泛化界

PAC 学习框架导读 (上)

Jiaxuan Zou

西安交通大学数学与统计学院

2025 年 5 月 1 日

1. 一致情况
2. 不一致情况

1 一致情况

2 不一致情况

什么是 PAC 学习框架？

当我们设计从例子中学习的算法时，我们面临许多基本问题：什么可以被高效学习？哪些知识本质上难以掌握？PAC 学习框架为我们提供了一个理论框架来探讨这些问题。它帮助我们理解学习什么样的样本，什么样的模型才能成功学习，并评估它们的效率和准确性。

符号对照表

符号	说明
$\mathcal{X}$	所有可能的样本或实例的集合，也称为输入空间
$\mathcal{Y}$	所有可能的标签或目标值的集合
$c:\mathcal{X}\to\mathcal{Y}$	从输入空间到标签空间的映射，即概念
$\mathcal{C}$	我们希望学习的概念类，是一组概念
$S=(x_1,\ldots,x_m)$	从分布 D 中独立同分布抽取的样本
$c(x_1),\ldots,c(x_m)$	基于特定目标概念 $c\in\mathcal{C}$ 的标签
$h_S\in H$	从假设集 H 中选择的假设
$R(h)$	假设 h 的泛化误差，也称为真实误差

表 1: 符号说明

定义 2.1 泛化误差

给定一个假设 $h \in H$，一个目标概念 $c \in C$，和一个基础分布 D，h 的泛化误差或风险 $R(h)$ 定义为：

$$R(h) = \Pr_{x \sim D}[h(x) \neq c(x)] = \mathbb{E}_{x \sim D}\left[1_{h(x) \neq c(x)}\right],$$

其中 1_ω 是事件 ω 的指示函数。
泛化误差是一个假设对于未知数据的预测准确性的度量。由于分布 D 和目标概念 c 通常是未知的，泛化误差对学习者来说不是直接可访问的。然而，学习者可以在标记样本 S 上测量假设的经验误差。

定义 2.2 经验误差

给定一个假设 $h \in H$，一个目标概念 $c \in C$，和一个样本 $S = (x_1, \ldots, x_m)$，h 的经验误差或经验风险 $\hat{R}(h)$ 定义为

$$\hat{R}(h) = \frac{1}{m} \sum_{i=1}^{m} 1_{h(x_i) \neq c(x_i)}.$$

经验误差是 $h \in H$ 在样本 S 上的平均误差，而泛化误差是基于分布 D 的期望误差。在本章及后续章节中，我们将看到与这两个量相关的一些高概率保证。

对于固定的 $h \in H$，基于独立同分布样本 S 的经验误差的期望等于泛化误差：

$$\mathbb{E}[\hat{R}(h)] = R(h).$$

定义 2.2 经验误差

$$\begin{aligned}\mathbb{E}_{S\sim D^m}[\hat{R}(h)] &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{S\sim D^m}[1_{h(x_i)\neq c(x_i)}]\\ &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{x\sim D}[1_{h(x)\neq c(x)}]\\ &= R(h).\end{aligned}$$

对于样本 S 中的任何 x 都成立。因此，

$$\mathbb{E}_{S\sim D^m}[\hat{R}(h)] = \mathbb{E}_{x\sim D}[1_{h(x)\neq c(x)}] = R(h).$$

定义 2.3 PAC 学习

概念类 C 是 PAC 可学习的，如果存在算法 $\mathcal{A}$ 和多项式函数 $\text{poly}(\cdot,\cdot,\cdot)$，使得对于任何 $\epsilon > 0$ 和 $\delta > 0$，对于所有分布 D 和任何目标概念 $c \in C$，以下成立：

$$\Pr_{S\sim D^m}[R(h_S) \leq \epsilon] \geq 1-\delta \quad \text{for} \quad m \geq \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c)).$$

如果 $\mathcal{A}$ 运行时间也是多项式，则 C 是有效 PAC 可学习的。

PAC 学习的关键点

PAC 框架关注概念类的可学习性，而非单个概念。它不依赖于分布 D 的具体形式，且训练和测试样本都来自同一分布 D。这使得泛化成为可能。

示例 2.1 学习轴对齐矩形

考虑实例集为平面上的点，$\mathcal{X} = \mathbb{R}^2$，概念类 C 是所有位于 $\mathbb{R}^2$ 中的轴对齐矩形的集合。每个概念 c 是特定轴对齐矩形内的点集。学习问题包括使用标记的训练样本，以小误差确定目标轴对齐矩形。我们将展示轴对齐矩形的概念类是 PAC 可学习的。

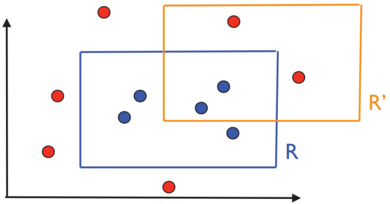

图 1: 问题示意图

算法描述

给定标记样本 S，算法 A 返回包含标记为 1 的点的最紧密的轴对齐矩形 $R' = R_S$。根据定义，R_S 不会产生任何假阳性，因为其点必须包含在目标概念 R 中。因此，R_S 的误差区域包含在 R 中。

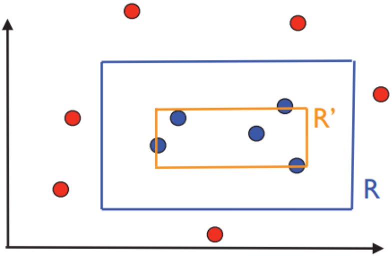

图 2: 算法返回的假设

误差分析

设 $R \in C$ 为目标概念。固定 $\epsilon > 0$。令 $\Pr[R_S]$ 表示由 R_S 定义的区域的概率质量，即随机抽取的点落入 R_S 的概率。由于算法的错误仅可能由于落入 R_S 的点造成，我们可以假设 $\Pr[R_S] > \epsilon$。由于 $\Pr[R_S] > \epsilon$，我们可以定义四个矩形区域 r_1, r_2, r_3, r_4 沿着 R_S 的边，每个区域的概率至少为 $\epsilon/4$。这些区域可以通过从空矩形开始，沿一边增加其大小直到其分布质量至少为 $\epsilon/4$ 来构建。

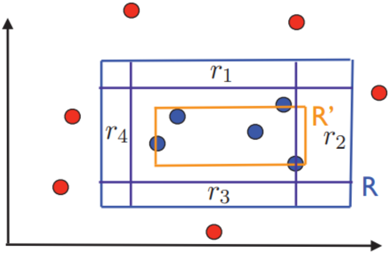

图 3: 算法返回的假设

概率分析

观察到如果 R_S 满足所有这些区域，则因为它是矩形，它将在每个区域中有一个边（几何论证）。其误差区域，即它未覆盖的 R 的区域，因此包含在这些区域中，并且不能有超过 ϵ 的概率质量。通过反证法，如果 $\mathcal{R}(R_S) > \epsilon$，则 R_S 必须至少错过一个区域 $r_i, i \in [1, 4]$。因此，我们可以写出

$$\Pr_{S \sim D^m}[\mathcal{R}(R_S) > \epsilon] \leq \Pr_{S \sim D^m}[\bigcup_{i=1}^{4}\{R_S \cap r_i = \emptyset\}]$$

$$\leq \sum_{i=1}^{4} \Pr_{S \sim D^m}[(R_S \cap r_i = \emptyset)] \quad \text{(并集界)}$$

$$\leq 4(1-\epsilon/4)^m \quad (\text{因为 } \Pr[r_i] > \epsilon/4)$$

$$\leq 4e^{-m\epsilon/4},$$

样本复杂度

对于任何 $\delta > 0$，为了确保 $\Pr_{S\sim D^m}[R(R_S) > \epsilon] \leq \delta$，我们可以施加

$$4e^{-m\epsilon/4} \leq \delta \Leftrightarrow m \geq \frac{4}{\epsilon}\log\frac{4}{\delta}.$$

因此，对于任何 $\epsilon > 0$ 和 $\delta > 0$，如果样本大小 m 大于 $\frac{4}{\epsilon}\log\frac{4}{\delta}$，则 $\Pr_{S\sim D^m}[R(R_S) > \epsilon] \leq 1-\delta$。此外，算法的计算成本是常数。这证明了轴对齐矩形的概念类是 PAC 可学习的，并且 PAC 学习轴对齐矩形的样本复杂度是 $O(\frac{1}{\epsilon}\log\frac{1}{\delta})$。

定理 2.1 泛化界

设 H 是从 $\mathcal{X}$ 到 $\mathcal{Y}$ 的有限函数集。算法 $\mathcal{A}$ 对于任何目标概念 $c \in H$ 和独立同分布样本 S 返回一致假设 h_S：$\hat{R}(h_S) = 0$。对于任何 $\epsilon, \delta > 0$，不等式 $\Pr_{S \sim D^m}[R(h_S) > \epsilon] \leq 1 - \delta$ 成立当且仅当

$$m \geq \frac{1}{\epsilon}\left(\log|H| + \log\frac{1}{\delta}\right).$$

泛化界限为：

$$R(h_S) \leq \frac{1}{m}\left(\log|H| + \log\frac{1}{\delta}\right).$$

证明

固定 $\epsilon > 0$。我们不知道算法 $\mathcal{A}$ 选择的一致假设 $h_S \in H$ 是哪一个。这个假设依赖于训练样本 S。因此，我们需要一个适用于所有一致假设的一致收敛界限，包括 h_S。我们将限制某个 $h \in H$ 一致且误差不超过 ϵ 的概率：

$$
\begin{aligned}
&\Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \\
&= \Pr\left[\bigvee_{h \in H} (h \in H, \hat{R}(h) = 0 \wedge R(h) > \epsilon)\right] \\
&\leq \sum_{h \in H} \Pr[\hat{R}(h) = 0 \wedge R(h) > \epsilon] \\
&\leq \sum_{h \in H} \Pr[\hat{R}(h) = 0 \mid R(h) > \epsilon].
\end{aligned}
$$

(条件概率的定义)

继续证明

现在，考虑任何假设 $h \in H$ 且 $R(h) > \epsilon$。那么，h 在独立同分布抽取的训练样本 S 上一致的概率，即在 S 上没有错误的概率，可以限制为：

$$\Pr[\hat{R}(h) = 0 \mid R(h) > \epsilon] \leq (1 - \epsilon)^m.$$

因此，我们有：

$$\sum_{h \in H} \Pr[\hat{R}(h) = 0 \mid R(h) > \epsilon] \leq |H|(1 - \epsilon)^m.$$

对于任何 $\delta > 0$，为了确保 $\Pr[\exists h \in H : \hat{R}(h) = 0 \wedge R(h) > \epsilon] \leq \delta$，我们可以施加

$$|H|(1 - \epsilon)^m \leq \delta \Leftrightarrow m \geq \frac{1}{\epsilon} \log \frac{|H|}{\delta}.$$

这证明了对于任何 $\epsilon > 0$ 和 $\delta > 0$，如果样本大小 m 大于 $\frac{1}{\epsilon} \log \frac{|H|}{\delta}$，则 $\Pr_{S \sim D^m}[R(h_S) > \epsilon] \leq 1 - \delta$。

定理 2.1 的含义

$$m \geq \frac{1}{\epsilon}\left(\log|H| + \log\frac{1}{\delta}\right).$$

泛化界限为：

$$R(h_S) \leq \frac{1}{m}\left(\log|H| + \log\frac{1}{\delta}\right).$$

该定理表明，当假设集 H 有限时，一致算法 $\mathcal{A}$ 是 PAC 学习算法。样本复杂度由 $1/\epsilon$ 和 $1/\delta$ 的多项式主导。泛化误差随着样本大小 m 递减，学习算法从更大的标记训练样本中受益。上界随着 $|H|$ 增加，但依赖性是对数的。

PAC 学习与不同概念类

我们现在使用定理 2.1 来分析各种概念类的 PAC 学习。

示例 2.2 布尔字面量的合取

考虑学习最多 n 个布尔字面量的合取的概念类 C_n。例如，对于 $n=4$，合取 $x_1 \wedge \overline{x}_2 \wedge x_4$ 是一个概念。正例之一：$(1,0,0,1)$。负例之一：$(1,0,0,0)$

我们可以使用一个简单的算法来找到一个一致的假设：对于每个正例 $(b_1,\ldots,b_n)$，如果 $b_i=1$，则排除 $\overline{x}_i$；如果 $b_i=0$，则排除 x_i。未被排除的字面量的合取即为一致假设。

0	I	I	0	I	I	+
0	I	I	I	I	I	+
0	0	I	I	0	I	-
0	I	I	I	I	I	+
I	0	0	I	I	0	-
0	I	0	0	I	I	+
0	I	?	?	I	I	

样本复杂度界限

由于 $|H| = 3^n$，根据定理 2.1，对于任何 $\epsilon > 0$ 和 $\delta > 0$，样本大小 m 应满足：

$$m \geq \frac{1}{\epsilon}\left(\log|H| + \log\frac{1}{\delta}\right)$$

代入 $|H| = 3^n$，得到：

$$m \geq \frac{1}{\epsilon}\left(n\log 3 + \log\frac{1}{\delta}\right)$$

这表明最多 n 个布尔字面量的合取类是 PAC 可学习的。计算复杂度也是多项式的，因为每个样本的训练成本是 $O(n)$。例如，对于 $\delta = 0.02, \epsilon = 0.1$，和 $n = 10$，需要至少 149 个样本来保证 99% 的准确度和 98% 的置信度。

不一致情况的保证

通常，可能不存在与训练样本一致的假设，尤其是在复杂学习问题中。然而，即使假设在训练样本上存在少量错误，也可能有用。本节将探讨这种情况下的学习保证。

Hoeffding 不等式

Hoeffding 不等式是概率论中用于估计随机变量偏离期望值的概率的重要工具。对于独立随机变量序列，它提供了如下界限：

$$\Pr\left(\left|\sum_{i=1}^{m} X_i - \sum_{i=1}^{m} \mu_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2s^2}\right)$$

其中 s^2 是变量方差的和，t 是任意正数。

推论 2.1

固定 $\epsilon > 0$ 并设 S 表示大小为 m 的独立同分布样本。对于任何假设 $h : X \to \{0, 1\}$，以下不等式成立：

$$\Pr_{S \sim D^m}\left[\left|\hat{R}(h) - R(h)\right| \geq \epsilon\right] \leq 2\exp(-2m\epsilon^2)$$

证明：由 Hoeffding 不等式可得证。■

单个假设的泛化界限

$$\Pr_{S \sim D^m}\left[\left|\hat{R}(h) - R(h)\right| \geq \epsilon\right] \leq 2\exp(-2m\epsilon^2)$$

将上式右侧设置为等于 δ 并求解 ϵ，我们立即得到单个假设的以下界限：

$$m \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$$

这意味着，为了确保单个假设的经验误差和泛化误差之间的差异不超过 ϵ 的概率至少为 $1-\delta$，我们需要至少 m 个样本。

推论 2.2 单个假设的泛化界限

对于单个假设 $h : \mathcal{X} \to \{0, 1\}$ 和任意 $\delta > 0$，以下不等式以至少 $1 - \delta$ 的概率成立：

$$R(h) \le \hat{R}(h) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

这个推论表明，单个假设的泛化误差可以通过增加样本数量 m 来估计和控制。

定理 2.2：有限假设集的泛化界（不一致情况）

设 H 是一个有限的假设集。对于任意 $\delta > 0$，以至少 $1-\delta$ 的概率，以下不等式成立：

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}.$$

这个定理为我们提供了在有限假设集情况下，经验风险和真实风险之间的界限。

定理 2.2 的证明

证明思路是利用并集界（union bound）和 Hoeffding 不等式。设 $h_1, \ldots, h_{|H|}$ 是 H 的元素。对每个假设应用推论 2.2，我们得到：

$$\Pr\left[\exists h \in H \mid |\hat{R}(h) - R(h)| > \epsilon\right] \leq \sum_{h \in H} \Pr\left[|\hat{R}(h) - R(h)| > \epsilon\right].$$

$$\leq 2|H| \exp(-2m\epsilon^2).$$

这个不等式表明，所有假设的经验风险和真实风险之间的偏差超过 ϵ 的概率被限制在这个表达式内。

设置 δ

为了使上述概率不超过 δ，我们设置：

$$2|H|\exp(-2m\epsilon^2) \leq \delta.$$

解这个不等式，我们可以得到 ϵ 的表达式：

$$\epsilon \geq \sqrt{\frac{\log|H| + \log\frac{2}{\delta}}{2m}}.$$

这完成了定理的证明。

定理 2.2 的含义

$$\forall h \in H, \quad R(h) \leq \hat{R}(h) + \sqrt{\frac{\log |H| + \log \frac{2}{\delta}}{2m}}.$$

定理 2.2 告诉我们，对于有限的假设集 H，真实风险 $R(h)$ 可以通过经验风险 $\hat{R}(h)$ 加上一个与样本大小 m、假设集大小 $|H|$ 和置信度 δ 相关的项来近似。这个结果对于理解和设计学习算法非常重要。

这可以看作是奥卡姆剃刀原则的一个实例：没有必要，不应增加复杂性，换句话说，最简单的解释是最好的。在这种情况下，可以表达为：在其他条件相同的情况下，一个更简单（更小）的假设集更好。