
作业 2：学习理论

清华大学软件学院
机器学习, 2023 年秋季学期

1 介绍

本次作业需要提交说明文档 (PDF 形式)。注意事项如下:

- 本次作业总分为 110 分, 若得分超过 100 分, 则按照 100 分截断。
- 作业按点给分, 因此请在说明文档中按点回答, 方便助教批改。
- 友情提示: 即使无法完成某一小题, 该题结论也可以作为后面小题的条件。
- 对于证明题而言, 需要说清楚重要 (不) 等式或引理的名字, 例如 “利用 sup 的次可加性”, “利用 2.3 题的结论” 等。如果手写作业请务必保证字迹清晰。
- 不要使用他人的作业, 也不要向他人公开自己的作业, 否则处罚很严厉, 会扣至-100 (倒扣本次作业的全部分值)。发现疑似抄袭将采用口试等方式进行查证。
- 统一文件的命名: {学号}_{姓名}_hw2.pdf

2 概率近似正确 (Probably Approximately Correct) (25pt)

课件给出了 PAC 学习的一般框架, 本题会介绍一个具体的实例——同心圆学习问题, 来帮助你更好地理解 PAC 学习理论。

如图1(a)所示, 同心圆学习问题的输入空间是二维平面上的所有点, 即 $x \in \mathcal{X} = \mathbb{R}^2$; 标签 $y \in \mathcal{Y} = \{0, 1\}$, 即平面上的点被分成正例 ($y = 1$) 或负例 ($y = 0$)。大小为 n 的训练样本集 $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ 由分布 $D_{\mathcal{X} \times \mathcal{Y}}$ (以下简记为 D) 独立同分布 (i.i.d.) 采样生成。具体地, x 独立同分布服从 $D_{\mathcal{X}}$, 标签 y 关于输入 x 的条件分布由某个未知的、以原点为圆心的同心圆半径 r 决定: 记

$$C_r = \{x \in \mathcal{X} \mid \|x\|_2 \leq r\},$$

则 $\mathbb{P}[y = 1|x] = \mathbb{I}[x \in C_r]$, 即所有正例点必然落在以原点为圆心, 半径为 r 的圆的内部或边界上, 而所有负例点必然落在该圆外部。

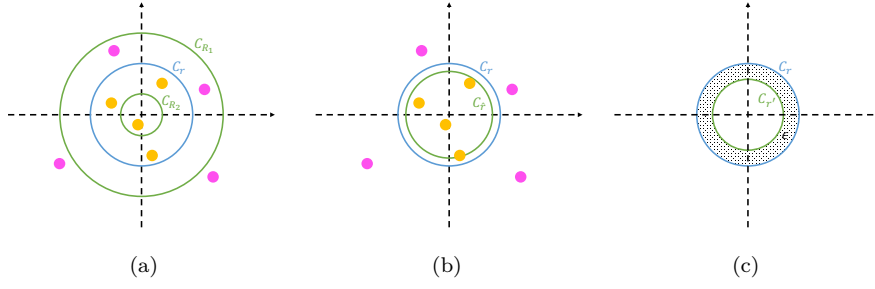


图 1: **(a)** 目标同心圆 C_r 及两种可能的同心圆样例 C_{R_1} 和 C_{R_2} 。圆点表示训练样本点，黄色的圆点表示正例，粉色的圆点表示负例。**(b)** 算法返回的同心圆 $C_{\hat{r}}$ 必然在 C_r 内部。**(c)** 在 C_r 的内部取同心圆 $C_{r'}$ ，使得样本落在圆环区域的概率为 ϵ 。

该学习问题定义为：给定训练集 $\mathcal{D}_n \sim D^n$ ，找到一个期望误差

$$\mathcal{E}(\hat{r}) = \mathbb{E}_{x \sim D_{\mathcal{X}}} [\mathbb{I}[\|x\|_2 \leq r] \neq \mathbb{I}[\|x\|_2 \leq \hat{r}]]$$

足够小的半径 \hat{r} 。从图1(a)中可以看出， \hat{r} 的误差 $\mathcal{E}(\hat{r})$ 对应的区域为 C_r 与 $C_{\hat{r}}$ 之间的圆环部分。接下来我们会逐步证明同心圆学习问题是 PAC-可学习的。

基于对数据生成分布 D 的先验知识，我们令假设空间 \mathcal{H} 是 \mathbb{R}^2 中以原点为圆心的同心圆的集合： $\mathcal{H} = \{h_{\xi} \mid \xi \geq 0\}$, $h_{\xi}(x) = \mathbb{I}[x \in C_{\xi}]$ 。我们给出以下学习算法 \mathcal{A} ：给定大小为 n 的训练集 \mathcal{D}_n ，返回包含 \mathcal{D}_n 中所有正例点的半径最小的假设，即相应的半径为

$$\hat{r} = \max_{(x,y) \in \mathcal{D}_n, y=1} \|x\|_2$$

(见图1(b))。显然，该算法必然满足 $\hat{r} \leq r$ ，误差 $\mathcal{E}_D(\hat{r})$ 对应的区域必定在圆 r 内。

固定 $\epsilon > 0$ ，不妨假设 $\mathbb{P}[C_r] > \epsilon$ (当 $\mathbb{P}[C_r] \leq \epsilon$ 时，很容易证明 PAC-可学习的条件)。在 C_r 的内部，构造一个新的同心圆 $C_{r'}$ ，使得 $\mathbb{P}[C_r \setminus C_{r'}] = \epsilon$ ，即 C_r 与 $C_{r'}$ 所夹的区域的概率为 ϵ 。

1. 证明误差 $\mathcal{E}_D(\hat{r})$ 大于 ϵ 当且仅当 $\hat{r} < r'$ 。(5pt)

2. 证明：

$$\mathbb{P}_{\mathcal{D}_n} [\mathcal{E}_D(\hat{r}_{\mathcal{D}_n}) > \epsilon] \leq \exp(-n\epsilon)$$

提示：对于任意 $x \in \mathbb{R}$, $1 - x \leq e^{-x}$ 成立。(5pt)

3. 证明：该问题是 PAC-可学习的。提示：用 PAC-可学习的定义。(5pt)

实际当中的**数据分布**都是存在噪音的。考虑以下情况：平面上的所有的负例点 ($y = 0$) 的标签都保持不变 $y' = 0$ ，所有的正例点 ($y = 1$) 的标签以概率 $\eta \in (0, \frac{1}{2})$ 变成 $y' = 0$ ，以 $1 - \eta$ 的概率仍然为 $y' = 1$ 。记这样的分布为 D_η 。对于学习算法而言， η 是未知的。假设其上界 η' 是已知的，即 $\eta \leq \eta' \leq 1/2$ 。学习算法依然返回包含 \mathcal{D}_n 中所有正例点 ($y' = 1$) 的最小的同心圆。

4. 当存在噪音时，求 $\mathcal{E}_{D_\eta}(r)$ 和 $\mathcal{E}_{D_\eta}(r')$ (用 $\mathbb{P}(C_r), \eta, \epsilon$ 表示)。(5pt)

5. 给出当存在噪音时， $\mathbb{P}_{\mathcal{D}_n \sim D_\eta^n} [\mathcal{E}(\hat{r}) - \mathcal{E}(r) > \epsilon]$ 的上界，并证明此时该问题依然是 PAC-可学习的。(5pt)

3 没有免费午餐定理 (No-Free-Lunch Theorem) (20pts)

在第一堂课中，我们直观地介绍了 No-Free-Lunch Theorem，本题将从**泛化理论**的角度证明此定理。定理内容为：考虑输入域 \mathcal{X} 上基于 01-loss 的二分类问题。对任意的 $n < |\mathcal{X}|/2$ 及任意的学习算法 $A: \mathcal{D}_n \mapsto h_{\mathcal{D}_n}$ ，总存在一个 $\mathcal{X} \times \mathcal{Y}$ 上的分布 D ，使得：

- 存在一个标签函数 $f: \mathcal{X} \rightarrow \{0, 1\}$ ，满足 $L_D(f) = 0$ ；
- 对 \mathcal{D}_n 的选择，有至少 $1/7$ 的概率，使得 $L(A(\mathcal{D}_n)) = L(h_{\mathcal{D}_n}) \geq 1/8$ 。

也即，无论学习算法如何，总存在一个“棘手”的分布，使得该学习算法总有较大（至少 $1/7$ ）的概率学到较差（ $L \geq 1/8$ ）的假设函数。出于简化，我们首先固定下来一个大小为 $2n$ 的 \mathcal{X} 的子集 C ，并将输入限制在 C 上进行研究。直观上我们可以感受到，由于学习算法只能观察到 C 中一半的样本，因此学到的假设 $A(\mathcal{D}_n)$ 可能会与标签函数 f 在**未采样到**的样本上发生矛盾。回答以下问题：

1. 记所有可能的标签函数 $f: C \rightarrow \{0, 1\}$ 为 f_1, f_2, \dots, f_T ，并定义分布 $D^{(i)}$ 满足样本 $(x, y) \in C \times \{0, 1\}$ 的概率为：

$$\mathbb{P}_{D^{(i)}}[(x, y)] = \begin{cases} \frac{1}{2n}, & \text{if } y = f_i(x) \\ 0, & \text{otherwise} \end{cases}$$

试求 $L_{D^{(i)}}(f_i)$ 。(1pt)

2. 考虑从 $D^{(i)}$ 中采样训练集的过程，容易发现这一过程只和输入的采样有关。设从 C 有放回地依次中采样 n 个元素组成的所有可能序列为 S_1, S_2, \dots, S_k ，并在标签函数 f_i 下得到数据

集 $S_1^{(i)}, S_2^{(i)}, \dots, S_k^{(i)}$ 。即，记 $S_j = (S_{j,1}, S_{j,2}, \dots, S_{j,n})$ ，则

$$S_j^{(i)} = \left((S_{j,1}, f_i(S_{j,1})), (S_{j,2}, f_i(S_{j,2})), \dots, (S_{j,n}, f_i(S_{j,n})) \right)。$$

令 $\mathcal{D}_n^{(i)} \sim (D^{(i)})^n$ ，试理解 $\mathcal{D}_n^{(i)}$ 和 $S_1^{(i)}, S_2^{(i)}, \dots, S_k^{(i)}$ 的关系，并证明：

$$\max_{1 \leq i \leq T} \mathbb{E}_{\mathcal{D}_n^{(i)}} [L_{D^{(i)}}(A(\mathcal{D}_n^{(i)}))] \geq \min_{1 \leq j \leq k} \frac{1}{T} \sum_{i=1}^T L_{D^{(i)}}(A(S_j^{(i)}))。$$

提示：注意到最大值 \geq 平均值 \geq 最小值。(5pt)

3. 对于某个固定的 j ，记 v_1, v_2, \dots, v_p 为 C 中未在 S_j 中出现的输入元素，则显然有 $p \geq 2n - n = n$ 。试证明：

$$\frac{1}{T} \sum_{i=1}^T L_{D^{(i)}}(A(S_j^{(i)})) \geq \frac{1}{2pT} \sum_{r=1}^p \sum_{i=1}^T \mathbb{I}[A(S_j^{(i)})(v_r) \neq f_i(v_r)]。$$

提示：先证明 $L_{\mathcal{D}^i}(h) \geq \frac{1}{2p} \sum_{r=1}^p \mathbb{I}[h(v_r) \neq f_i(v_r)]$ 。(5pt)

4. 考虑到算法 A 只能观察到 S_j^i ，因此无论 v_1, v_2, \dots, v_r 的标签如何，算法 A 找到的假设函数 $A(S_j^i)$ 总是固定的。因此在固定的输入样本 S_j 之下，我们总能将所有可能的标签函数两两配对，使得每对函数 f_i 和 $f_{i'}$ 在 S_j 上的标签完全相同，但是在 v_1, v_2, \dots, v_r 上的标签完全相反。借此证明：

$$\max_{1 \leq i \leq T} \mathbb{E}_{\mathcal{D}_n^i} [L_{D^{(i)}}(A(\mathcal{D}_n^{(i)}))] \geq \frac{1}{4}。$$

提示：借助第 2, 3 题的结论。(5pt)

5. 试证明 No-Free-Lunch Theorem。

提示：对 $[0, 1]$ 中的随机变量 Z ，有 $\mathbb{P}[Z > a] = \frac{\mathbb{E}[Z] - a}{1 - a}$ 。无需证明此引理。(2pt)

6. 试说明：若 \mathcal{X} 为无限集，假设空间 \mathcal{H} 为 $\mathcal{X} \rightarrow \{0, 1\}$ 的所有函数，则该分类问题不总是 PAC-可学习的。(2pt)

4 类别分布的学习 (15pt)

往年考试题。现有 m 个类别, 在这些类别上共有 k 种类别分布, 第 i 种分布服从

$$\mathbb{P}_i[X = j] = p_{i,j},$$

其中 $\sum_{j=1}^m p_{i,j} = 1, \forall 1 \leq i \leq m$ 。现在对每种分布, 我们都独立同分布 (*i.i.d.*) 地采样了 n 条样本, 并基于这些样本作出每种分布的估计

$$\hat{p}_{i,j} = \frac{1}{n} \sum_{s=1}^n \mathbb{I}[X_i^{(s)} = j],$$

其中 $X_i^{(s)}$ 是从第 i 个分布中采样的第 s 条样本。令

$$\mathbf{p}_i = (p_{i,1}, p_{i,2}, \dots, p_{i,m}), \quad \hat{\mathbf{p}}_i = (\hat{p}_{i,1}, \hat{p}_{i,2}, \dots, \hat{p}_{i,m})$$

试证明:

1. $\forall \delta > 0$, 以至少 $1 - \delta$ 的概率,

$$\forall 1 \leq i \leq k, \quad \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1 \leq m \sqrt{\frac{1}{2n} \log \frac{2mk}{\delta}}.$$

提示: 使用集中不等式。(7pt)

2. $\forall \delta > 0$, 以至少 $1 - \delta$ 的概率,

$$\forall 1 \leq i \leq k, \quad \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_1 \leq \sqrt{\frac{2}{n} \log \frac{k \cdot 2^m}{\delta}}.$$

提示: 先注意到

$$\forall \mathbf{v} \in \mathbb{R}^m, \|\mathbf{v}\|_1 = \sup_{\mathbf{u} \in \{-1, +1\}^m} \mathbf{u}^\top \mathbf{v},$$

再使用集中不等式。(8pt)

5 有限假设空间的泛化界 (15pt)

令 \mathcal{H} 为一有限（或可数）假设空间： $h: \mathcal{X} \mapsto \{0, 1\}$ ，且 p 是 \mathcal{H} 上的概率测度，即

$$\sum_{h \in \mathcal{H}} p(h) = 1, \quad p(h) \geq 0, \quad \forall h \in \mathcal{H}$$

p 可以表示假设空间上的**先验概率**，即学习算法选择某一特定假设 h 的概率。

证明：对任意 $\delta > 0$ ，以至少 $1 - \delta$ 的概率，以下不等式成立：

$$\forall h \in \mathcal{H}, \quad \mathcal{E}(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + \sqrt{\frac{\log \frac{1}{p(h)} + \log \frac{2}{\delta}}{2n}}$$

并将以上误差界与课件上给出的误差界：

$$\forall h \in \mathcal{H}, \quad \mathcal{E}(h) \leq \hat{\mathcal{E}}_{\mathcal{D}_n}(h) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{2}{\delta}}{2n}}$$

进行比较，谈谈自己的理解。提示：对 Hoeffding 不等式得到的结果进行换元 $\delta' = p(h)\delta$ 。(15pt)

6 无限假设空间的泛化界 (35pt)

6.1 Rademacher 复杂度

固定 $n \geq 1$ ，对于从输入集 \mathcal{X} 映射到 \mathbb{R} 的假设函数集 \mathcal{H} 和 \mathcal{H}' ，证明 Rademacher 复杂度的以下性质：

1. 对假设空间的单调性：若 $\mathcal{H} \subseteq \mathcal{H}'$ ，则 $\mathcal{R}_n(\mathcal{H}) \leq \mathcal{R}_n(\mathcal{H}')$ 。(3pt)
2. 缩放： $\forall \alpha \in \mathbb{R}$ ， $\mathcal{R}_n(\alpha \mathcal{H}) = |\alpha| \mathcal{R}_n(\mathcal{H})$ ，其中 $\alpha \mathcal{H} = \{\alpha h \mid \forall h \in \mathcal{H}\}$ 。(4pt)
3. 线性组合： $\mathcal{R}_n(\mathcal{H} + \mathcal{H}') \leq \mathcal{R}_n(\mathcal{H}) + \mathcal{R}_n(\mathcal{H}')$ ，其中 $\mathcal{H} + \mathcal{H}' = \{h + h' \mid \forall h \in \mathcal{H}, \forall h' \in \mathcal{H}'\}$ 。(4pt)
4. 凸包： $\mathcal{R}_n(\text{convex-hull}(\mathcal{H})) = \mathcal{R}_n(\mathcal{H})$ ，其中

$$\text{convex-hull}(\mathcal{H}) = \left\{ \sum_{j=1}^k \lambda_j h_j \mid \forall k \in \mathbb{N}_+, \sum_{j=1}^k \lambda_j = 1, h_1, h_2, \dots, h_k \in \mathcal{H} \right\}。 \quad (4pt)$$

提示：只需借助定义证明经验 Rademacher 复杂度的上述性质。

6.2 增长函数

定义从 \mathbb{R} 映射到 $\{+1, -1\}$ 函数族 $\mathcal{H} = \{h_{[a,b]} \mid a, b \in \mathbb{R}\}$, 其中

$$h_{[a,b]}(x) = \begin{cases} +1 & \text{for } x \in [a, b], \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

试给出函数族 \mathcal{H} 的增长函数 $\Pi_{\mathcal{H}}(n)$, 并用其导出 $\mathcal{R}_n(\mathcal{H})$ 的上界。

提示 1: 使用组合数。

提示 2: 后一问可直接使用课件上结果。(10pt)

6.3 VC 维

1. 平面上所有轴对齐矩形¹构成的假设空间 $\mathcal{H} = \{\mathbb{I}[a \leq x_1 \leq b, c \leq x_2 \leq d] \mid \forall a \leq b, \forall c \leq d\}$ 的 VC 维是多少?

提示: 通过举例的方式给出下界, 通过证明的方式给出上界, 证明用文字和图形表述即可。
不考虑点共线的情况。(5pt)

2. $\mathcal{H} = \{\mathbb{I}[\sin(x+a) > 0] \mid \forall a \in \mathbb{R}\}$ 的 VC 维是多少?

提示: $\mathbb{I}[\sin(x+a) > 0]$ 是周期函数。(5pt)

¹边和坐标轴平行的矩形, 当样本在矩形内部时为标签 1, 否则为标签 0