmin}} [L_S(h) + \epsilon_{k_h}(m, \delta w_{k_h})]$ commit to a δ beforehand

higher w_k for simpler hypothesis, tighter for complex hypo

- SRM algorithm

We can implement this minimization by a finite number of calls to an "ERM oracle", as long as our loss is lower-bounded by $a \le \ell(h, z)$ (typically $a = 0$):

function $\mathrm{SRM}_{\mathcal{H}}(S)$
  best $\leftarrow \infty$
  for $k = 1, 2, \ldots$ **do**
    $h_k \leftarrow \mathrm{ERM}_{\mathcal{H}_k}(S)$
    cand_loss $\leftarrow L_S(h_k) + \epsilon_k(m, w_k\delta)$
    if cand < best **then**
      $\hat{h} \leftarrow h_k$
      best $\leftarrow$ cand
    if $\min_{k'>k} a + \epsilon_{k'}(m, w_{k'}\delta) >$ best **then**
      break
  return $\hat{h}$

wired because depend on δ

Note that if we "decompose" as $\mathcal{H}_1 = \mathcal{H}$, then SRM becomes just $\mathrm{ERM}_{\mathcal{H}}$.

- general bound of SRM

non-uniform learnability because m function depend on h

$\hat{h}_S := SRM_{\mathcal{H}}(S) \in \underset{h\in\mathcal{H}}{\operatorname{argmin}} L_S(h) + \epsilon_{k_h}(m, \delta w_{k_h})$

$L_D(\hat{h}_S) \le L_S(\hat{h}_S) + \epsilon_{k_{\hat{h}_S}}(m, \delta w_{k_{\hat{h}_S}})$ (def. prob $\ge 1-\delta$)

$\le L_S(h^*) + \epsilon_{k_{h^*}}(m, \delta w_{k_{h^*}})$

some h can be learned with less data

$= L_D(h^*) + L_S(h^*) - L_D(h^*) + \epsilon_{k_{h^*}}(m, \delta w_{k_{h^*}})$

$\le L_D(h^*) + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}} + \epsilon_{k_{h^*}}(m, \delta w_{k_{h^*}})$ (prob $\ge 1-\delta$)

$\therefore \Pr(L_D(\hat{h}_S) - L_D(h^*) \le \epsilon_{k_{h^*}}(m, \delta w_{k_{h^*}}) + (b-a)\sqrt{\frac{1}{2m}\log\frac{2}{\delta}}) \ge 1-\delta$

$h^* \in \mathcal{H}$ is any fixed hypothesis

- Specific bound use Randmerchar

$R_k = \mathbb{E}_{S\sim D^m} \mathrm{Rad}((\ell \circ H_k)|_S)$

$L_D(h) \le L_S(h) + 2R_{k_h} + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{w_{k_h}\delta}}$ but depend on δ

$w_k = \frac{6}{\pi^2 k^2}$ $\log\frac{1}{w_{k_h}\delta} = \log(k_h^2 \cdot \frac{\pi^2}{6}\cdot\frac{1}{\delta}) = 2\log k_h + \log\frac{\pi^2}{6} + \log\frac{1}{\delta}$

$< 2\log k_h + \frac{1}{2} + \log\frac{1}{\delta}$

$= 2\log k_h + \log\frac{\sqrt{e}}{\delta}$

$\sqrt{\log\frac{1}{w_{k_h}\delta}} < \sqrt{2\log k_h} + \sqrt{\log\frac{\sqrt{e}}{\delta}}$

$\Pr(\forall h\in H.\ L_D(h) \le L_S(h) + 2R_{k_h} + (b-a)\sqrt{\frac{1}{m}\log k_h} + (b-a)\sqrt{\frac{1}{2m}\log\frac{\sqrt{e}}{\delta}}) \ge 1-\delta$

② $\hat{h}_S = SRM(S) \in \arg\min L_S(h) + 2R_{k_h} + (b-a)\sqrt{\frac{1}{m}\log k_h}$ does not commit to a δ

does not depend on δ, better in practice

$L_D(\hat{h}_S) \le L_S(\hat{h}_S) + 2R_{k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{m}\log k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{\sqrt{e}}{\delta}}$ ($SRM\ \hat{h}_S$) with $p \ge 1-\delta$ ← hold for $\forall h$

then

$L_D(\hat{h}_S) \le L_S(\hat{h}_S) + 2R_{k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{m}\log k_{\hat{h}_S}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$ ← in order to make 右邊

(holds with probability $1-\frac{\sqrt{e}}{3}\delta$)

$\le L_S(h^*) + 2R_{k_{h^*}} + (b-a)\sqrt{\frac{1}{m}\log k_{h^*}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$ $(1-\frac{\sqrt{e}}{3}\delta)$ ← union

Hoeffding $L_S(h^*) \le L_D(h^*) + (b-a)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$ $(1-\frac{\delta}{3})$

$\le L_D(h^*) + 2R_{k_{h^*}} + (b-a)\sqrt{\frac{1}{m}\log k_{h^*}} + (b-a)\sqrt{\frac{2}{m}\log\frac{3}{\delta}}$ $(1-\frac{\sqrt{e}+1}{3}\delta \ge 1-\delta)$

Comparison: if we know the correct H_{k_h} from the start and run ERM

ERM: $L_D(ERM_{H_{k^*}}) \le L_D(h^*) + 2R_{k_{h^*}} + (b-a)\sqrt{\frac{2}{m}\log\frac{1}{\delta}}$

- varying bounded losses

e.g. logistic regression $H_k = \{x \mapsto w\cdot x : \|w\| \le B_k\}$ $\Pr(\|x\| \le C) = 1$

$h^*(x) = w^*\cdot x$

if choose $B_k = 2^k$ then $2^{k-1} < \|w\| \le 2^k \Rightarrow k_h < \log_2(2\|w\|) \Rightarrow \sqrt{\log k_h} < \sqrt{\log\log_2(2\|w\|)}$

$(b-a)k_h < (2C\|w\|+1)$, thus (not commit to a δ)

$\hat{h}_S = h_{\hat{w}_S}$; $\hat{w}_S \in \arg\min_{w\in R^d} L_S(h_w) + \frac{4C\|w\|}{\sqrt{m}} + (2C\|w\|+1)\sqrt{\frac{1}{m}\log(\log_2(2\|w\|))}$

$L_D(\hat{h}_S) \le L_D(h^*) + \frac{4C\|w^*\|}{\sqrt{m}} + (2C\|w^*\|+1)\sqrt{\frac{1}{m}\log(\log_2(2\|w\|))} + (2C\|w^*\| + 2C\|\hat{w}_S\| + 2)\sqrt{\frac{1}{2m}\log\frac{3}{\delta}}$

$\approx$

- relationship to regularization

$$\vec{w}_S \in \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; L_D(h^*) + \frac{\lambda}{\sqrt{m}} \|w^*\|$$

SRM $\sim$ regularization

- non-uniform learnability:

we have shown a bound

$$\Pr\left(L_D(SRM_{\mathcal{H},s}(S)) \le L_D(h^*) + \varepsilon_{k^*}(m, w_{k^*}\delta) + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}\right) \ge 1-2\delta$$

the bound does not show PAC learnability because ε depends on h^* !!

Def (competes): $A(S)$ (ε, δ)-competes with $h \in \mathcal{H}$ on D with m samples, $\Pr_{S\sim D^m}(L_D(A(S)) \le L_D(h) + \varepsilon) \ge 1-\delta$

Def (non-uniform learning): A nonuniformly learns $\mathcal{H}$ if $\exists$ finite $m(\varepsilon, \delta, h)$, s.t. $\forall \varepsilon, \delta \in (0,1)$, $\forall h \in \mathcal{H}$, $\forall D$, $\forall m \ge m(\varepsilon, \delta, h)$, $A(S)$ (ε, δ)-competes with h on D

this is looser than PAC because m depends on h

PAC: $\exists m \ge m(\varepsilon, \delta)$ $\Pr(L_D(A(S)) \le \inf_h (L_D(h) + \varepsilon)) \ge 1-\delta$ for $\forall \varepsilon, \delta \in (0,1)$

If $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots$ with $VCdim(\mathcal{H}_k) < \infty$

then SRM nonuniformly learns $\mathcal{H}$ (in 0-1 binary classification)

Pf Let $\mathcal{H}_k = \{h \in \mathcal{H} : m(\frac{1}{8}, \frac{1}{7}, h) \le k\}$

$\therefore \mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots$

consider D realizable by $\mathcal{H}_k$

then $\Pr_{S\sim D^k}(L_D(A(S)) \le \frac{1}{8}) \ge \frac{1}{7}$ using $h^* \in \mathcal{H}_k$ with $L_D(h^*) = 0$

polynomial example

$\mathcal{H} = \{x \mapsto \operatorname{sgn}(w \cdot x) : x \in \mathbb{R}^d\}$

$VCdim(\mathcal{H}) = d$

$$\sup_{h\in\mathcal{H}} L_D(h) - L_S(h) \le \sqrt{\frac{2d}{m}((\log m - 1) - \log d)} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$$

?

countable

Assume $\mathcal{H} = \{h_1, h_2, \cdots\}$ — divide — singleton classes

Use $\mathcal{H}_k = \{h_k\}$

Then $\epsilon_k(m, w_k\delta) \le (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{w_k\delta}}$

$\le (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{w_k}} + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$

how to assign weights?

how to decide the order?

· Minimum Description Length

prefix-free

choose a weight according to a binary language for determining $\mathcal{H}$

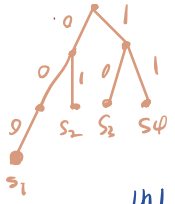
Kraft's inequality :

If $S \subseteq \{0,1\}^*$ (a set of binary strings) is prefix-free

(if '00' is avalid s, we don't have anything else that starts with '00')

(or read the data before '0' in C language)

then $\sum_{s \in S} 2^{-|s|} \le 1$ (probability distribution)



Then, we can choose a representation for $\mathcal{H}$ and assign $w_h = 2^{-|h|}$

$MDL_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta 2^{-|h|}}}$

$\sqrt{\frac{1}{2m}(\log(\frac{1}{\delta}) + |h|\log 2)}$

or $MDL_S \in \underset{h \in \mathcal{H}}{\operatorname{argmin}} L_S(h) + \sqrt{\frac{\log 2}{2}\frac{|h|}{m}}$

$\therefore L_D(MDL(S)) \le L_D(h^*) + (b-a)\left(\sqrt{\frac{\log 2}{2}\cdot\frac{|h^*|}{m}} + \sqrt{\frac{2}{m}\log\frac{2}{\delta}}\right)$

Occam's razor: if there are multiple explanations of the data ($L_S(h_1) = L_S(h_2)$)

prefer the simpler one (shortest explanation)

If we choose $|h|$ to be the length of the shortest possible implementation of h in some programming language, is known Kolmogorov Complexity

MAP inferance with a Kolmogorov prior

Margins Theory

we do ERM with 0-1 loss

VC theory

$\mathcal{H} = \{x \mapsto \text{sgn}(w \cdot x) : x \in \mathbb{R}^d\}$ $\quad VCdim(\mathcal{H}) = d$ known from previous lec

$\sup_{h \in \mathcal{H}} L_D(h) - L_S(h) \le \sqrt{\frac{2d}{m}(\log m + 1 - \log d)} + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}$ $\quad$ (prob $\ge 1-\delta$)

$L_D(\hat{h}_S) - \inf_{h \in \mathcal{H}} L_D(h) \le \sqrt{\frac{2d}{m}(\log m + 1 - \log d)} + \sqrt{\frac{2}{m}\log\frac{2}{\delta}}$ $\quad$ (prob $\ge 1-\delta$)

two problems:

① ERM with 0-1 loss is NP hard if D is not realizable by $\mathcal{H}$

② if d is really big, the bound doesn't tell anything until $\frac{m}{\log m} > 2d$

like in kernel methods, d is sometimes infite!

we like a better bound when in high dimension (big d) (hope it is not d-dependent)

Rademacher

If $\mathcal{H}_B = \{x \mapsto w \cdot x : \|w\| \le B\}$ then $\mathop{\mathbb{E}}_S \text{Rad}(\mathcal{H}_B|_S) \le \frac{B\sqrt{\mathbb{E}\|x\|^2}}{\sqrt{m}}$

but $\uparrow$ is continuous, logistic loss

$$\text{Rad}(\ell_{\text{logistic}} \circ \mathcal{H}_B|_S) \le 1 \cdot \text{Rad}(\mathcal{H}_B|_S)$$

$$\therefore L_D^{\text{logistic}}\left(\operatorname{argmin}_{h \in \mathcal{H}_B} L_D^{\text{logistic}}(h)\right) \le \frac{2B\sqrt{\mathbb{E}\|x\|^2}}{\sqrt{m}} + \left(B\sqrt{\mathbb{E}\|x\|^2} + 1\right)\sqrt{\frac{2}{m}\log\frac{2}{\delta}}$$

what if we want to bound on accuracy

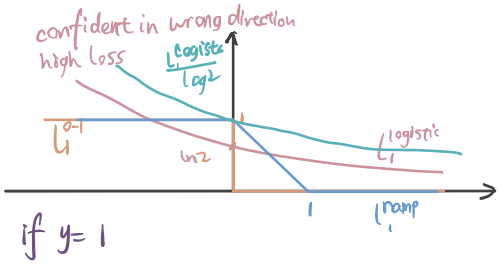
$\text{Rad}(\ell_{0\text{-}1} \circ \text{sgn} \circ \mathcal{H}_B)$

Let $\ell_{\text{surr}}(h, z) = \frac{1}{\log 2}\ell_{\text{logistic}}(h, z)$

$\forall h, z \quad \ell_{\text{surr}}(h, z) \ge \ell_{0\text{-}1}(h, z)$

$L_D^{\text{surr}}(h) \ge L_D^{0\text{-}1}(h)$

$$L_D^{0\text{-}1}(\text{sgn} \circ h) \le L_D^{\text{surr}}(h) \le L_S^{\text{surr}}(h) + 2\text{Rad}(\ell_{\text{surr}} \circ \mathcal{H}|_S) + (b-a)_{\text{surr}}\sqrt{\frac{2}{m}\log\frac{2}{\delta}}$$



if $y = 1$

0-1 loss $\quad \mathbb{1}[h \le 0]$

logistic loss (LR) $\quad \log(1 + \exp(-h))$

hinge loss (SVM) $\quad [1-h]_+$

ramp loss (s) $\quad [1-h]_+ - [s-h]_+$ $\quad (s=0)$