机器学习理论导引
作业一

2019 年 5 月 7 日

作业提交注意事项

(1) 请严格参照课程网站作业提交方法一节提交作业;

(2) 未按照要求提交作业，或提交作业格式不正确，将会被扣除部分作业分数;

(3) 截止时间后不接收作业，本次作业记零分。

1 [20pts] PAC Learning

讲义中已经证明了轴平行矩形的假设空间是可学习的。这启发我们，无限假设空间也可能是可学习的。本题目给出另一个可学习的无限假设空间的简单的例子。

令 $\mathcal{H}$ 表示一维的阈值函数构成的假设空间，记为 $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$。此处 $h_a : \mathbb{R} \mapsto \{0,1\}$ 是阈值函数 $h_a(x) = \mathbb{I}_{[x<a]}$，仅当 $x < a$ 时取值为 1，否则为 0。显然，假设空间 $\mathcal{H}$ 无限大。假设目标概念 $c \in \mathcal{H}$，即该问题是可分的。

请证明: 假设空间 $\mathcal{H}$ 是 PAC 可学习的, 使用 ERM 算法, 样本复杂度 $m_{\mathcal{H}}(\epsilon,\delta) \le \lceil \log(2/\delta)/\epsilon \rceil$.
提示 1：即证明当 $m > \log(2/\delta)/\epsilon$ 时，以 $1-\delta$ 的概率使得泛化误差 $\mathcal{L}_{\mathcal{D}}(h_S) < \epsilon$.
提示 2：在目标概念 h^* 对应的 a^* 左右设置泛化误差带 $a_0 < a^* < a_1$，使得

$$\Pr_{x\sim\mathcal{D}}[x \in (a_0, a^*)] = \Pr_{x\sim\mathcal{D}}[x \in (a^*, a_1)] = \epsilon.$$

Proof.

由于 $\mathcal{H}$ 是 PAC 可学习的，所以目标概念 c^* 存在于假设空间 $\mathcal{H}$ 中，而由 $\mathcal{H}$ 的特征可知，存在一个与目标概念 c^* 相对应的实值 a^*，使得在分布 $\mathcal{D}$ 中所有小于 a^* 的样例其真实标记均为 *1*，同时所有不小于 c^* 的样例其真实标记均为 *0*。取 a_0 和 $a_1(a_0 < a^* < a_1)$，满足：

$$\Pr_{x\sim\mathcal{D}}[x \in (a_0, a^*)] = \Pr_{x\sim\mathcal{D}}[x \in (a^*, a_1)] = \epsilon$$

基于经验误差最小化原则，将算法设计为：如果数据集 D^m 存在标记为 *1* 的样本，则将标记为 *1* 的样本中最大的样本对应的假设输出，若所有样本的标记均为 *0*，则将最小的样本对应的假设输出，进一步假设 $\Pr[x < a*] : \Pr[x \ge a*] = \kappa : (1-\kappa)$, 则有（还需要分类讨论一下）:

(1) 当 $\kappa \ge \epsilon$ 时，

$$\begin{aligned}
\Pr[E(h) > \epsilon] &= \Pr[(\exists x \in D^m x < a_0 \wedge \forall x \in D^m x \ \notin (a_0, a^*)) \vee \forall x \in D^m x > a_1] \\
&= \Pr[\exists x \in D^m x < a_0 \wedge \forall x \in D^m x \ \notin (a_0, a^*)] + \Pr[\forall x \in D^m x > a_1] \\
&= \left(\sum_{i=1}^{m} C_m^i (\kappa-\epsilon)^i (1-\kappa)^{m-i}\right) + (1-\kappa-\epsilon)^m \\
&= \left(\sum_{i=0}^{m} C_m^i (\kappa-\epsilon)^i (1-\kappa)^{m-i} - C_m^0 (\kappa-\epsilon)^0 (1-\kappa)^m\right) + (1-\kappa-\epsilon)^m \\
&= ((1-\epsilon)^m - (1-\kappa)^m) + (1-\kappa-\epsilon)^m \\
&< (1-\epsilon)^m + (1-\epsilon)^m \\
&= 2(1-\epsilon)^m \le 2\exp(-m\epsilon) \le \delta \\
\Longrightarrow m &> \frac{1}{\epsilon}\ln\frac{2}{\delta}
\end{aligned}$$

(2) 当 $\kappa < \epsilon$ 时，$\Pr[E(h) > \epsilon] = \Pr[\forall x \in D^m x > a_1] = (1-\kappa-\epsilon)^m \le \exp(-m\epsilon) \le \delta \Longrightarrow m > \frac{1}{\epsilon}\ln\frac{1}{\delta}$。

故样本复杂度 $m > \max(\frac{1}{\epsilon}\ln\frac{2}{\delta}, \frac{1}{\epsilon}\ln\frac{1}{\delta}) = \frac{1}{\epsilon}\ln\frac{2}{\delta}$，即 $m_{\mathcal{H}}(\epsilon,\delta) \le \lceil \frac{1}{\epsilon}\ln\frac{2}{\delta} \rceil$。

□

2 [30pts] VC Dimension

本题目探讨有限假设空间情况下的 VC 维的性质。

(1) **[10pts]** 请证明：对于有限假设空间 $\mathcal{H}$，VC 维满足 $\mathrm{VC}(\mathcal{H}) \leq \lfloor \log(|\mathcal{H}|) \rfloor$.

(2) **[10pts]** 上面已经证明了对于有限假设空间的 VC 维上界，然而实际的 VC 维可能会远小于这个上界。请在样本空间 $\mathcal{X} = [0,1]$ 上构造一个无限假设空间 $\mathcal{H}$ 使得 $\mathrm{VC}(\mathcal{H}) = 1$.

(3) **[10pts]** 请在样本空间 $\mathcal{X} = [0,1]$ 上构造一个有限假设空间 $\mathcal{H}$ 使得 $\mathrm{VC}(\mathcal{H}) = \lfloor \log_2(|\mathcal{H}|) \rfloor$.

Proof.

(1) 任给数据集 D^m，若要 $\mathcal{H}$ 能够打散 D^m，则意味着 $\forall \mathrm{b} \in \{0,1\}^m \exists h \in \mathcal{H} \forall x_i \in D^m (h(x_i) = b_i)$，其中 $\{0,1\}^m$ 是所有 m 维 01 向量构成的集合，继而有 $|\mathcal{H}| \geq |\{0,1\}^m| = 2^m \Rightarrow m \leq \log|\mathcal{H}|$，也就是 $VC(\mathcal{H}) \leq \lfloor \log|\mathcal{H}| \rfloor$ 。

(2) 假设空间 $\mathcal{H} = \{h_a : a \in \mathcal{X}\}$，其中 $h_a(x) = \mathbb{I}(x = a)$，易见对于任意两个样本 x_1, x_2，不存在假设 $h \in \mathcal{H}$ 使得 $h(x_1) = h(x_2) = 1$，因此 $\mathcal{H}$ 无法打散任何大小为 *2* 的样本集合，即 $VC(\mathcal{H}) = 1$。

(3) 任取 $\mathcal{X}$ 中的 $\lfloor \log|\mathcal{H}| \rfloor$ 个样本 $x_1, x_2, ..., x_{\lfloor \log|\mathcal{H}| \rfloor}$，将 $\{0,1\}^{\lfloor \log|\mathcal{H}| \rfloor}$ 中的元素按字典序从小到大排列，其中第 i 个向量记为 b_i。同时将 $\mathcal{H}$ 中的第 i 个假设 h_i 定义为 $h_i(x_j) = b_i^j$ *(*b_i^j 为向量 b_i 的第 j 个元素*)*。易见这样的 $\mathcal{H}$ 可以打散 $\{x_1, x_2, ..., x_{\lfloor \log|\mathcal{H}| \rfloor}\}$ 的任意子集，故 $VC(\mathcal{H}) = \lfloor \log_2(|\mathcal{H}|) \rfloor$。

□

3 [20pts] Rademacher Complexity of the Two-Function Hypothesis Set

考虑只包含两个函数的假设集 $\mathcal{F}$，具体表示为 $\mathcal{F}=\{f_1,f_2\}$. 假设对于任意的样本 $x\in\mathcal{X}$，有 $f_1(x)=+1$ 以及 $f_2(x)=-1$ 成立. 对于大小为 m 的样本集 $S=\{x_1,\ldots,x_m\}$，其中 $x_i\in\mathcal{X},\forall i\in\{1,\ldots,m\}$. 试证明：假设集 $\mathcal{F}$ 关于样本集 S 的经验 Rademacher 复杂度上界为 $1/\sqrt{m}$，即

$$\widehat{\mathfrak{R}}_S(\mathcal{F})\le\frac{1}{\sqrt{m}}. \tag{3.1}$$

Proof.

仿照课程第 *4* 章的内容，借助 *hoeffding* 不等式，可以获得一个渐进复杂度一样的界，不过常数大一点：

$$\begin{aligned}
\exp[tE_\sigma(\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sigma_i f(z_i))] &\le E_\sigma[\exp(t\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sigma_i f(z_i))]\\
&= E_\sigma[\sup_{f\in\mathcal{F}}\exp(t\sum_{i=1}^m\sigma_i f(z_i))]\\
&\le \sum_{f\in\mathcal{F}}E_\sigma[\exp(t\sum_{i=1}^m\sigma_i f(z_i))]\\
&= \sum_{f\in\mathcal{F}}\prod_{i=1}^m E_{\sigma_i}[\exp(t\sigma_i f(z_i))]\\
&\le \sum_{f\in\mathcal{F}}\prod_{i=1}^m \exp[\frac{t^2(2f(z_i)^2)}{8}]\\
&= 2\exp(\frac{mt^2}{2})\\
\Longrightarrow E_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sigma_i f(z_i)] &\le \frac{\ln 2}{t}+\frac{mt}{2}\\
\Longrightarrow E_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sigma_i f(z_i)] &\le (\frac{\ln 2}{t}+\frac{mt}{2})_{min}=\sqrt{2m\ln 2}\\
\Longrightarrow \widehat{\mathfrak{R}}_S(\mathcal{F})=E_\sigma[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\sigma_i f(z_i)] &\le \sqrt{\frac{2\ln 2}{m}}
\end{aligned}$$

如果直接按照定义可以证明 $\widehat{\mathfrak{R}}_S(\mathcal{F})=\frac{1}{2^{m-1}}\binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}$，设 Σ_+ 为所有 σ_i 取值为 $+1$ 的集合，Σ_- 为所有 σ_i 取值为 -1 的集合，以下按奇偶对 m 进行讨论：

当 $m=2k+1$ 时，

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &= E_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sigma_i f(z_i)]\\
&= \sum_{i=0}^k\Pr(|\Sigma_+|=i)\frac{1}{m}(2k+1-2i)+\sum_{i=0}^k\Pr(|\Sigma_-|=i)\frac{1}{m}(2k+1-2i)\\
&= 2\sum_{i=0}^k\frac{\binom{2k+1}{i}}{2^{2k+1}}\frac{1}{2k+1}(2k+1-2i)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2^{2k}}\sum_{i=0}^{k}\binom{2k+1}{i}(1-\frac{2i}{2k+1}) \\
&= \frac{1}{2^{2k}}\sum_{i=0}^{k}\binom{2k+1}{i} - \frac{1}{2^{2k}}\sum_{i=1}^{k}\frac{2}{2k+1}(2k+1)\binom{2k}{i-1} \\
&= 1-\frac{2}{2^{2k}}\sum_{i=1}^{k}\binom{2k}{i-1} = 1-\frac{2}{2^{2k}}\sum_{i=0}^{k}\binom{2k}{i} \\
&= 1-\frac{2}{2^{2k}}\frac{1}{2}(2^{2k}-\binom{2k}{k}) = 1-\frac{1}{2^{2k}}(2^{2k}-\binom{2k}{k}) \\
&= \frac{1}{2^{2k}}\binom{2k}{k} = \frac{1}{2^{m-1}}\binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}
\end{aligned}
$$

当 $m=2k$ 时，

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &= E_\sigma[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)] \\
&= \sum_{i=0}^{k}\Pr(|\Sigma_+|=i)\frac{1}{m}(2k-2i)+\sum_{i=0}^{k-1}\Pr(|\Sigma_-|=i)\frac{1}{m}(2k-2i) \\
&= \sum_{i=0}^{k}\frac{1}{2^{2k}}\binom{2k}{i}\frac{2k-2i}{2k}+\sum_{i=0}^{k-1}\frac{1}{2^{2k}}\binom{2k}{i}\frac{2k-2i}{2k} \\
&= 2\sum_{i=0}^{k-1}\frac{1}{2^{2k}}\binom{2k}{i}(1-\frac{i}{k}) \\
&= \frac{1}{2^{2k-1}}\left[\sum_{i=0}^{k-1}\binom{2k}{i}-\frac{1}{k}\sum_{i=0}^{k-1}i\binom{2k}{i}\right] \\
&= \frac{1}{2^{2k-1}}\left[\frac{1}{2}(2^{2k}-\binom{2k}{k})-\frac{2k}{k}\sum_{i=1}^{k-1}\binom{2k-1}{i-1}\right] \\
&= \frac{1}{2^{2k-1}}\left[2^{2k-1}-\frac{1}{2}\binom{2k}{k}-2\sum_{i=0}^{k-2}\binom{2k-1}{i}\right] \\
&= \frac{1}{2^{2k-1}}\left[2^{2k-1}-\frac{1}{2}\binom{2k}{k}-(2^{2k-1}-2\binom{2k-1}{k})\right] \\
&= \frac{1}{2^{2k-1}}\left[2\binom{2k-1}{k}-\frac{1}{2}\binom{2k}{k}\right] = \frac{1}{2k}\left[4\frac{(2k-1)!}{k!(k-1)!}-\frac{(2k)!}{k!k!}\right] \\
&= \frac{1}{2^{2k}}\left[\frac{1}{k!k!}(4k(2k-1)!-(2k)!)\right] = \frac{1}{2^{2k}}\frac{(2k)!}{k!k!} \\
&= \frac{1}{2^{2k}}\binom{2k}{k} = \frac{1}{2^{2k-1}}\binom{2k-1}{k-1} = \frac{1}{2^{m-1}}\binom{m-1}{\lfloor\frac{m-1}{2}\rfloor}
\end{aligned}
$$

由于当 $m=2k$ 和 $m=2k+1$ 时，$\widehat{\mathfrak{R}}_S(\mathcal{F})$ 均等于 $\frac{1}{2^{2k}}\binom{2k}{k}$，且 $\frac{1}{\sqrt{2k+1}}<\frac{1}{\sqrt{2k}}$，故可以统一两者的讨论，仅需证明 $\frac{1}{2^{2k}}\binom{2k}{k}\le\frac{1}{\sqrt{2k+1}}$ 对任意 $k>0$ 均成立即可。容易验证对 $k=1$ 命题成立，归纳假设 $k=n$ 时不等式成立，则当 $k=n+1$ 时 $\frac{1}{2^{2n+2}}\binom{2n+2}{n+1}=\frac{1}{2^{2n+2}}\frac{(2n+2)!}{(n+1)!(n+1)!}\le\frac{(2n+1)(2n+2)}{4(n+1)^2}\frac{1}{\sqrt{2n+1}}=\sqrt{\frac{(2n+1)^2(2n+2)^2}{(2n+2)^4(2n+1)}}=\sqrt{\frac{1}{2n+3}\frac{(2n+2)^2-1}{(2n+2)^2}}<\sqrt{\frac{1}{2n+3}}$，由此可知对 $k>0$，不等式 $\frac{1}{2^{2k}}\binom{2k}{k}\le\frac{1}{\sqrt{2k+1}}$ 恒成立，因此 $\widehat{\mathfrak{R}}_S(\mathcal{F})\le\frac{1}{\sqrt{m}}$。

□

4 [30pts] Rademacher Complexity Property

固定正整数 $m \geq 1$，对任意实数 $\alpha \in \mathbb{R}$ 以及由 $\mathcal{X} \mapsto \mathbb{R}$ 的映射组成的任意两个假设集 $\mathcal{H}_1$、$\mathcal{H}_2$，试证明下列关于 Rademacher 复杂度的等式/不等式成立。

(1) **[10pts]** $\mathfrak{R}_m(\alpha\mathcal{H}_1) = |\alpha|\mathfrak{R}_m(\mathcal{H}_1)$.

(2) **[10pts]** $\mathfrak{R}_m(\mathcal{H}_1+\mathcal{H}_2) = \mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)$, 其中假设集 $\mathcal{H}_1+\mathcal{H}_2$ 具体可表达为 $\mathcal{H}_1+\mathcal{H}_2 = \{h_1+h_2 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$.

(3) **[10pts]** $\mathfrak{R}_m(\mathcal{H}) \leq \mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)$, 其中假设集 $\mathcal{H}$ 定义为 $\mathcal{H} = \{\max(h_1,h_2) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$.

提示：最后一问中你可能会用到 Talagrand's Lemma（又称为 Contraction Lemma）. 具体可参见参考文献 [1] 中 Lemma 26.9（书第 26 章，pp. 381-382）[参考链接].

Proof.

(1)

$$
\begin{aligned}
\mathfrak{R}_m(\alpha\mathcal{H}) &= E_{D^m,\sigma}[\sup_{h\in\alpha\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)] \\
&= E_{D^m,\sigma}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i \alpha h(x_i)] \\
&= \begin{cases} \alpha E_{D^m,\sigma}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)], & \alpha > 0 \\ (-\alpha) E_{D^m,\sigma}[\sup_{h\in\mathcal{H}} \frac{1}{m}\sum_{i=1}^m -\sigma_i h(x_i)], & \alpha \leq 0 \end{cases} \\
&= |\alpha| E_{D^m,\sigma}[\sup_{h\in\alpha\mathcal{H}} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)] \\
&= |\alpha|\mathfrak{R}_m(\mathcal{H})
\end{aligned}
$$

(2)

$$
\begin{aligned}
\mathfrak{R}_m(\mathcal{H}_1+\mathcal{H}_2) &= E_{D^m,\sigma}[\sup_{h\in\mathcal{H}_1+\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^m \sigma_i h(x_i)] \\
&= E_{D^m,\sigma}[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(h_1(x_i)+h_2(x_i)\right)] \\
&= E_{D^m,\sigma}[\sup_{h_1\in\mathcal{H}_1} \frac{1}{m}\sum_{i=1}^m \sigma_i h_1(x_i) + \sup_{h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^m \sigma_i h_2(x_i)] \\
&= E_{D^m,\sigma}[\sup_{h_1\in\mathcal{H}_1} \frac{1}{m}\sum_{i=1}^m \sigma_i h_1(x_i)] + E_{D^m,\sigma}[\sup_{h_2\in\mathcal{H}_2} \frac{1}{m}\sum_{i=1}^m \sigma_i h_2(x_i)] \\
&= \mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)
\end{aligned}
$$

(3) 由 *Contraction Lemma* 有 $\mathfrak{R}_m(|\mathcal{H}_1-\mathcal{H}_2|) \leq \mathfrak{R}_m(\mathcal{H}_1-\mathcal{H}_2)$,

$$
\mathfrak{R}_m(\mathcal{H}) = \mathfrak{R}_m(\max(\mathcal{H}_1,\mathcal{H}_2))
$$

$$
\begin{aligned}
&= E_{D^m,\sigma}[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(\max(h_1(x_i),h_2(x_i))\right)] \\
&= E_{D^m,\sigma}[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(\frac{1}{2}(h_1(x_i)+h_2(x_i)+|h_1(x_i)-h_2(x_i)|)\right)] \\
&= E_{D^m,\sigma}[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(\frac{1}{2}(h_1(x_i)+h_2(x_i)+|h_1(x_i)-h_2(x_i)|)\right)] \\
&= \frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)\right)+\frac{1}{2}\left(\mathfrak{R}_m(|\mathcal{H}_1-\mathcal{H}_2|)\right) \\
&\leq \frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)\right)+\frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1-\mathcal{H}_2)\right) \\
&= \frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)\right)+\frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(-\mathcal{H}_2)\right) \\
&= \frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)\right)+\frac{1}{2}\left(\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)\right) \\
&= \mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)
\end{aligned}
$$

$\square$

Reference

[1] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.