CLT Homework 1 20200130 Yujun Kim

#2.7

(a) The training pt learner recieves is generated from h_* and randomly changes with probability η.
By def. of $d(h^*)$, $d(h^*) = \eta$ (or deduced from (b) with $R(h^*)=0$)

(b) $d(h) = \mathbb{P}(h(x) = h^*(x))\eta + \mathbb{P}(h(x) \neq h^*(x))(1-\eta)$
$= (1-R(h))\eta + R(h)(1-\eta)$
$= \eta + (1-2\eta)R(h)$

(c) $d(h)$ is increasing with respect to $R(h)$, as $1-2\eta > 0$.
Hence, if $R(h) > \varepsilon$, then $d(h) > \eta + (1-2\eta)\varepsilon$.
Thus, $d(h) - d(h^*) = d(h) - \eta > (1-2\eta)\varepsilon > (1-2\eta')\varepsilon = \varepsilon'$.

(d) $\hat{d}(h^*) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h^*(x_i) \neq y_i}$.
Then, $\mathbb{E}_{S\sim D^m}[\hat{d}(h^*)] = d(h^*)$, and indicator functions are bounded on $[0,1]$.

By Hoeffding, $\mathbb{P}_{S\sim D^m}[\hat{d}(h^*) - d(h^*) \geq \frac{\varepsilon'}{2}] \leq e^{-2m(\frac{\varepsilon'}{2})^2}$
If $m \geq \frac{2}{\varepsilon'^2}\log\frac{2}{\delta}$, then $\mathbb{P}_{S\sim D^m}[\hat{d}(h^*) - d(h^*) \geq \frac{\varepsilon'}{2}] \leq \frac{\delta}{2}$.

Considering complement, done.

(e). Indeed, for any h, $\mathbb{P}[d(h) - \hat{d}(h) \geq \frac{\varepsilon'}{2}] \leq e^{-2m(\frac{\varepsilon'}{2})^2}$ (By Hoeffding)
$\Rightarrow \mathbb{P}[d(h) - \hat{d}(h) \geq \frac{\varepsilon'}{2}$ for some $h \in H] \leq |H| e^{-2m(\frac{\varepsilon'}{2})^2}$
If $m \geq \frac{2}{\varepsilon'^2}(\log|H| + \log\frac{2}{\delta})$, then $\mathbb{P}[d(h) - \hat{d}(h) \geq \frac{\varepsilon'}{2}$ for some $h \in H] \leq \frac{\delta}{2}$.
Thus, $\mathbb{P}[\forall h \in H, d(h) - \hat{d}(h) \leq \frac{\varepsilon'}{2}] \geq \mathbb{P}[\forall h \in H, d(h) - \hat{d}(h) < \frac{\varepsilon'}{2}]$
$\geq 1 - \frac{\delta}{2}$.

(f) $\hat{d}(h) - \hat{d}(h^*) = [\hat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)]$

By union bound, with $m \geq \max\{\frac{2}{\varepsilon'^2}\log\frac{2}{\delta}, \frac{2}{\varepsilon'^2}(\log|H| + \log\frac{2}{\delta})\}$
(↳ from (d), (e)) $= \frac{2}{\varepsilon'^2}(\log|H| + \log\frac{2}{\delta})$,

$$\mathbb{P}[\underbrace{\hat{d}(h^*) - d(h^*) \leq \frac{\varepsilon'}{2} \text{ and } \forall h \in H,\ d(h) - \hat{d}(h) \leq \frac{\varepsilon'}{2}}_{(*)}] \geq 1 - \delta.$$

If $(*)$, then for $h \in H$ with $R(h) > \varepsilon$,

$$\begin{aligned}\hat{d}(h) - \hat{d}(h^*) &\geq -\frac{\varepsilon'}{2} + d(h) - d(h^*) - \frac{\varepsilon'}{2} \\ &= -\varepsilon' + d(h) - d(h^*) \\ &\geq 0 \qquad \text{(by (c), as } \varepsilon' = \varepsilon(1-2\eta)).\end{aligned}$$

Thus, $\underbrace{\hat{d}(h) - \hat{d}(h^*) \geq 0 \ \forall h \in H \text{ with } R(h) > \varepsilon}_{(**)}$.

$(*)$ implies $(**)$.

$$\mathbb{P}[(**)] \geq \mathbb{P}[(*)] \geq 1 - \delta, \text{ given } m \geq \frac{2}{\varepsilon'^2}(\log|H| + \log\frac{2}{\delta}) = \frac{2}{\varepsilon^2(1-2\eta)^2}(\log|H| + \log\frac{2}{\delta}).$$

#2.10 $\mathcal{Z}$ is finite, $\mathcal{A}$ is PAC-learning algorithm.

① Define $\mathcal{D}$: Let $\mathcal{D} \sim \text{Unif}(\mathcal{Z})$. Then, polynomial p s.t. $k \geq p(\frac{1}{\varepsilon}, \frac{1}{\delta}, n)$ implies $\mathbb{P}_{S \sim \mathcal{D}^k}[R(h_S) \leq \varepsilon] \geq 1 - \delta$. Here, h_S is given by $\mathcal{A}$.

② Show consistency. Here, $R(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{h(x_i) \neq y_i}$, as $\mathcal{D}$ is $\text{Unif}(\mathcal{Z})$.

Note that $\mathcal{Z} = \{(x_i, y_i)\}_{i=1}^{m}$.

Take $\varepsilon = \frac{1}{m+1}$. Then, $R(h_S) \leq \varepsilon$ implies h_S being consistent to $\mathcal{Z}$.
($\because R(h_S) = \frac{j}{m}$ for $0 \leq j \leq m$, $R(h_S) \leq \frac{1}{m+1} \Rightarrow R(h_S) = 0$).

Thus, for $k \geq p(\frac{1}{\varepsilon} = m+1, \frac{1}{\delta}, n) = g(m, \frac{1}{\delta}, n)$ (g: another polynomial)

$$\mathbb{P}_{S \sim \mathcal{D}^k}[R(h_S) = 0] = \mathbb{P}_{S \sim \mathcal{D}^k}[R(h_S) \leq \varepsilon = \frac{1}{m+1}] \geq 1 - \delta.$$

Thus, using $k \geq g(m, \frac{1}{\delta}, n)$ samples with algorithm $\mathcal{A}$ suffices.

#2.12 (pf of bound)

$\mathcal{H} \subseteq [\mathcal{X} \to \mathcal{Y} = \{0,1\}]$

By corollary 2.10, $\forall h \in \mathcal{H}$,

$$\underset{S\sim D^m}{\mathbb{P}}\left[\hat{R}_S(h) - R(h) \le -\varepsilon\right] \le e^{-2m\varepsilon^2}.$$

Let $\varepsilon_h := \sqrt{\frac{\log\frac{1}{p(h)} + \log\frac{1}{\delta}}{2m}}$ Then,

$$\underset{S\sim D^m}{\mathbb{P}}\left[R(h) > \hat{R}_S(h) + \sqrt{\frac{\log\frac{1}{p(h)} + \log\frac{1}{\delta}}{2m}}\right] \le e^{-2m\varepsilon_h^2} = p(h)\,\delta.$$

union bound

$$\mathbb{P}\left[\exists h \in \mathcal{H},\ R(h) > \hat{R}_S(h) + \varepsilon_h\right] \le \sum_{h\in\mathcal{H}} p(h)\,\delta = \delta, \text{ as } \sum_{h\in\mathcal{H}} p(h) = 1.$$

Thus, $\mathbb{P}\left[\forall h \in \mathcal{H},\ R(h) \le \hat{R}_S(h) + \varepsilon_h\right] \ge 1-\delta$

(Comparison)

For inconsistent finite case, $\frac{1}{p(h)}$ was replaced by $|\mathcal{H}|$.
If $p(h) = \frac{1}{|\mathcal{H}|}$, i.e. prior for $\mathcal{H}$ is selected uniformly over $|\mathcal{H}|$, the two results matches.

If prior distribution is non uniform, then

$$\sup_{h\in\mathcal{H}} \varepsilon_h \ge \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2m}} =: \varepsilon_{\mathcal{H}}, \text{ as } \frac{1}{p(h)} > \frac{1}{|\mathcal{H}|} \text{ for some } h \in \mathcal{H}.$$

Hence, uniform ε-bound over $h \in \mathcal{H}$ gets looser,
but $\inf_{h\in\mathcal{H}} \varepsilon_h \le \varepsilon_{\mathcal{H}}$ so, it contains more component wise data on bounds over $h \in \mathcal{H}$.

The result in this problem is strict generalization of finite uniform case.

#3.11

$$\mathcal{X}=\mathbb{R}^{n_1}, \quad \mathcal{H}=\left\{x\mapsto \sum_{j=1}^{n_2} w_j\,\sigma(u_j\cdot x) : \|w\|_1\le\Lambda' ;\ \|u_j\|_2\le\Lambda\ \forall j\in[n_2]\right\}$$

(a)

$$\hat{R}_S(\mathcal{H})=\mathop{\mathbb{E}}_{\sigma_i\in\{\pm1\},\, i\in[m]}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right], \text{ for } S=(x_1,\dots,x_m).$$

$$=\mathop{\mathbb{E}}_{\sigma_i\in\{\pm1\}}\left[\sup_{\substack{\|w\|_1\le\Lambda',\\ \|u_j\|_2\le\Lambda}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\sum_{j=1}^{n_2}w_j\,\sigma(u_j\cdot x_i)\right]$$

$$=\mathop{\mathbb{E}}_{\sigma_i\in\{\pm1\}}\left[\sup_{\substack{\|w\|_1\le\Lambda',\\ \|u_j\|_2\le\Lambda}}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\sum_{j=1}^{n_2}w_j\,\sigma(u_j\cdot x_i)\right|\right], \text{ by symmetry over } \sigma_i$$

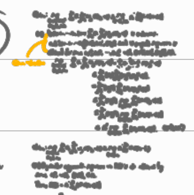
$$=\mathbb{E}\left[\sup\frac{1}{m}\left|\sum_{j=1}^{n_2}w_j\sum_{i=1}^{m}\sigma_i\sigma(u_j\cdot x_i)\right|\right]$$

$$=\mathbb{E}\left[\sup_{\substack{\|w\|_1\le\Lambda',\\ \|u_j\|_2\le\Lambda}}\frac{1}{m}\sum_{j=1}^{n_2}\left|w_j\sum_{i=1}^{m}\sigma_i\sigma(u_j\cdot x_i)\right|\right], \text{ by symmetry over choosing signs of } w_j \cdots ①$$

$$\le\mathbb{E}\left[\sup_{\|w\|_1\le\Lambda'}\frac{1}{m}\sum_{j=1}^{n_2}|w_j|\underbrace{\sup_{\|u\|_2\le\Lambda}\left|\sum_{i=1}^{m}\sigma_i\sigma(u\cdot x_i)\right|}_{(*)}\right] \cdots ②$$

$$=\frac{\Lambda'}{m}\mathbb{E}\left[\sup_{\|u\|_2\le\Lambda}\left|\sum_{i=1}^{m}\sigma_i\sigma(u\cdot x_i)\right|\right]$$

Also, $u\mapsto\sum_{i=1}^{m}\sigma_i\sigma(u\cdot x_i)$ is continuous, $\{u \mid \|u\|_2\le\Lambda\}$ is compact.
By extrem value thm, supremum at (*) is attained by some $\|u^*\|_2\le\Lambda$.
i.e. $\left|\sum_{i=1}^{m}\sigma_i\sigma(u^*\cdot x_i)\right|=\sup_{\|u\|_2\le\Lambda}\left|\sum_{i=1}^{m}\sigma_i\sigma(u\cdot x_i)\right|$.

X 

Then, by letting $u_j=u^*\ \forall j\in[n_2]$ and taking supremum over w in ①,
we see that ① $\ge$ ②. Hence, $\hat{R}_S(\mathcal{H})=\frac{\Lambda'}{m}\mathbb{E}\left[\sup_{\|u\|_2\le\Lambda}\left|\sum_{i=1}^{m}\sigma_i\sigma(u\cdot x_i)\right|\right]$.

(b)

Use Talagrand's lemma with $\Phi=\sigma$, $h(x)=u\cdot x$, over $h\in\{x\mapsto u\cdot x \mid \|u\|_2\le\Lambda\}=\mathcal{H}$.
Then, $\hat{R}_S(\mathcal{H})=\frac{\Lambda'}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{m}\sigma_i(\Phi\circ h)(x_i)\right|\right]$ (by (a))

$$\le\frac{L\Lambda'}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^{m}\sigma_i h(x_i)\right|\right] \quad (\text{by Talagrand})$$

$$=\frac{L\Lambda'}{m}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{\|u\|_2\le\Lambda}\left|\sum_{i=1}^{m}\sigma_i u\cdot x_i\right|\right]$$

$$= \frac{L\Lambda'}{m} \mathbb{E}\left[\sup_{\substack{\|u\|_2 \le \Lambda \\ s \in \{\pm 1\}}} s \sum_{i=1}^{m} \sigma_i u \cdot x_i\right] \quad \text{(By symmetry over sign of } u\text{)}$$

$$= L\Lambda' \mathbb{E}\left[\sup_{\substack{\|u\|_2 \le \Lambda \\ s \in \{\pm 1\}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i s \, u \cdot x_i\right]$$

$$= L\Lambda' \mathbb{E}\left[\sup_{h \in \mathcal{H}''} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i)\right]$$

$$= L\Lambda' \hat{R}_S(\mathcal{H}')$$

(c)

① $\left|\sum_{i=1}^{m} \sigma_i h(x_i)\right| = \left|\sum_{i=1}^{m} \sigma_i s(u \cdot x_i)\right|$

$$= \left| u \cdot \sum_{i=1}^{m} \sigma_i x_i \right|$$

$$\le \|u\|_2 \cdot \left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2 \quad \text{(By Cauchy-Schwartz)}$$

$$\le \Lambda \left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2, \quad \forall h \in \mathcal{H}', \ h(x) = s(u \cdot x)$$

② Take $u = \frac{\Lambda}{\|\sum_{i=1}^{m} \sigma_i x_i\|_2} \sum_{i=1}^{m} \sigma_i x_i$, $s = \frac{\sum_{i=1}^{m} \sigma_i u \cdot x_i}{|\sum_{i=1}^{m} \sigma_i u \cdot x_i|} = \text{sign}\left(\sum_{i=1}^{m} \sigma_i u \cdot x_i\right)$.

Then, $\|u\|_2 = \Lambda$, $s \in \{\pm 1\}$ with

$$\sum_{i=1}^{m} \sigma_i s(u \cdot x_i) = \left|\sum_{i=1}^{m} \sigma_i u \cdot x_i\right| = \left| u \cdot \sum_{i=1}^{m} \sigma_i x_i \right| = \Lambda \left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2$$

By ①.②, $\sup_{h \in \mathcal{H}'} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x_i) = \frac{\Lambda}{m} \left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2$.

Thus, $\hat{R}_S(\mathcal{H}') = \frac{\Lambda}{m} \mathbb{E}_{\sigma}\left[\left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2\right]$.

(d)

$$\hat{R}_S(\mathcal{H}') = \frac{\Lambda}{m} \mathbb{E}_{\sigma}\left[\left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2\right]$$

$$\le \frac{\Lambda}{m} \sqrt{\mathbb{E}_{\sigma}\left[\left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2^2\right]} \quad \text{(By Jensen)}$$

(e)

If $\|x\|_2 \le r \ \forall x \in S$, $\mathbb{E}_{\sigma} \left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2^2 = \mathbb{E}\left[\sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i \sigma_j x_i \cdot x_j\right]$

$$= \mathbb{E}\left[\sum_{i=1}^{m} \|x_i\|^2\right] \le m r^2$$

Thus $\mathbb{E}\left\|\sum_{i=1}^{m} \sigma_i x_i\right\|_2^2 \le m r^2 \Rightarrow \hat{R}_S(\mathcal{H}') \le \frac{\Lambda}{m}\sqrt{m r^2} = \frac{\Lambda r}{\sqrt{m}}$.

$$\Rightarrow \hat{R}_S(\mathcal{H}) \le L\Lambda' \hat{R}_S(\mathcal{H}') \le \frac{L\Lambda'\Lambda r}{\sqrt{m}}$$

#3.23

(a) $\Pi_{\mathcal{H}}(m) = \max_{\{x_1,\dots,x_m\}\subseteq X} |\{(h(x_1),\dots,h(x_m)) : h\in\mathcal{H}\}|$

For $S=\{x_1,\dots,x_m\}\subseteq X$, let's denote $C(S)=(C(x_1),\dots,C(x_m))$, and $\mathcal{C}(S)=\{C(S): c\in\mathcal{C}\}$.

$\Rightarrow |\mathcal{C}(S)| = |\{c(S): c=c_1\cap c_2 \text{ for } c_1\in\mathcal{C}_1, c_2\in\mathcal{C}_2\}|$.

Suppose $c(S)\neq c'(S)$ for $c=c_1\cap c_2$, $c'=c_1'\cap c_2'$ and $c_1,c_1'\in\mathcal{C}_1$, $c_2,c_2'\in\mathcal{C}_2$.

Then, either $c_1(S)\neq c_1'(S)$ or $c_2(S)\neq c_2'(S)$ (think contrapositive).

Hence, there is injection $\varphi:\mathcal{C}(S)\to\mathcal{C}_1(S)\times\mathcal{C}_2(S)$

$c(S)\mapsto(c_1(S),c_2(S))$ for $c=c_1\cap c_2$, $c_i\in\mathcal{C}_i$.

among many possible choice of $(c_1(S),c_2(S))$, use axiom of choice

Thus, $|\mathcal{C}(S)|\le|\mathcal{C}_1(S)\times\mathcal{C}_2(S)| = |\mathcal{C}_1(S)||\mathcal{C}_2(S)|\le\Pi_{\mathcal{C}_1}(m)\Pi_{\mathcal{C}_2}(m)$. $\cdots(*)$

Take maximum on $(*)$ over all $S\subseteq\mathcal{X}$ with $|S|=m$. Then, $\Pi_{\mathcal{C}}(m)\le\Pi_{\mathcal{C}_1}(m)\Pi_{\mathcal{C}_2}(m)$

(b) $\mathrm{VCDim}(\mathcal{H}) = \max\{m: \Pi_{\mathcal{H}}(m)=2^m\}$. (Maximum size of shattered set)

$\Pi_{\mathcal{H}}(m)\le\left(\frac{em}{d}\right)^d$ for $d=\mathrm{VCDim}(\mathcal{H})$.

(Step 1) By induction on s with (a), $\Pi_{\mathcal{C}_s}(m)\le(\Pi_{\mathcal{C}}(m))^s\le\left(\frac{em}{d}\right)^{ds}$.

If $\Pi_{\mathcal{C}_s}(m)=2^m$, then $2^m\le\left(\frac{em}{d}\right)^{ds}$

For $m=2ds\log_2(3s)$, $\begin{cases} 2^m = 2^{2ds\log_2(3s)} = (3s)^{2ds} & \cdots ① \\ \left(\frac{em}{d}\right)^{ds} = (2es\log_2(3s))^{ds} < \left(2es\cdot\frac{9s}{2e}\right)^{ds} = (3s)^{2ds} & \cdots ② \end{cases}$

see lemma below.

first consider case this value is integer.

Thus, $\left(\frac{em}{d}\right)^{ds} < (3s)^{2ds}\le 2^m$ for $m=2ds\log_2(3s)$. by ①,②

Hence, $\Pi_{\mathcal{C}_s}(m)\le\left(\frac{em}{d}\right)^{ds}<2^m$ for $m=2ds\log_2(3s)$.

i.e. set of size $m=2ds\log_2(3s)$ is not shattered by $\mathcal{C}_s$.

(Step 2) To consider the case where $2ds\log_2(3s) =: \alpha\notin\mathbb{Z}$,

let $\varphi(t)=2^t$, $\psi(t)=\left(\frac{et}{d}\right)^{ds}$. We prove $\varphi(t)>\psi(t)\ \forall t\ge\alpha$.

$\because$) From above, $\varphi(\alpha)>\psi(\alpha)$.

linear approximation at $t=\alpha$.

$\log(\varphi(t)) = t\log 2$, $\log\psi(t) = ds\log\left(\frac{et}{d}\right)\le\frac{ds}{\alpha}(t-\alpha)+\log(\psi(\alpha))$.

log ftn is concave.

Let $l_1(t)=t\log 2$, $l_2(t)=\frac{ds}{\alpha}(t-\alpha)+\log\psi(\alpha)$.

/ /

Then, $l_1(\alpha) = \log(\varphi(\alpha)) > \log(\psi(\alpha)) = l_2(\alpha)$.

$l_1'(t) = \log 2$, $l_2'(t) = \frac{ds}{\alpha}$.

$\log 2 > \frac{ds}{\alpha} \iff \log 2 > \frac{1}{2\log_2 3s}$

$\iff \log 3s = (\log_2 3s)(\log 2) > \frac{1}{2}$, which is true $\forall s \geq 1$.

Thus, $l_1(t) > l_2(t)$ $\forall t \geq \alpha$.

$\Rightarrow \varphi(t) = e^{l_1(t)} > e^{l_2(t)} \geq e^{\log(\psi(t))} = \psi(t)$ $\forall t \geq \alpha$.

Thus, for $m = \lceil \alpha \rceil \in \mathbb{N}$, we have $m \geq \alpha$ so that

$2^m = \varphi(m) > \psi(m) = \left(\frac{em}{d}\right)^{ds}$.

i.e. Any subset of size $\lceil \alpha \rceil$ is not shattered by $\mathcal{C}_s$.

Thus, $VCDim(\mathcal{C}_s) \leq \lceil \alpha \rceil - 1 < \alpha = 2ds \log_2(3s)$

<u>Lemma</u>// $\log_2(3x) < \frac{9x}{2e}$ for $x \geq 2$.

(pf) $\log_2 3x = f(x)$ is concave. Thus, f is below its tangent line at $x=2$.

$f(x) = \frac{\log 3x}{\log 2}$. $f'(x) = \frac{1}{x \log 2}$. $f'(2) = \frac{1}{2\log 2}$.

$f(x) \leq \frac{1}{2\log 2}(x-2) + \frac{\log 6}{\log 2} < \frac{x}{2\log 2} < \frac{9x}{2e}$ (as $e < 9\log 2$)