
作业 2：学习理论

清华大学软件学院
机器学习, 2024 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**总分值为 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) (35pt)

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——轴对齐矩形学习问题，来帮助你更好地理解 PAC 学习理论。

如图1(a)所示，轴对齐矩形¹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 $D_{\mathcal{X} \times \mathcal{Y}}$ 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轴对齐矩形 R 决定： $\Pr[y = 1|x] = \mathbb{I}[x \in R]$ ，即所有正例点必然落在 R 内部，而所有负例点必然落在该矩形外部。为方便起见，下文我们将输入 $x \sim D_{\mathcal{X}}$ 落在任一矩形 R 内的概率记作 $\Pr[R]$ 。

¹边和坐标轴平行的矩形

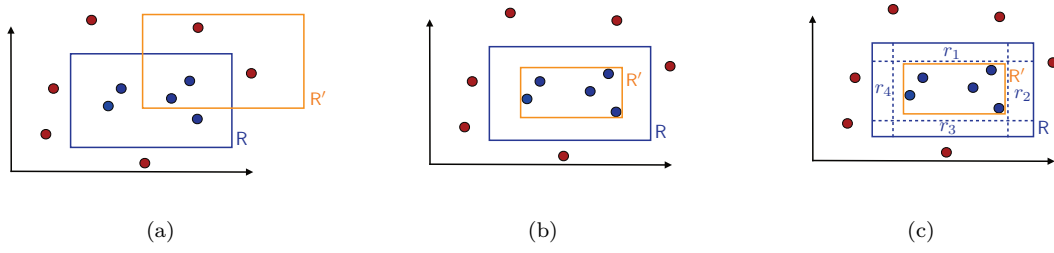


图 1: (a) 目标矩形 R 和可能的矩形 R' 。圆点表示训练样本点，蓝色的圆点表示正例，红色的圆点表示负例。(b) 算法返回的矩形 $R' = R_{\mathcal{D}_n}$ 必然在 R 内部。(c) 沿着矩形区域 R 的四条边，构造的四个新的矩形区域 r_1, r_2, r_3 和 r_4 。

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

$$\mathcal{E}(R') = \mathbb{E}_{x \sim D_{\mathcal{X}}} [\mathbb{I}[x \in R] \neq \mathbb{I}[x \in R']]$$

足够小的矩形 R' 。从图1(a)中可以看出， R' 的误差 $\mathcal{E}(R')$ 对应的区域包括：在矩形 R' 内但在矩形 R 外的区域（假阳性），在矩形 R 内但在矩形 R' 外（假阴性）。接下来我们会逐步证明轴对齐矩形学习问题是 PAC-可学习的。

基于对数据生成分布 $D_{\mathcal{X} \times \mathcal{Y}}$ 的先验知识，我们令假设空间 \mathcal{H} 是 \mathbb{R}^2 中所有轴对齐矩形的集合： $\mathcal{H} = \{R = [l, r] \times [b, t] \mid l \leq r, b \leq t\}$ 。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 中所有正例点的最小的轴对齐矩形 $R' = R_{\mathcal{D}_n}$ （见图1(b)）。显然，该算法不会产生假阳性，误差 $\mathcal{E}(R_{\mathcal{D}_n})$ 对应的区域必定在矩形 R 内。

固定 $\epsilon > 0$ ，不妨假设 $\Pr[R] > \epsilon$ （当 $\Pr[R] \leq \epsilon$ 时，很容易证明 PAC-可学习的条件）。沿着矩形区域 R 的四条边，分别构造四个新的矩形区域 r_1, r_2, r_3 和 r_4 （见图1(c)），使得其与 R 的一条边重合，且是概率 $\Pr[r_i] \geq \epsilon/4$ 的最小矩形。例如 $r_4 = [l, s_4] \times [b, t]$ 与矩形 R 的左边重合，满足 $s_4 = \inf \{s : \Pr[l, s] \times [b, t] \geq \epsilon/4\}$ 。 [l, s4] 表示 r4的水平范围，左边界是 l，右边界是 s4

1. 证明如果误差 $\mathcal{E}(R_{\mathcal{D}_n})$ 大于 ϵ ，那么 $R_{\mathcal{D}_n}$ 必然和某个 r_i 不重叠，即

$$\{\mathcal{E}(R_{\mathcal{D}_n}) > \epsilon\} \subset \bigcup_{i=1}^4 \{R_{\mathcal{D}_n} \cap r_i = \emptyset\}$$

提示：使用反证法。(5pt)

2. 计算矩形 $R_{\mathcal{D}_n}$ 和给定区域 r_i 没有任何重叠的概率

$$\Pr_{\mathcal{D}_n \sim D^n} [\{R_{\mathcal{D}_n} \cap r_i = \emptyset\}]$$

尽可能紧的上界。提示：用 ϵ 和 n 表示。(5pt)

²下文中简记为 D^n

3. 计算矩形 $R_{\mathcal{D}_n}$ 和 r_1, r_2, r_3 和 r_4 中的某一个没有重叠区域的概率

$$\Pr_{\mathcal{D}_n \sim D^n} [\cup_{i=1}^4 \{R_{\mathcal{D}_n} \cap r_i = \emptyset\}]$$

尽可能紧的上界。提示：使用 union bound。(5pt)

4. 证明：

$$\Pr_{\mathcal{D}_n \sim D^n} [\mathcal{E}(R_{\mathcal{D}_n}) > \epsilon] \leq 4 \exp(-n\epsilon/4)$$

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

5. 证明： \mathbf{R}^2 上所有轴对齐矩阵组成的假设空间 \mathcal{H} 是 PAC-可学习的。提示：用 PAC-可学习的定义。(5pt)

实际当中的数据集都是存在噪音的。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 $y' = 1$ 。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 ($y' = 1$) 的最小的轴对齐矩形。

6. 计算此时矩形 $R_{\mathcal{D}_n}$ 和给定区域 r_i 没有任何重叠的概率

$$\Pr_{\mathcal{D}_n \sim D^n} [\{R_{\mathcal{D}_n} \cap r_i = \emptyset\}]$$

尽可能紧的上界。提示：用 ϵ, η' 和 n 表示。(5pt)

7. 给出当存在噪音时， $\Pr_{\mathcal{D}_n \sim D^n} [\mathcal{E}(R_{\mathcal{D}_n}) > \epsilon]$ 的上界，并证明此时 \mathcal{H} 依然是 PAC-可学习的。(5pt)

3 集中不等式 (10pt)

取值在 $\{1, 2, \dots, K\}$ 的 X_1, X_2, \dots, X_n 是服从 Multinoulli 分布的独立随机变量, $\Pr[X_i = k] = p_k$, 参数 $\mathbf{p} = (p_1, \dots, p_K)$ 满足 $p_k \geq 0, \sum p_k = 1$ 。我们可以使用经验分布 $\hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_K)$:

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^n \mathbb{I}[X_i = k], \quad k = 1, 2, \dots, K$$

来估计 \mathbf{p} 。求证: 对任意 $\delta > 0$, 以至少 $1 - \delta$ 的概率,

$$\|\mathbf{p} - \hat{\mathbf{p}}\|_{\infty} = \max_k |p_k - \hat{p}_k| \leq \sqrt{\frac{1}{2n} \log \frac{2K}{\delta}}$$

提示: 使用 Hoeffding 不等式或者 McDiarmid 不等式。

4 有限假设空间的泛化界 (15pt)

假设 D 是定义在 \mathcal{X} 上的一个数据分布, $f: \mathcal{X} \rightarrow \{+1, -1\}$ 是打标函数 (labeling function)。定义分布 D 上的标签偏置 p_+ 为:

$$p_+ = \Pr_{x \sim D}[f(x) = +1].$$

定义 S 是从分布 D 独立同分布采样出的一个大小为 n 的数据集。我们可以使用 $\hat{p}_+ = \frac{1}{n} \sum_{x \in S} \mathbb{I}[f(x) = +1]$ 。求证: 对于任意 $\delta > 0$, $|p_+ - \hat{p}_+| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$ 成立的概率至少为 $1 - \delta$ 。

提示: 参考课件中对有限假设空间的泛化误差界的推导, 使用 Hoeffding 不等式进行证明即可。

5 无限假设空间的泛化界 (50pt)

5.1 Rademacher 复杂度

往年考试题。

1. 定义只包含两个函数的函数族 $\mathcal{G} = \{g_{-1}, g_{+1}\}$, 其中 g_{-1} 是恒取 -1 的常数函数, g_{+1} 是恒取 $+1$ 的常数函数, $S = (z_1, z_2, \dots, z_m)$ 是从分布 \mathcal{D} 采样的大小为 m 的数据集。请直接给出 \mathcal{G} 的 VC 维, 并推导 Rademacher 经验复杂度 $\hat{\mathcal{R}}_S(\mathcal{G})$ 的上界。(10pt)

提示: 对于凸函数 f , $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$ 。(Jensen 不等式)

2. 定义函数族 $\mathcal{G}: \mathcal{Z} \rightarrow [0, 1]$, 课件中对于 Rademacher 复杂度的上界推导是基于对函数 $\Phi(S) = \sup_{g \in \mathcal{G}} (\mathbb{E}_{\mathcal{D}}[g] - \hat{\mathbb{E}}_S[g])$ 的分析, 其中 $S = (z_1, z_2, \dots, z_m)$ 是从分布 \mathcal{D} 采样的大小为 m 的数据集, $\hat{\mathbb{E}}_S[g] = \frac{1}{m} \sum_{i=1}^m g(z_i)$ 。

在本题中你将再次使用 McDiarmid 不等式, 基于对函数 $\Psi(S) = \sup_{g \in \mathcal{G}} (\mathbb{E}_{\mathcal{D}}[g] - \hat{\mathbb{E}}_S[g] - 2\hat{\mathcal{R}}_S(\mathcal{G}))$ 的分析, 推导出 Rademacher 复杂度更紧的上界。要求给出完整证明, 以及使用概率近似正确框架表述最终的结论。(15pt)

提示: 课件上已经证明了 $\mathbb{E}_S[\Phi(S)] \leq 2\mathcal{R}_m(\mathcal{G})$, 该条件可直接使用。

5.2 增长函数

定义从 \mathbf{R} 映射到 $\{+1, -1\}$ 的阈值函数族 $\mathcal{H} = \{x \rightarrow \mathbb{I}[x \geq \theta] | \theta \in \mathbf{R}\} \cup \{x \rightarrow \mathbb{I}[x \leq \theta] | \theta \in \mathbf{R}\}$, 请求出函数族 \mathcal{H} 的增长函数, 并导出 $\mathcal{R}_n(\mathcal{H})$ 的上界。(10pt)

提示: 可以直接使用课件上的结论。

5.3 VC 维

与第2题中的轴对齐矩阵学习问题类似, 我们考虑凸 d 边形学习问题。设假设空间 $\mathcal{H} = \{\mathbb{I}[x \in P] | P \text{ is a convex } d\text{-gon}\}$ 是二维平面 \mathbf{R}^2 上所有凸 d 边形分类函数的集合。证明: \mathcal{H} 的 VC 维是 $2d + 1$ 。(不考虑数据中存在三点共线的情况)(15pt)

提示 1: 通过举例的方式给出下界, 再证明上界; 用文字和图形表述即可。

提示 2: 考虑点是否都排列在同一个凸多边形上, 进行分类讨论。