机器学习理论研究导引
作业二

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 **2024/04/17 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/e1e2127e038245b0b445/

(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 [100pts] 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) **[50pts]** 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) **[50pts]** 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. 考虑如下数据集 $D = \{x_1, x_2, \dots, x_m\} \subset \mathcal{X}$, 其中 $x_i \leq x_j, \forall i, j \in [m], i \leq j$。对于上述区间，仅有一个样本输出为 *1* 的情况有 m 种，此时对于任意的 $x_i \in D$, 该区间为 $a \in (x_{i-1}, x_i], b \in [x_i, x_{i+1})$。同理可得, 仅有两个样本输出为 *2* 的情况有 $(m-1)$ 种,...; 有 m 个样本输出为 *1* 的情况有 *1* 种，此时 $a \leq x_1, b \geq x_m$; 没有样本输出为 *1* 的情况有 *1* 种。则 $\mathcal{H}$ 的增长函数为:

$$
\begin{aligned}
\Pi_{\mathcal{H}}(m) &= \max_{|D|=m} |\mathcal{H}_{|D}| && (1.1)\\
&= \max_{\{x_1, x_2, \dots, x_m\} \in \mathcal{X}} |\{(h(x_1, \dots, h(x_m)))|h \in \mathcal{H}\}| && (1.2)\\
&= m + (m-1) + \cdots + 1 + 1 && (1.3)\\
&= \frac{m^2 + m + 2}{2} && (1.4)
\end{aligned}
$$

2. 由教材上的说明及上题中假设空间的构造可得，令 $D = \{1, 2\}$, 显然 $\mathcal{H}$ 能打散 D, 因此 $VC(\mathcal{H}) \geq 2$。对于任意大小为 *3* 的数据集 $D' = \{x_1, x_2, x_3\}$。若令 $x_1 < x_2 < x_3$ 则分类结果 *(+1,-1,+1)* 必不可能实现，因为当 $h_{a,b}(x_1) = +1$ 且 $h_{a,b}(x_3) = +1$ 时，$h_{a,b}(x_2)$ 必为 *+1*。因此，$\mathcal{H}$ 不能打散任何大小为 *3* 的数据集。于是根据 VC 维的定义可得 $VC(\mathcal{H}) = 2$。
根据引理 *3.1* 可得:

$$
\begin{aligned}
\Pi_{\mathcal{H}}(m) &\leq \sum_{i=0}^{2} \binom{m}{i} && (1.5)\\
&= C_m^0 + C_m^1 + C_m^2 && (1.6)\\
&= 1 + m + \frac{m(m-1)}{2} && (1.7)\\
&= \frac{m^2 + m + 2}{2} && (1.8)
\end{aligned}
$$

由上题结论可得，存在 VC 维为 *2* 的函数空间 $\mathcal{H}$ 使得 $\Pi_{\mathcal{H}}(m) = \sum_{i=0}^{d} \binom{m}{i}$, 即对于任意 VC 维 d, *Sauer* 引理是是紧的。