机器学习理论研究导引
作业二

陈晟, MG21330006

2022 年 3 月 30 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2022/04/05 23:59:59**, 截止时间后不再接收作业, 本次作业记零分;

(2) 作业提交方式：使用此 LaTex 模板书写解答，只需提交编译生成的 pdf 文件，将 pdf 文件上传到以下 ftp 服务器的指定位置：
地址：sftp://210.28.132.67:22，用户名：mlt2022，密码：mltspring2022@nju
文件夹位置：/C:/Users/mlt2022/hw_submissions/hw2_submission/；

(3) pdf 文件命名方式：学号-姓名-作业号-v 版本号, 例 MG1900000-张三-2-v1；如果需要更改已提交的解答，请在截止时间之前提交新版本的解答，并将版本号加一；

(4) 未按照要求提交作业，或 pdf 命名方式不正确，将会被扣除部分作业分数.

1 [25pts] Rademacher Complexity Property

固定正整数 $m \geqslant 1$, 对任意实数 $\alpha \in \mathbb{R}$ 以及由 $\mathcal{X} \to \mathbb{R}$ 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) **[5pts]** 若 $\mathcal{H}_1$ 中仅包含一个假设, 即 $\mathcal{H}_1 = \{h_1\}$, 则 $\mathfrak{R}_m(\mathcal{H}_1) = 0$.

(2) **[5pts]** $\mathfrak{R}_m(\alpha\mathcal{H}_1) = |\alpha|\mathfrak{R}_m(\mathcal{H}_1)$.

(3) **[5pts]** $\mathfrak{R}_m(\mathcal{H}_1 + \mathcal{H}_2) = \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$.

(4) **[10pts]** $\mathfrak{R}_m(\mathcal{H}) \leqslant \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$, 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等式 $\max(a, b) = \frac{1}{2}(a + b + |a - b|)$ 以及 Talagrand's Lemma (又称为 Contraction Lemma). 关于 Talagrand's Lemma, 可参见《Understanding Machine Learning: From Theory to Algorithms》 Lemma 26.9（书第 26 章, pp. 381-382）

Proof.

此处用于写解答 (中英文均可) □

(1) 若 G 为假设集，则

$\because \mathfrak{R}_m(G) = E_s\left[E_\sigma\left[sup_{u\in G_{|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right] \le E_s\left[\frac{\sqrt{m}\sqrt{2ln|G_{|S}|}}{m}\right]$

$\therefore \mathfrak{R}_m(G) \le \sqrt{\frac{2ln\prod_G(m)}{m}}$

当 $G = \mathcal{H}_1$ 时，显而易见

$\mathfrak{R}_m(H1) \le \sqrt{\frac{2ln\prod_G(m)}{m}} = \sqrt{\frac{2lnm}{m}}$

由于 $d\frac{2lnx}{x} = \frac{2-2lnx}{x^2} = 0$

$x = e$ 为集大值点，$x \ge 1$ 时也是最大值

而 $\frac{2lne}{e} = 1$

所以 $\frac{2lnm}{m} \le 1, m \ge 1, m \in N$

即 $\mathfrak{R}_m(\mathcal{H}_1) < 1, \mathfrak{R}_m(\mathcal{H}_1) = 0$

(2)

$\because \mathfrak{R}_m(\alpha\mathcal{H}_1) = E_s\left[E_\sigma\left[sup_{u\in\alpha\mathcal{H}_{1|s}} \frac{1}{m}\sum_{i=1}^m \alpha\sigma_i u_i\right]\right]$

$\therefore \mathfrak{R}_m(\mathcal{H}_1) = E_s\left[\alpha E_\sigma\left[sup_{u\in\mathcal{H}_{1|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right] = \alpha E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{1|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right]$

即 $\mathfrak{R}_m(\alpha\mathcal{H}_1) = |\alpha|\mathfrak{R}_m(\mathcal{H}_1)$

(3)

$\because \mathfrak{R}_m(\mathcal{H}_1 + \mathcal{H}_2) = E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{1|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right] + E_\sigma\left[sup_{u\in\mathcal{H}_{2|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right] = E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{1|s}} \frac{1}{m}\sum\right.\right.$
$E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{2|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right]$

$\therefore \mathfrak{R}_m(\mathcal{H}_1 + \mathcal{H}_2) = \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$

(4)

假设集 $\mathcal{H}$ 定义为 $\mathcal{H} = \{\max(h_1, h_2) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$

$\because \mathfrak{R}_m(\mathcal{H}) = E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right]$

$\therefore \mathfrak{R}_m(\mathcal{H}) = E_s\left[E_\sigma\left[sup_{u\in\{\max(h_1,h_2):h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2\}_{|s}} \frac{1}{m}\sum_{i=1}^m \sigma_i u_i\right]\right]$

由 *Talagrand's Lemma*

$\mathfrak{R}_m\left(\mathcal{H}\right)=E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{|s}}\frac{1}{m}\sum_{i=1}^{m}\sigma_iu_i\right]\right]\leq\mathfrak{R}_m\left(\mathcal{H}_1+\mathcal{H}_2\right)=E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{1|s}}\frac{1}{m}\sum_{i=1}^{m}\sigma_iu_i\right]+E_\sigma\left[sup_{u\in\right.\right.$
$E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{1|s}}\frac{1}{m}\sum_{i=1}^{m}\sigma_iu_i\right]\right]+E_s\left[E_\sigma\left[sup_{u\in\mathcal{H}_{2|s}}\frac{1}{m}\sum_{i=1}^{m}\sigma_iu_i\right]\right]$

2 [25pts] VC Dimension of Voting

考虑 VC 维为 d 的假设空间 $\mathcal{H}$, 其中 $h \in \mathcal{H} : \mathcal{X} \to \{-1, +1\}$, 令 $\mathcal{M}$ 表示由 $\mathcal{H}$ 中任意 $k \geqslant 1$ 个假设根据多数投票法生成的假设所组成的假设空间, 即

$$\mathcal{M} = \left\{ h(\boldsymbol{x}) = \underset{y \in \{-1,+1\}}{\arg\max} \sum_{i=1}^{k} \mathbb{I}(h_i(\boldsymbol{x}) = y) : h_1, \cdots, h_k \in \mathcal{H} \right\}. \tag{2.1}$$

若 $kd \geqslant 4$, 试证明 $\mathcal{M}$ 的 VC 维有上界 $O(kd\ln(kd))$.

Proof.

此处用于写解答 *(中英文均可)*

$\mathcal{M}$ 为假设 $VC(\mathcal{M}) = max\{m : \prod_{\mathcal{M}} = 2^m\}$

$VC(\mathcal{M}) = d$ 表明存在大小为 d 的示例集能被假设空间打散

由于 $\mathcal{M}$ 是由 k 个假设根据多数投票法生成的假设，并且 VC 维为 d

假设能被 $\mathcal{M}$ 打散的最大样本集大小为 d ，则 $\prod_{\mathcal{F}} = 2^d$ ，由

$\prod_{\mathcal{M}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$

$\prod_{\mathcal{F}}(m) \leq \prod_{\mathcal{F}^{(1)}}(m) \cdot \prod_{\mathcal{F}^{(2)}}(m)$ $\mathcal{F}^{(1)} \subset y_1^x$ 和 $\mathcal{F}^{(2)} \subset y_2^x$,$\mathcal{F}$ 是它们的笛卡尔乘积

以及 $\prod_{\mathcal{H}}(m) \leq \left(\frac{e \cdot m}{d}^d\right)$

由 $\prod_{\mathcal{F}}(m) \leq \prod_{i=1}^{l} \prod_{\mathcal{F}^{(i)}}(m) \leq \prod_{i=1}^{l} \prod_{j=1}^{d_i} \left(\frac{e \cdot m}{d_{i-1}+1}\right)^{d_{i-1}+1}$

综上令 $N = \sum_{i=1}^{l} \sum_{j=1}^{d_i} (d_{i-1}+1) = kd$

$\therefore \prod_{\mathcal{F}}(m) \leq (e \cdot m)^k d$

$2^d \leq (de)^{kd}$

$\therefore \mathcal{M}$ 的 VC 维有上界 $O(kd\ln(kd))$

3 [25pts] 一维阈值函数的经验 Rademacher 复杂度

令 $\mathcal{H}=\{h_a : a\in\mathbb{R}\}$ 表示一维阈值函数 $h_a(x)=\mathbb{I}(x\le a)$ 构成的假设空间，集合 D 包含实轴上 m 个不同的点，本题将引导大家估计 $\mathcal{H}$ 关于 D 的经验 Rademacher 复杂度 $\hat{\mathfrak{R}}_D(\mathcal{H})$.

(1) **[10pts]** 试证明：$\hat{\mathfrak{R}}_D(\mathcal{H})=O\left(\sqrt{\frac{\log m}{m}}\right)$. 提示：你可能需要使用第三章课件中的某个定理或推论.

(2) **[15pts]** 事实上，第一问中对经验 Rademacher 复杂度的估计并非是最紧的，这意味着将广泛成立的定理应用在特定问题时，不一定能获得最好的结果. 本问我们将利用一个关于随机游走的结论给出一个准确的估计.

Definition 1 (一维随机游走)**.** *假设在实轴上有一个点，其初始位置 $x_0=0$. 一个***一维随机游走***是一个总共进行 n 轮的过程，在每一轮中，该点以概率 p 向正方向前进 1 个单位，以概率 $1-p$ 向负方向前进 1 个单位. 令 σ_i，$i=1,2,\ldots,n$ 为定义在 $\{-1,+1\}$ 上的随机变量，且 $P(\sigma_i=+1)=p$，那么该点第 k 轮过后所在的位置可以表示为 $x_k=\sum_{i=1}^k\sigma_i$.*

Theorem 1 (一维随机游走的最远距离期望)**.** *在一个 n 轮的一维随机游走中，若 $p=1/2$，那么该点在整个过程中离初始位置最远的距离的期望为 $\Theta(\sqrt{n})$，即：*

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{i\in\{1,2,\ldots,n\}} x_i\right]=\Theta\left(\sqrt{n}\right) .$$

试利用上述定义和定理证明：$\hat{\mathfrak{R}}_D(\mathcal{H})=\Theta\left(\frac{1}{\sqrt{m}}\right)$.

(3) **[20pts*]** 对于一维阈值函数这个特殊的假设空间，我们还可以使用一些组合数学的技巧，先写出这个 Rademacher 复杂度的组合数表达式，再对组合数的大小进行估计，可以获得和第 (2) 问相同的结果. **这一问不占基本分值，感兴趣的同学可以尝试按照下面的步骤完成证明. 完成 Step 2 和 Step 3 部分或全部证明的同学，将在本学期的作业中获得一些额外的分数（不会超过作业分数的上限）.**

Step 1. 在经验 Rademacher 复杂度的表达式中，实际上期望运算就是将 $\boldsymbol{\sigma}$ 的 2^m 种取值下内部表达式的值求出了算术平均. 因此记 A_i 为使得内部表达式的取值为 i 的 $\boldsymbol{\sigma}$ 的取值总数，那么有 $\hat{\mathfrak{R}}_D(\mathcal{H})=\frac{\sum_{i=0}^m i*A_i}{m\cdot 2^m}$.
Step 2. 定义 $B_k=\sum_{i=0}^k A_i$，那么 B_k 可以由一个等价的格子计数问题 (lattice path enumeration) 的相关结论结合数学归纳法给出. 进一步可以写出 A_k 的表达式.
Step 3. 最后需要估计一系列组合数的求和的大小。通过对组合数对称性的一些观察，可以通过简单的变形，在求和式内部凑出裂项结构，从而将求和消去。在使用 Stirling 公式对组合数进行估计后，分子剩余部分恰好为 $\sqrt{m}\cdot 2^m$，和分母的 $m\cdot 2^m$ 相除即可完成证明.

参考文献：定理 1 来自论文 How Far Might We Walk at Random；第 (3) 问 Step 2 中关于格子计数问题的结论可参考文章 Lattice Path Enumeration 的定理 10.1.3。

Proof. *此处用于写解答 (中英文均可)*

(1) $\because \hat{\mathfrak{R}}_D(\mathcal{H}) = \frac{1}{m}E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\mathbb{I}(x_i \le a)\right] \le \frac{1}{m}E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\right]$
阈值函数 $\because \prod_{\mathcal{H}}(m) = 2^m$
$\therefore \frac{1}{m}E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\right] \le \frac{1}{m}\left(mlogm\right)^{\frac{1}{2}}$
即 $\hat{\mathfrak{R}}_D(\mathcal{H}) \le \frac{1}{m}\left(mlogm\right)^{\frac{1}{2}}$
$\hat{\mathfrak{R}}_D(\mathcal{H}) = O\left(\sqrt{\frac{\log m}{m}}\right)$

(2) 由一维随机游走的最远距离期望：在一个 n 轮的一维随机游走中，若 $p = 1/2$，那么该点在整个过程中离初始位置最远的距离的期望为 $\Theta\left(\sqrt{n}\right)$，即 *Theorem 1*：

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{i\in\{1,2,\ldots,n\}} x_i\right] = \Theta\left(\sqrt{n}\right) \ .$$

可以类比本题目，$\mathcal{H} = \{h_a : a \in \mathbb{R}\}$ 表示一维阈值函数 $h_a(x) = \mathbb{I}(x \le a)$ 构成的假设空间，由于 D 是实轴上的点，实轴上有无数的点，所以 a 取任何值都可以看作是将整个实轴分成了相等的两份，因此，D 上的某一点 x_i，$\mathbb{I}(x_i \le a) = 1$ 的概率就为 $\frac{1}{2}$，即可以用到题目所给的 *Theorem 1*

由 *(1)* $\because \hat{\mathfrak{R}}_D(\mathcal{H}) = \frac{1}{m}E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\mathbb{I}(x_i \le a)\right] \le \frac{1}{m}E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\right]$
$E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\mathbb{I}(x_i \le a)\right] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{i\in\{1,2,\ldots,m\}} x_i\right] = \Theta\left(\sqrt{m}\right)$
因而可得：$\hat{\mathfrak{R}}_D(\mathcal{H}) = \frac{1}{m}E_\sigma\left[sup\sum_{m}^{i=1}\sigma_i\mathbb{I}(x_i \le a)\right] = \frac{1}{m}\Theta\left(\sqrt{m}\right)$
即 $\hat{\mathfrak{R}}_D(\mathcal{H}) = \Theta\left(\sqrt{1/\sqrt{m}}\right)$

4 [25pts] VC Dimension and the Number of Parameters

回顾课件中精确计算 VC 维的几个例子:

- 阈值函数的假设空间为 $\mathcal{H} = \{\mathrm{sign}(\mathbb{I}(x \leqslant a) - 0.5) : a \in \mathbb{R}\}$, 假设有 1 个参数, 且 $\mathcal{H}$ 的 VC 维为 1;
- 区间函数的假设空间为 $\mathcal{H} = \{\mathrm{sign}(\mathbb{I}(x \in [a,b]) - 0.5) : a, b \in \mathbb{R}, a \leqslant b\}$, 假设有 2 个参数, 且 $\mathcal{H}$ 的 VC 维为 2;
- 课件中给出了 $\mathbb{R}^d$ 中线性超平面的假设空间, 假设有 $d+1$ 个参数, 且假设空间的 VC 维为 $d+1$.

在这些例子中, 假设的参数个数等于假设空间的 VC 维. 该结论是否对任意假设空间都成立呢? 试给出下列各小问的假设空间的数学表达式, 假设的参数个数, VC 维, 并给出 VC 维的证明.

(1) **[10pts]** 所有经过 $(0,0)$ 点的正弦函数, 函数值大于等于 0 时标记为 $+1$, 否则为 -1.

(2) **[15pts]** 所有正三角形, 三角形边缘与内部为 $+1$, 外部为 -1.

提示: 正三角形的 VC 维可能较难证明, 可以尝试给出尽可能紧的上界与下界.

Solution.

此处用于写解答 (中英文均可)

(1) 根据题目可得,所有经过 (0,0) 点的正弦函数的假设空间:

$\mathcal{H} = \{sign(m * sin(ax + k\pi)) : m \in \mathbb{R}, a \in \mathbb{R}, k \in \mathbb{N}\}$

根据假设的表达式可得,参数有 3 个,分别是 m,a,k

要验证 VC 维,就要看是否有大小为 3 的示例集能将其打散,即:

$\exists D = \{x_1, x_2, x_3\}$, 显然无论 x_1, x_2, x_3 三者的关系如何,都可以由对应的假设 $h_{m,a,k}$ 满足其所有分类集合如 $1, 1, -1, -1, 1, -1$ 等,这是由于正弦函数是周期函数,m 可以实现任意宽度上的变化,a 可以实现函数的水平轴反转,k 可以实现水平方向的平移,使在一个区域内既可能大于 0 也可能小于 0

显然,正弦函数是可以被大小为 3 的示例集打散的。

但是当 d>3 时,可以和 d=3 时做类似的推断,无论 d 取任何值,正弦函数都存在假设 $h_{m,a,k}$ 能满足其不同的分类集合

因此,所有经过 (0,0) 点的正弦函数 $\mathcal{H} = \{sign(m * sin(ax + k\pi)) : m \in \mathbb{R}, a \in \mathbb{R}, k \in \mathbb{N}\}$

$VC(\mathcal{H}) = \infty$

(2) 把一个正三角形的底边放在二维坐标轴的横轴上,并且原点为正三角形底边的中点,则可以通过边长来确定正三角形边以及内部所构成的区域函数

设正三角形边长为 L,则区域为: $y >= 0 \cap y \geq \sqrt{3}x - \frac{L}{2} \cap y \leq -\sqrt{3}x + \frac{L}{2}$

$a.\mathcal{H} = \left\{sign((x, y) \in y >= 0 \cap y \geq \sqrt{3}x - \frac{L}{2} \cap y \leq -\sqrt{3}x + \frac{L}{2}) : l \in \mathbb{R}, l > 0\right\}$

或者说,可以用一个坐标点和三角形边长来确定该正三角形的区域 D,即 $D(a, b, l)$,a,b 为三角形的一个点的坐标,l 为正三角形边长,因此

$b.\mathcal{H} = \{sign((x, y) \in D(a, b, l)) : a, b, l \in \mathbb{R}, l > 0\}$

如果以 a. 来看,假设空间只有一个参数 l

显然，若 D 为一维，即 $D=(x_1,y_1)$ 就可以将其假设空间打散，若 D 为二维，简单思考也能满足

但是若 D 为 *8* 维，即即 $D=(x_1,y_1),(x_2,y_2),(x_3,y_3)...(x_8,y_8)$, 则会出现问题

首先考察任意 *8* 个点的凸包. 如果有点在凸包内, 那么要凸包上的点在里面, 凸包里的点在外面，这显然是不可能的. 否则就是 *8* 个点都在凸包上, 取不相邻的 *4* 个在里面, 另外不相邻的 *4* 个就要在外面，由于在外面至少要在三角形一条边的外面，根据鸽笼原理，至少有两个点在同一边的外面。这样势必那两点间的应该在里面的点也会被切出去，就会发生矛盾

所以 $VC(\mathcal{H})=8$