机器学习理论研究导引
作业三

魏沐昊 **********

2024 年 4 月 26 日

作业提交注意事项

(1) 本次作业提交截止时间为 **2024/05/01 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/f86c262c888f4685a11d/

(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;

(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 [100pts] VC Dimension and Generalization Bound

(1) **[50pts]** 试给出轴平行矩形假设空间基于 VC 维的泛化误差上界.

(2) **[50pts]** 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.

Solution.

(1). 轴平行矩形 *(Axis-Parallel Rectangle,APR)* 是平面 $\mathbb{R}^2$ 上四条边均与坐标轴平行的矩形区域, 易知轴平行矩形的假设空间是无限假设空间。由习题 *3.1.(1)* 可知轴平行矩形的假设空间的 *VC* 维为 *4*, 证明如下: 对于假设空间中的任意一个轴平行矩形, 总能找到四个点位于轴平行矩形的内部 *(*不论是任意相同 *label* 两点的连线构成矩形的边或者构成矩形的对角线*)*, 然而存在一个大小为 *5* 的样本集, 假设其中四个正类样本可恰好构成一个轴平行矩形而负类样本点位于该矩形之中, 则无论如何都无法再假设空间中找到一个合适的轴平行矩形将该样本分开。因此, 轴平行矩形的假设空间的 *VC* 维为 *4*。

由定理 *4.3* 可得, 对轴平行矩形的假设空间 $\mathcal{H}$, 对 $m>d=4$ 和 $0<\delta<1$ 有:

$$P(|E(h)-\hat{E}(h)|\leq\sqrt{\frac{32\log\frac{em}{2}+8\log\frac{4}{\delta}}{m}})\geq 1-\delta$$

此时 $\epsilon=\sqrt{\frac{32\log\frac{em}{2}+8\log\frac{4}{\delta}}{m}}$, 忽略 δ 和常数项可得:

$$E(h)\leq\hat{E}(h)+O(\sqrt{\frac{\log m}{m}})$$

以至少 $1-\delta$ 的概率成立。

(2). 由书中式 *(2.23)* 可得:

$$\epsilon\geq\frac{4}{m}\log\frac{4}{\delta}$$

此时收敛速度为 $O(\frac{1}{m})$。

可以发现, 基于 *VC* 维的泛化误差上界比 *(2.23)* 更 “松”, 收敛速度也更慢, 原因是:*VC* 维的定义是分布无关的, 因此基于 *VC* 维的分析结果是分布无关的、数据独立的, 也就是说对于任意分布都成立, 这使得基于 *VC* 维的分析结果通常具有一定的 “普适性”; 但另一方面, 由于没有考虑数据本身, 基于 *VC* 维的分析结果通常比较 “松”。而书中式 *(2.23)* 是针对于轴平行矩形这一特定问题进行分析的, 因此误差上界更 “紧”。