机器学习理论研究导引
作业二

庄镇华 502022370071

2023 年 4 月 15 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2023/04/26 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 文件提交至南大网盘: https://box.nju.edu.cn/u/d/efa16754c94d4904b95d/

(3) pdf 文件命名方式: 学号-姓名-作业号-v 版本号, 例 MG1900000-张三-1-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;

(5) 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] Grouth Function and Sauer's Lemma

Sauer 引理提供了增长函数的上界, 本题旨在探讨该上界是否是紧的.

(1) **[10pts]** 令 $\mathcal{H} = \{h_{a,b} \mid a \leqslant b, h_{a,b}(x) = \mathbb{I}(x \in [a,b])\}$ 为 $\mathbb{R}$ 上所有区间函数构成的函数空间. 计算 $\mathcal{H}$ 的增长函数 $\Pi_{\mathcal{H}}(m)$.

(2) **[15pts]** 对于任意 VC 维 d, Sauer 引理是否是紧的? 即是否存在 VC 维为 d 的函数空间 $\mathcal{H}$, 使得 $\Pi_{\mathcal{H}}(m) = \sum_{i=0}^{d} \binom{m}{i}$?

Solution.

(1) 将 $x_1, x_2, \cdots, x_m$ 进行大小排序，仅有 *1* 个样本输出为 *1* 的情况有 m 种，仅有 *2* 个样本输出为 *1* 的情况有 *m-1* 种，$\cdots$，所有样本输出为 *1* 的情况有 *1* 种，所有样本输出为 *0* 的情况有 *1* 种，则

$$
\begin{aligned}
\Pi_{\mathcal{H}}(m) &= \max_{|D|=m} |\mathcal{H}_{|D}| = \max_{\{x_1,\ldots,x_m\}\subset\mathcal{X}} |\{(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_m)) \,|h \in \mathcal{H}\}| \\
&= 1 + m + m - 1 + \cdots + 1 \\
&= \frac{m^2 + m + 2}{2}
\end{aligned}
$$

(2) 首先计算区间函数的 *VC* 维, 其中 $h_{a,b}(x) = \mathbb{I}(x \in [a,b])$。令 $D = \{1, 2\}$, 易知 $\mathcal{H}$ 能打散 D , 因此 $VC(\mathcal{H}) \geqslant 2$。对于任意大小为 *3* 的样本集 $D' = \{x_1, x_2, x_3\}$, 不妨设 $x_1 < x_2 < x_3$, 则分类结果 $(1, 0, 1)$ 不能被 $\mathcal{H}$ 中的任何区间函数实现, 因为当 $h_{a,b}(x_1) = 1$ 且 $h_{a,b}(x_3) = 1$ 时, 必有 $h_{a,b}(x_2) = 1$。所以 $\mathcal{H}$ 无法打散任何大小为 *3* 的样本集, 于是根据 VC 维的定义可知区间函数的 VC 维为 *2*,

根据 *Sauer* 引理有, $\Pi_{\mathcal{H}}(m) \leq \binom{m}{0} + \binom{m}{1} + \binom{m}{2} = 1 + m + \frac{m(m-1)}{2} = \frac{m^2+m+2}{2}$, 而根据第 *(1)* 问, $\Pi_{\mathcal{H}}(m) = \frac{m^2+m+2}{2}$, 因此存在 *VC* 维为 *2* 的区间函数空间 $\mathcal{H}$, 使得 $\Pi_{\mathcal{H}}(m) = \sum_{i=0}^{d} \binom{m}{i}$, 也即对于任意 *VC* 维 d, *Sauer* 引理是紧的。

2 [20pts] 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) **[5pts]** $\mathfrak{R}_m(\mathcal{H}) \leqslant \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$, 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)

Solution.

(1) $\mathfrak{R}_m(\mathcal{H}_1) = \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_1}\frac{1}{m}\sum_{i=1}^m \sigma_i h\left(\boldsymbol{x}_i\right)\right]\right]$，又因为 σ_i 是 $\{-1,1\}$ 上服从均匀分布的随机变量，$\mathcal{H}_1 = \{h_1\}$，所以 $\mathfrak{R}_m(\mathcal{H}_1) = \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i h_1\left(\boldsymbol{x}_i\right)\right]\right] = \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\boldsymbol{\sigma}}\left[\sigma_i\right] h_1\left(\boldsymbol{x}_i\right)\right] = 0$

(2) $\mathfrak{R}_m(\alpha\mathcal{H}_1) = \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_1}\frac{1}{m}\sum_{i=1}^m \alpha\sigma_i h\left(\boldsymbol{x}_i\right)\right]\right]$
$= \mathbb{E}_{X\subset\mathcal{X}:|X|=m}|\alpha|\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_1}\frac{1}{m}\sum_{i=1}^m \sigma_i h\left(\boldsymbol{x}_i\right)\right]\right]$
$= |\alpha|\mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_1}\frac{1}{m}\sum_{i=1}^m \sigma_i h\left(\boldsymbol{x}_i\right)\right]\right]$
$= |\alpha|\mathfrak{R}_m(\mathcal{H}_1)$

(3) $\mathfrak{R}_m(\mathcal{H}_1 + \mathcal{H}_2) = \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h_1\in\mathcal{H}_1,h_2\in\mathcal{H}_2}\frac{1}{m}\sum_{i=1}^m \sigma_i(h_1\left(\boldsymbol{x}_i\right) + h_2\left(\boldsymbol{x}_i\right))\right]\right]$
$= \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_1}\frac{1}{m}\sum_{i=1}^m \sigma_i h\left(\boldsymbol{x}_i\right)\right]\right] + \mathbb{E}_{X\subset\mathcal{X}:|X|=m}\left[\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_2}\frac{1}{m}\sum_{i=1}^m \sigma_i h\left(\boldsymbol{x}_i\right)\right]\right]$
$= \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$

(4) $\mathfrak{R}_m(\mathcal{H}) = \mathfrak{R}_m(\max(\mathcal{H}_1,\mathcal{H}_2)) = \mathfrak{R}_m(\frac{\mathcal{H}_1+\mathcal{H}_2+|\mathcal{H}_1-\mathcal{H}_2|}{2}) = \frac{\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)+\mathfrak{R}_m(|\mathcal{H}_1-\mathcal{H}_2|)}{2}$

由 *Talagrand's Lemma*,即设 f 为 *L-Lipschitz* 连续函数,并且 $\mathcal{H}$ 是假设集,那么 $\mathfrak{R}_m(f\circ\mathcal{H}) \leq L\cdot\mathfrak{R}_m(\mathcal{H})$;

易知函数 *f(x)=|x|* 为 *1-Lipschitz* 连续函数，因此 $\mathfrak{R}_m(|\mathcal{H}_1 - \mathcal{H}_2|) \leq 1\cdot\mathfrak{R}_m(\mathcal{H}_1 - \mathcal{H}_2) = \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(-\mathcal{H}_2) = \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$;

因此 $\mathfrak{R}_m(\mathcal{H}) = \frac{\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)+\mathfrak{R}_m(|\mathcal{H}_1-\mathcal{H}_2|)}{2} \leq \frac{\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)+\mathfrak{R}_m(\mathcal{H}_1)+\mathfrak{R}_m(\mathcal{H}_2)}{2} = \mathfrak{R}_m(\mathcal{H}_1) + \mathfrak{R}_m(\mathcal{H}_2)$。

3 [25pts] VC Dimension and PAC Learnability

证明: 若概念类 $\mathcal{C}$ 的 VC 维是无穷, 则 $\mathcal{C}$ 不是 PAC 可学的.

Proof.

首先直接给出 ***No-Free-Lunch* 定理**。设 A 为针对域 $\mathcal{X}$ 上损失为 *0*、*1* 的二分类任务的任意算法，$m < |\mathcal{X}|/2$ 为训练集样本大小，那么存在 $\mathcal{X} \times \{0,1\}$ 上的分布 $\mathcal{D}$ 使得 *1)* 存在函数 $f : \mathcal{X} \to \{0,1\}$ 使得 $L_{\mathcal{D}}(f) = 0$; *2)* 存在样本集 $S \sim \mathcal{D}^m$，使得至少有 1/7 的概率 $L_{\mathcal{D}}(A(S)) \geq 1/8$ 成立。

然后利用**反证法**证明原命题。若概念类 $\mathcal{C}$ 是 *PAC* 可学的，那么存在算法 A 使得对于任意 $\epsilon, \delta > 0$，存在 $M(\epsilon, \delta)$ 对于任意分布 $\mathcal{D}$ 满足：如果 $m > M(\epsilon, \delta)$，有 $P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*) + \epsilon) < \delta$，其中 $h^* = arg\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$。

令 $\epsilon < 1/8, \delta < 1/7, m > \mathcal{M}(\epsilon, \delta)$，因为 $VC(\mathcal{H}) = \infty$，所以 $\mathcal{H}$ 可以打散 $x_1, ..., x_{2m} \in \mathcal{X}$，根据 *No-Free-Lunch* 定理，存在 $x_1, ..., x_{2m}$ 及其标签上的分布 $\mathcal{D}$ 满足：存在 $f : \mathcal{X} \to \{0,1\}$ 使得 $L_{\mathcal{D}}(f) = 0, P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) < 1/7$ 成立。

因为 $\mathcal{D}$ 为 $\{x_1, ..., x_{2m}\}$ 及其标签上的分布，并且 $\mathcal{H}$ 可以打散这个样本集，因此 $L_{\mathcal{D}}(h^*) = 0$, $P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*)+\epsilon) \geq P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > 1/8) > 1/7 > \delta$,这与 $P_{S\sim\mathcal{D}^m}(L_{\mathcal{D}}(A(S)) > L_{\mathcal{D}}(h^*) + \epsilon) < \delta$ 矛盾，因此假设不成立，原命题得证。

4 [30pts] 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) **[20pts]** 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.

Proof.

假设的参数个数不一定等于假设空间的 VC 维。4(1) 中参数个数为 1 但是假设空间的 VC 维为无穷。

(1) 假设空间的数学表达式为 $\mathcal{H}=\{sign(\sin(\alpha x)):\alpha\in\mathbb{R}\}$, 假设的参数个数为 1, VC 维为 ∞, 事实上，对于任意大小 m，选择样本点 $x_i=10^{-i}$, 其中 $i=1,\ldots,m$，对于任意给定的标签 $y_1,y_2,\ldots,y_m$, 其中 $y_i\in\{-1,1\}$，只需要设定 $\alpha=\pi(1+\sum_{i=1}^{m}\frac{(1-y_i)10^i}{2})$ 即可满足条件，由 m 的任意性可知该假设空间的 VC 维为无穷。

(2) 假设空间的数学表达式为 $\mathcal{H}=\{h_{(a_1,b_1,a_2,b_2)}\}$，此处点 $(a_1,b_1),(a_2,b_2)$ 不重合，且为正三角形的两个顶点，约定其第三个顶点位于右下方。

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2)=\begin{cases}+1 & \text{如果 } (x_1,x_2) \text{ 位于定义的正三角形边缘与内部}\\ -1 & \text{其余情况}\end{cases}$$

假设的参数个数为 4, VC 维为 6 或 7。

易知任意三角形的 VC 维为 7, 因此正三角形的 VC 维一定小于等于 7, 又容易得出存在 6 个点（取正六边形的 6 个顶点）可以被正三角形空间打散，因此正三角形的 VC 维大于等于 6, 因此正三角形的 VC 维是 6 或 7。

下面首先证明任意三角形的 VC 维为 7。圆上七等分点所有可能的标签都可以使用三角形打散，因为在所有情况中，负标签最多形成 3 个连续块，而三角形的一条边可以隔出一个块。但是，任意 8 个点都没办法被打散，如果存在一个点在其余点的凸包内，则无法将该点标记为负例，将其他点标记为正例；否则，无法以交替的 +, -, +, -, +, -, +, -顺序标记这些点。