机器学习理论研究导引
作业三

庄镇华 502022370071

2023 年 5 月 5 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2023/05/18 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/6395c6d776cf4177b78b/

(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 [20pts] VC Dimension and Generalization Bound

试给出轴平行矩形假设空间基于 VC 维的泛化误差上界. 该误差界与书中式 (2.23) 相比有什么差别? 试解释该差别的原因.

定理 4.3 若假设空间 $\mathcal{H}$ 的有限 VC 维为 d, $h \in \mathcal{H}$, 则对 $m > d$ 和 $0 < \delta < 1$ 有

$$P\left(\left|E(h) - \widehat{E}(h)\right| \leqslant \sqrt{\frac{8d\ln\frac{2cm}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1 - \delta.$$

Solution.

易知轴平行矩形假设空间为无限假设空间，首先证明轴平行矩阵假设空间 $\mathcal{H}$ 的 VC 维为 4，我们需要找到 4 个可以被假设空间 $\mathcal{H}$ 打散的点，并且证明不存在 5 个点可以被 $\mathcal{H}$ 打散。其中 4 个可以被 $\mathcal{H}$ 打散的点如左图所示。不失一般性，如右图所示，在任意 5 个点 $C = \{c_1, \cdots, c_5\}$ 的情景下，我们可以选中最左边、最右边、最上边和最下边的点，我们记 c_5 为那个没被选中的点，并且给定标签为 $(1, 1, 1, 1, 0)$，显然不存在任何一个轴平行矩形可以得到这组标签，因此 C 不能被 $\mathcal{H}$ 打散，因此 $VC(\mathcal{H}) = 4$。

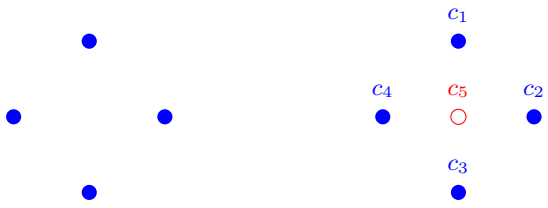

图 1: 左, 能被轴平行矩形打散的 4 个点; 右, 任意轴平行矩形不能将 $c5$ 标记为 0, 其余标记为 1

已经知道轴平行矩形假设空间的 VC 维为 4，结合书中定理 4.3 可以得到该假设空间基于 VC 维的泛化误差上界：对于 $0 < \delta < 1$，以至少 $1 - \delta$ 的概率有

$$E(h) \leqslant \widehat{E}(h) + \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}} = \widehat{E}(h) + \sqrt{\frac{32\ln\frac{em}{2} + 8\ln\frac{4}{\delta}}{m}}$$

在给定 δ 与 $\mathcal{H}$ 之后，我们考虑泛化误差上界关于样本数目增加时的收敛情况是在取等条件下考虑问题，即 $\epsilon = \sqrt{\frac{32\ln\frac{em}{2} + 8\ln\frac{4}{\delta}}{m}}$，在其他因素确定的情况下，误差上界 ϵ 与 $\sqrt{\frac{logm}{m}}$ 成正比，因此收敛速度为 $O(\sqrt{\frac{logm}{m}})$。

书中式 (2.23) 为 $m \geqslant \frac{4}{\epsilon}\ln\frac{4}{\delta}$，同理，取等号时 $\epsilon = \frac{4}{m}\ln\frac{4}{\delta}$，收敛速度为 $O(\frac{1}{m})$。

可以发现基于 VC 维的误差上界比书中式 (2.23) 的误差上界松很多，收敛速度也慢很多，这是因为 VC 维的定义与数据分布无关，因此基于 VC 维的分析结果是分布无关、数据独立的，也就是说对任意数据分布都成立。这使得基于 VC 维的分析结果具有一定的“普适性”；但另一方面，由于没有考虑数据自身，基于 VC 维的分析结果通常比较“松”，而书中式 (2.23) 的分析结果是针对轴平行矩形这一特定问题分析的，因此其误差上界较紧。

2 [80pts] VC-dimension Generalization in Realizable Case

定理 4.3 若假设空间 $\mathcal{H}$ 的有限 VC 维为 d, $h \in \mathcal{H}$, 则对 $m > d$ 和 $0 < \delta < 1$ 有

$$P\left(\left|E(h) - \widehat{E}(h)\right| \leqslant \sqrt{\frac{8d\ln\frac{2cm}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1 - \delta.$$

书中定理 4.3 基于 VC 维给出了泛化误差上界, 该上界对 VC 维 d 与样本量 m 的依赖为 $O(\sqrt{(m/d)^{-1}\log(m/d)})$. 本题将探讨可分情形下这一泛化界是否能得到提升. 令 $\mathcal{H}$ 为一 VC 维为 d 的假设空间, 目标概念 $c \in \mathcal{H}$ 在假设空间中.

(1) **[15pts]** $S, S' \sim \mathcal{D}^m$ 是两个从分布 $\mathcal{D}$ 中独立同分布采样得到的样本集. 记 $R(h)$ 与 $\hat{R}_S(h)$ 分别为假设 h 的泛化误差、在样本集 S 上的经验误差. $\mathcal{H}_S = \{h \in \mathcal{H} \mid \hat{R}_S(h) = 0\}$ 为在样本集 S 上经验误差为 0 的假设构成的集合. 对任意 $h_0 \in \mathcal{H}_S$, 证明:

$$\Pr\left[\sup_{h\in\mathcal{H}_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\epsilon}{2}\right] \geqslant \Pr\left[B(m,\epsilon) > \frac{m\epsilon}{2}\right]\Pr\left[R(h_0) > \epsilon\right],$$

其中 $B(m,\epsilon)$ 是服从参数为 (m,ϵ) 的二项分布的随机变量.

(2) **[15pts]** 设 $m\epsilon \geqslant 8$. 对任意 $h_0 \in \mathcal{H}_S$, 证明:

$$\Pr\left[R(h_0) > \epsilon\right] \leqslant 2\Pr\left[\sup_{h\in\mathcal{H}_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\epsilon}{2}\right] \tag{2.1}$$

(3) **[15pts]** 样本集 S 与 S' 可以视为将具有 $2m$ 个样本的样本集 T 均匀随机拆分为两个子集得到的. 此时, 式 (2.1) 右侧可改写为

$$\Pr\left[\sup_{h\in\mathcal{H}_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\epsilon}{2}\right] = \Pr_{T\sim\mathcal{D}^{2m}, T\to(S,S')}\left[\exists h \in \mathcal{H} : \hat{R}_S(h) = 0 \wedge \hat{R}_{S'}(h) > \frac{\epsilon}{2}\right].$$

令假设 h' 满足 $\hat{R}_T(h') > \frac{\epsilon}{4}$ 并令 $l > \frac{m\epsilon}{2}$ 为假设 h' 在样本集 T 上预测错误的样本总数. 证明所有 l 个犯错样本都在 S' 中的概率不超过 2^{-l}, 即

$$\Pr_{T\sim\mathcal{D}^{2m}, T\to(S,S')}\left[\hat{R}_S(h') = 0 \wedge \hat{R}_{S'}(h') > \frac{\epsilon}{2} \;\middle|\; \hat{R}_T(h') > \frac{\epsilon}{4}\right] \leqslant 2^{-l}.$$

(4) **[15pts]** 对任意假设 $h \in \mathcal{H}$, 证明

$$\Pr_{T\sim\mathcal{D}^{2m}, T\to(S,S')}\left[\hat{R}_S(h) = 0 \wedge \hat{R}_{S'}(h) > \frac{\epsilon}{2}\right] \leqslant 2^{-\frac{\epsilon m}{2}}.$$

(5) **[20pts]** 基于上述结论, 试给出并证明一个可分情形下的大概率泛化界. 并将这一泛化界与一般情形下的 $O(\sqrt{(m/d)^{-1}\log(m/d)})$ 进行对比.

Proof.

(1) 令 $h_0 \in \mathcal{H}_S$, 则

$$Pr\left[\sup_{h\in\mathcal{H}_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\epsilon}{2}\right] \geqslant Pr\left[\left|\hat{R}_S(h_0) - \hat{R}_{S'}(h_0)\right| > \frac{\epsilon}{2}\right]$$

由 $\mathcal{H}_S = \{h \in \mathcal{H} \mid \hat{R}_S(h) = 0\}$ 为在样本集 S 上经验误差为 0 的假设构成的集合，所以

$$Pr\left[\left|\hat{R}_S(h_0) - \hat{R}_{S'}(h_0)\right| > \frac{\epsilon}{2}\right] = Pr\left[\hat{R}_{S'}(h_0) > \frac{\epsilon}{2}\right]$$

由 $Pr[A] \geq Pr[A \wedge B] = Pr[A \mid B]Pr[B]$，所以

$$Pr\left[\hat{R}_{S'}(h_0) > \frac{\epsilon}{2}\right] \geq Pr\left[\hat{R}_{S'}(h_0) > \frac{\epsilon}{2} \mid R(h_0) > \epsilon\right] Pr[R(h_0) > \epsilon]$$

考虑 $Pr\left[\hat{R}_{S'}(h_0) > \frac{\epsilon}{2} \mid R(h_0) > \epsilon\right] Pr[R(h_0) > \epsilon]$ 的含义为在真实泛化误差至少为 ϵ 的条件下在 m 个样本上的经验误差至少为 $\epsilon/2$ 的概率，而当真实泛化误差为 ϵ 时，该概率即为事件 $B(m, \epsilon) > m\epsilon/2$ 的概率，因而

$$Pr\left[\hat{R}_{S'}(h_0) > \frac{\epsilon}{2} \mid R(h_0) > \epsilon\right] \geq Pr\left[B(m, \epsilon) > \frac{m\epsilon}{2}\right]$$

综上可得

$$Pr\left[\sup_{h \in \mathcal{H}_S} \left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\epsilon}{2}\right] \geqslant Pr\left[B(m, \epsilon) > \frac{m\epsilon}{2}\right] Pr[R(h_0) > \epsilon].$$

(2) 只需证明 $Pr\left[B(m, \epsilon) > \frac{m\epsilon}{2}\right] \geq \frac{1}{2}$ 即可，对于随机变量 $B(m, \epsilon)$，可知其均值为 $m\epsilon$，方差为 $m\epsilon(1-\epsilon)$。由 *Chebyshev* 不等式可知，

$$Pr\left[|B(m, \epsilon) - m\epsilon| \geq \frac{m\epsilon}{2}\right] \leq \frac{m\epsilon(1-\epsilon)}{(m\epsilon/2)^2} = \frac{4(1-\epsilon)}{m\epsilon}$$

因而

$$Pr\left[B(m, \epsilon) \leq \frac{m\epsilon}{2}\right] = Pr\left[-(B(m, \epsilon) - m\epsilon) \geq \frac{m\epsilon}{2}\right] \leq Pr\left[|B(m, \epsilon) - m\epsilon| \geq \frac{m\epsilon}{2}\right] \leq \frac{4(1-\epsilon)}{m\epsilon}$$

又因为 $m\epsilon \geq 8$，所以

$$Pr\left[B(m, \epsilon) \leq \frac{m\epsilon}{2}\right] \leq \frac{4(1-\epsilon)}{m\epsilon} \leq \frac{4}{m\epsilon} \leq \frac{1}{2}$$

所以

$$Pr\left[B(m, \epsilon) > \frac{m\epsilon}{2}\right] \geq \frac{1}{2}$$

代入 *(1)* 中的结论，可以得到

$$Pr[R(h_0) > \epsilon] \leqslant 2Pr\left[\sup_{h \in \mathcal{H}_S} \left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\epsilon}{2}\right]$$

(3) 在样本集 T 上预测错 l 个样本的组合数为 $\binom{2m}{l}$，所有 l 个犯错样本都在 S' 的组合数为 $\binom{m}{l}$，因而所有 l 个犯错样本都在 S' 中的概率为

$$\frac{\binom{m}{l}}{\binom{2m}{l}} = \frac{m!}{l!(m-l)!} \cdot \frac{l!(2m-l)!}{(2m)!} = \prod_{i=0}^{l-1} \frac{m-i}{2m-i} \leq \prod_{i=0}^{l-1} \frac{m-i}{2m-2i} \leq \frac{1}{2^l}$$

因而

$$Pr_{T \sim \mathcal{D}^{2m}, T \to (S, S')}\left[\hat{R}_S(h') = 0 \wedge \hat{R}_{S'}(h') > \frac{\epsilon}{2} \;\middle|\; \hat{R}_T(h') > \frac{\epsilon}{4}\right] \leqslant 2^{-l}.$$

(4)

$$\begin{aligned}
&Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\right]\\
=&Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\wedge\hat{R}_T(h)>\frac{\epsilon}{4}\right]\\
=&Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\mid\hat{R}_T(h)>\frac{\epsilon}{4}\right]Pr\left[\hat{R}_T(h)>\frac{\epsilon}{4}\right]\\
\leq&Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\mid\hat{R}_T(h)>\frac{\epsilon}{4}\right]\\
\leq&2^{-l}\leq 2^{-\frac{m\epsilon}{2}}
\end{aligned}$$

(5) (4) 中证明了任意假设 $h\in\mathcal{H}$，都有下式成立，

$$Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\right]\leqslant 2^{-\frac{\epsilon m}{2}}.$$

那么根据联合界不等式，存在假设 $h\in\mathcal{H}$ 使得该式成立的概率

$$Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\exists h\in\mathcal{H}:\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\right]\leq\Pi_{\mathcal{H}}(2m)2^{-\frac{m\epsilon}{2}}$$

联合 *(1)∼(4)* 的结论，对于假设 $h\in\mathcal{H}$ 满足存在数据集 $S\sim\mathcal{D}^m$ 使得 $\widehat{R}_S(h)=0$，取任意 $\epsilon>0$ 使得 $m\epsilon\geq 8$，有

$$\begin{aligned}
&Pr[R(h)>\epsilon]\\
\leqslant&2Pr\left[\sup_{h\in\mathcal{H}_S}\left|\hat{R}_S(h)-\hat{R}_{S'}(h)\right|>\frac{\epsilon}{2}\right]\\
=&2Pr_{T\sim\mathcal{D}^{2m},T\to(S,S')}\left[\exists h\in\mathcal{H}:\hat{R}_S(h)=0\wedge\hat{R}_{S'}(h)>\frac{\epsilon}{2}\right]\\
\leq&2\Pi_{\mathcal{H}}(2m)\cdot 2^{-\frac{m\epsilon}{2}}\\
\leq&2\left(\frac{2em}{d}\right)^d 2^{-\frac{m\epsilon}{2}}
\end{aligned}$$

令上式右边等于 δ 解得对应的 ϵ

$$\epsilon=\frac{1}{m}\Big(d\log\frac{2em}{d}+\log\frac{1}{\delta}+\log 2\Big)\frac{2}{\log 2}$$

即可分情况下的样本收敛率为 $O(\frac{\log(m/d)}{m/d})$,远快于一般情况下的收敛率 $O(\sqrt{(m/d)^{-1}\log(m/d)})$。