第 4 章 学习的可行性

4.1 学习是不可能的吗？

简单二分类问题

定义 $\mathcal{X}=\{0,1\}^3, \mathcal{Y}=\{\circ,\times\}$，可以枚举所有候选 f 作为 $\mathcal{H}$。

问题 从 $\mathcal{H}$ 中选择 $g\in\mathcal{H}$ 使得对所有 $g(\mathbf{x}_n)=y_n$（类似于 PLA），$g\approx f$？

没有免费的午餐

- 在 $\mathcal{D}$ 内 $g\approx f$：当然！
- 在 $\mathcal{D}$ 外 $g\approx f$：不！（但那才是我们真正想要的！）

从 $\mathcal{D}$ 学习（推断 $\mathcal{D}$ 外的东西）注定失败，如果有任何“未知”的 f 可能发生。

$\mathbf{x}$	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

图 4.1.1: 简单分类问题示意图

定理 4.1.1 (无免费午餐定理（No Free Lunch）)

在机器学习中，不存在单一算法能够在所有问题上都达到最优性能。即对于任何给定的学习算法，都存在至少一个问题，在该问题上该算法的性能不会优于随机猜测。 ♡

4.2 概率论提供了关键性的解决办法

推断未知事物：推断橙色概率

罐子 假设橙色概率 = μ，绿色概率 = $1-\mu$，其中 μ 未知

样本 独立抽取 N 个弹珠，橙色比例 = ν，绿色比例 = $1-\nu$，现在 ν 已知

问题 样本 ν 能否说明样本外 μ 的情况？

可能与概率

样本 ν 是否能说明样本外 μ 的情况？

- 可能不能：样本可能大部分是绿色而罐子大部分是橙色
- 可能能：样本 ν 可能接近未知的 μ

引理 4.2.1 (Hoeffding 引理)

设随机变量 X 满足 $\mathbb{E}[X]=0$，且几乎处处 $X\in[a,b]$，则对于任意实数 $\lambda\in\mathbb{R}$，有如下不等式成立：

$$\mathbb{E}\left[e^{\lambda X}\right]\leq\exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

♡

证明 考虑任意满足 $X\in[a,b]$ 且 $\mathbb{E}[X]=0$ 的随机变量。令 $c=b-a$。

我们令函数 $\phi(\lambda)=\mathbb{E}[e^{\lambda X}]$，其为 X 的矩母函数。由于 $e^{\lambda x}$ 是关于 x 的凸函数，且 $X\in[a,b]$，我们可以使用凸函数的最大值原理：在固定期望的前提下，$\mathbb{E}[e^{\lambda X}]$ 在 X 取两点分布时最大。

因此，考虑如下构造的两个点的分布：

$$X=\begin{cases}a, & \text{概率}p=\frac{b}{b-a}\\ b, & \text{概率}1-p=\frac{-a}{b-a}\end{cases}\quad\Rightarrow\mathbb{E}[X]=pa+(1-p)b=0$$

于是有：

$$\mathbb{E}[e^{\lambda X}]\leq pe^{\lambda a}+(1-p)e^{\lambda b}=\frac{b}{b-a}e^{\lambda a}+\frac{-a}{b-a}e^{\lambda b}$$

令 $c=b-a$，将 $a=-\frac{c}{2},b=\frac{c}{2}$（无损一般性，平移不影响证明），则：

$$\mathbb{E}[e^{\lambda X}]\leq\frac{1}{2}e^{-\lambda c/2}+\frac{1}{2}e^{\lambda c/2}=\cosh\left(\frac{\lambda c}{2}\right)$$

由 $\cosh(x)\leq\exp\left(\frac{x^2}{2}\right)$ 对任意实数 x 成立，因此：

$$\mathbb{E}[e^{\lambda X}]\leq\cosh\left(\frac{\lambda c}{2}\right)\leq\exp\left(\frac{\lambda^2c^2}{8}\right)$$

即：

$$\mathbb{E}[e^{\lambda X}]\leq\exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

■

定理 4.2.1 (霍夫丁不等式)

设罐中橙色概率为 μ，样本容量为 N，样本中橙色比例为 ν。则对任意 $\varepsilon>0$ 与 $N>0$，

$$\mathbb{P}\left[|\nu-\mu|>\varepsilon\right]\leq2\exp(-2\varepsilon^2N).$$

该不等式适用于弹珠、硬币、民调等场景，且与 μ 无关，无需先验知识。当 N 足够大时，样本比例 ν 以高概率接近真值 μ，即

$$\nu\approx\mu\quad\text{（Probably Approximately Correct）.}$$

♡

证明 设 $X_1,X_2,\ldots,X_n$ 是独立随机变量，且对每个 i，有 $X_i\in[a_i,b_i]$。定义 $S_n=\sum_{i=1}^{n}X_i$，我们考虑尾部概率：

$$\mathbb{P}(S_n-\mathbb{E}[S_n]\geq t)$$

对任意 $\lambda>0$，由 Chernoff 结合 Markov 不等式 $\mathbb{P}(X\geq a)\leq\frac{\mathbb{E}[X]}{a}$ 可得：

$$\mathbb{P}(S_n-\mathbb{E}[S_n]\geq t)=\mathbb{P}(e^{\lambda(S_n-\mathbb{E}[S_n])}\geq e^{\lambda t})\leq\frac{\mathbb{E}[e^{\lambda(S_n-\mathbb{E}[S_n])}]}{e^{\lambda t}}$$

由于 X_i 独立：

$$\mathbb{E}[e^{\lambda(S_n-\mathbb{E}[S_n])}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda(X_i-\mathbb{E}[X_i])}]$$

接下来对每个 i，应用 Hoeffding 引理：若 $X \in [a,b]$，则对任意 $\lambda \in \mathbb{R}$，有

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

因此，

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge t) \le \exp\left(-\lambda t + \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i-a_i)^2\right)$$

最小化右边对 λ，令导数为零得最优值 $\lambda = \frac{4t}{\sum_{i=1}^{n}(b_i-a_i)^2}$，代入得：

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \ge t) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

将 $t = n\epsilon$ 得到：

$$\mathbb{P}\left(\bar{X}_n - \mathbb{E}[\bar{X}_n] \ge \epsilon\right) \le \exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}\right)$$

设 $X_1, X_2, \ldots, X_N$ 是独立同分布随机变量，且每个变量取值范围为 $[0,1]$。定义样本均值为：

$$\nu = \frac{1}{N}\sum_{i=1}^{N} X_i, \quad \mu = \mathbb{E}[X_i]$$

根据 Hoeffding 不等式，对于任意 $\varepsilon > 0$，有：

$$\mathbb{P}\left(|\nu - \mu| \ge \varepsilon\right) \le \exp\left(-2\varepsilon^2 N\right)$$

■

4.3 与学习的关联

从弹珠到学习误差

设定 未知目标函数 $f : \mathcal{X} \to \mathcal{Y}$，固定假设 $h : \mathcal{X} \to \mathcal{Y}$，数据分布 P。

未知真误差

$$E_{\text{out}}(h) = \mathbb{P}_{\mathbf{x}\sim P}\big[h(\mathbf{x}) \ne f(\mathbf{x})\big].$$

已知样本误差 给定独立同分布训练集

$$\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}, \quad y_n = f(\mathbf{x}_n),$$

样本误差为

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\big[h(\mathbf{x}_n) \ne y_n\big].$$

命题 4.3.1 (形式保证)

对任一固定的假设 h 与足够大的样本规模 N，样本内误差 $E_{\text{in}}(h)$ 以高概率逼近样本外误差 $E_{\text{out}}(h)$（误差不超过 ε），即

$$\mathbb{P}\big[\big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \varepsilon\big] \leq 2\exp\big(-2\varepsilon^2 N\big).$$

该不等式与“罐子”类比相同，对所有 N 与 ε 成立，且不依赖未知的 $E_{\text{out}}(h)$，也无需知道 f 或分布 P。因此“$E_{\text{in}}(h) = E_{\text{out}}(h)$”是 PAC 正确的。

若进一步满足 $E_{\text{in}}(h) \approx E_{\text{out}}(h)$ 且 $E_{\text{in}}(h)$ 很小，则可推出 $E_{\text{out}}(h)$ 亦很小，即 h 以高概率逼近 f 于分布 P。

单一假设的验证与学习结论

验证 对任意固定假设 h，当数据量足够大时，

$$E_{\text{in}}(h) \approx E_{\text{out}}(h).$$

能否宣称“学得不错” $(g \approx f)$?

是！ 若算法 $\mathcal{A}$ 只考虑这一个 h 且 $E_{\text{in}}(h)$ 很小，则 $g = h$ 满足 PAC 正确性。

否！ 实际上 $\mathcal{A}$ 需在整个假设集 $\mathcal{H}$ 中选 g（如 PLA），而大多数固定 h 的 $E_{\text{in}}(h)$ 并不小；若强制取该 h 为 g，则 $E_{\text{in}}(g)$ 大概率很大，无法保证 PAC。

结论 真正的学习要求算法在 $\mathcal{H}$ 中主动挑选 g，而非被迫接受某个固定 h。

笔记 [单一假设验证 vs 真正学习]

- 若固定一个假设 h，当样本数足够大时，由 Hoeffding 不等式可得：$E_{\text{in}}(h) \approx E_{\text{out}}(h)$，可认为 h 通过了“验证”。
- 但这并不意味着我们“学得好”，因为学习算法并非接受一个 h，而是从假设集 $\mathcal{H}$ 中选择最优的 g。
- 实际上，大多数 $h \in \mathcal{H}$ 的 $E_{\text{in}}(h)$ 都不会很小；如果算法仅凭训练误差小就选出 g，可能只是偶然“运气好”。
- 要想保证 $E_{\text{out}}(g)$ 也小，必须考虑整个 $\mathcal{H}$ 的容量，如使用 VC 维等工具对其复杂度进行控制。

结论：验证单一假设容易，真正学习困难。学习不仅要训练误差小，还要能推广到未知数据——这要求我们控制假设空间的复杂性，从而提升模型的泛化能力。

4.4 与真实学习的关联

命题 4.4.1 (坏样本与坏数据)

设固定假设 h 与样本 $\mathcal{D}$ 的大小为 N。若存在 $\mathcal{D}$ 使得样本误差 $E_{\text{in}}(h)$ 与真实误差 $E_{\text{out}}(h)$ 差距超过阈值 ε，则称 $\mathcal{D}$ 为 h 的坏样本。

i) 对单个 h，坏样本概率 $\mathbb{P}_{\mathcal{D}}\big[|E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon\big]$ 存在上界。

ii) 若假设集 $\mathcal{H} = \{h_1, \ldots, h_M\}$ 含 M 个假设，则存在某个 $h \in \mathcal{H}$ 出现坏样本的概率可通过联合界控制：

$$\mathbb{P}_{\mathcal{D}}\Big[\exists\, h \in \mathcal{H},\ |E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon\Big] \leq \sum_{m=1}^{M} \mathbb{P}_{\mathcal{D}}\big[|E_{\text{in}}(h_m) - E_{\text{out}}(h_m)| > \varepsilon\big].$$

iii) 当 M 增大时，上述上界随之增大；只有当 M 相对样本量 N 足够小时，算法 $\mathcal{A}$ 仍能高概率保证对所有 h 都有 $E_{\text{in}}(h) \approx E_{\text{out}}(h)$。

定理 4.4.1 (有限假设集的霍夫丁上界)

设假设集 $\mathcal{H} = \{h_1, \ldots, h_M\}$ 有限，样本 $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ 独立同分布且大小为 N。对任意 $\varepsilon > 0$，有

$$\mathbb{P}_{\mathcal{D}}\Big[\exists\, h \in \mathcal{H},\ \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \varepsilon\Big] \le 2M \exp\big(-2\varepsilon^2 N\big)\,.$$

该界对任意 M, N, ε 成立，且与未知的目标函数 f、分布 P 及所有 $E_{\text{out}}(h)$ 无关。因此，若算法 $\mathcal{A}$ 选取

$$g = \arg\min_{h\in\mathcal{H}} E_{\text{in}}(h),$$

则 $E_{\text{in}}(g) \approx E_{\text{out}}(g)$ 以高概率成立，即 PAC 成立。

笔记 [有限假设集与泛化能力]

- 对单个假设 h，Hoeffding 不等式可保证当样本足够大时，$E_{\text{in}}(h) \approx E_{\text{out}}(h)$，即“坏样本”的概率很小。
- 若学习算法从有限假设集 $\mathcal{H} = \{h_1, \ldots, h_M\}$ 中选出 $g = \arg\min_h E_{\text{in}}(h)$，需考虑整个假设集上出现“坏样本”的联合概率。
- 使用联合界（Union Bound）可得：出现某个 $h \in \mathcal{H}$ 使得 $|E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon$ 的概率上界为：

$$\mathbb{P}_{\mathcal{D}}\Big[\exists\, h \in \mathcal{H},\ \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \varepsilon\Big] \le 2M \exp\big(-2\varepsilon^2 N\big)\,.$$

- 上界表明，只有当假设集大小 M 相对样本量 N 不太大时，才能高概率地保证所有假设的泛化误差都不大。

结论：对于有限假设集，Hoeffding 不等式联合界为学习算法提供了 **PAC** 泛化保证。这说明控制假设空间规模是实现有效学习的关键步骤之一。

4.5 总结

笔记 [学习的可行性]

- 学习是不可能的吗？：在数据集 $\mathcal{D}$ 之外，绝对不存在“免费午餐”。
- 概率论提供了关键性的解决办法：在 $\mathcal{D}$ 之外，我们能做到“可能近似正确”（PAC）。
- 与学习的关联：若对某个固定的 h，经验误差 $E_{\text{in}}(h)$ 足够小，则可验证其性能。
- 与真实学习的关联：若假设空间 $|\mathcal{H}|$ 有限且最终选出的 g 满足 $E_{\text{in}}(g)$ 足够小，则学习可行。

笔记 [总体结论] 通过合理控制假设空间的规模，并利用概率不等式对训练误差与真实误差的差距进行界定，我们能够以高概率选出在整体数据分布上表现良好的假设，从而保证学习的可行性和泛化能力。

笔记 [精华弹幕] 我们很难学习到正确的结果，但是当我们的样本量足够大时，通过公式我们就知道样本与实际情况非常接近，也就能学到正确的经验，也就是说机器学习时有用的。