# 第 8 章 噪声与误差

## 8.1 噪声与概率目标

**定义 8.1.1 (机器学习中的噪声（Noise in Machine Learning）)**

在监督学习中，噪声（noise）泛指任何使得观测数据无法被目标模型完美重现的现象。维基百科将其归纳为两大类：

1) 随机噪声（**stochastic noise**）：由随机波动、测量误差或未建模的随机过程引起，表现为数据标签或特征值中的随机偏差。
2) 确定性噪声（**deterministic noise**）：当真实现象的复杂度超出模型表达能力时，数据中包含的、无法被模型捕获的系统性结构误差。

无论来源如何，这两种噪声都会导致模型试图去拟合其本无法建模的部分，从而引发过拟合。通常需要正则化、数据清洗或鲁棒算法来减缓其负面影响。♣

**定义 8.1.2 (概率弹珠模型（Probabilistic Marbles）)**

在 VC 理论中，数据被比喻为一袋弹珠：

- 样本：单个弹珠 $x \sim P(x)$ 从总体中 i.i.d. 抽取。
- 标签：
  - 确定性弹珠：颜色固定为 $f(x)$；误差事件为 $f(x) \neq h(x)$。
  - 概率弹珠（含噪声）：颜色随机，$y \sim P(y \mid x)$；误差事件为 $y \neq h(x)$。
- 共同点：无论 $x$ 还是 $(x, y)$，只要 i.i.d. 抽取，VC 界均可估计"橙色弹珠"(即泛化误差）的概率。♣

**命题 8.1.1 (学习目标的形式化)**

对任意输入 $x \in \mathcal{X}$，其目标分布 $P(y \mid x)$ 可分解为

$$P(y \mid x) = \begin{cases} P\big(f(x) \,|\, x\big) = 1 - \eta(x), \\ P\big(\bar{f}(x) \,|\, x\big) = \eta(x), \end{cases}$$

其中

- $f(x) \in \mathcal{Y}$ 为理想迷你目标，
- $\eta(x) \in [0, 1]$ 为翻转噪声水平。

则学习算法的目标可表述为：

$$\min_{h \in \mathcal{H}} \mathbb{E}_{x \sim P(x)} \Big[ \mathbb{P}_{y \sim P(y|x)} \big[ h(x) \neq y \big] \Big].$$

特别地，当 $\eta(x) \equiv 0$ 时，目标退化为

$$\min_{h \in \mathcal{H}} \mathbb{E}_{x \sim P(x)} \big[ \mathbb{I}\{h(x) \neq f(x)\} \big].$$

♠

## 8.2 误差度量

**定义 8.2.1 (误差度量（Error Measure）)**

最终假设 $g$ 与目标函数 $f$ 之间的误差度量 $E(g,f)$ 可按以下三种场景定义：

1) 样本外误差（**out-of-sample**）：

$$E_{\text{out}}(g)=\mathbb{E}_{x\sim P}\big[\mathbb{I}\{g(x)\neq f(x)\}\big].$$

2) 逐点误差（**pointwise**）：对单个输入 $x$ 计算

$$err(g,f,x)=\mathbb{I}\{g(x)\neq f(x)\}.$$

3) 分类误差（**classification**）：通常直接记为

$$0/1\text{ error}=\mathbb{I}\{\text{prediction}\neq\text{target}\},$$

也即逐点误差的特例。♣

**定义 8.2.2 (逐点误差度量（Pointwise Error Measure）)**

误差度量 $E(g,f)$ 常可写成逐点误差 $\text{err}(g(x),f(x))$ 的平均：

- 样本外误差

$$E_{\text{out}}(g)=\mathbb{E}_{x\sim P}\big[\text{err}(g(x),f(x))\big].$$

- 样本内误差

$$E_{\text{in}}(g)=\frac{1}{N}\sum_{n=1}^{N}\text{err}(g(x_n),f(x_n)).$$

两种常用的逐点误差

- **0/1** 误差（分类）

$$\text{err}(\hat{y},y)=\mathbb{I}[\hat{y}\neq y]\quad\text{只关心“对”或“错”。}$$

- 平方误差（回归）

$$\text{err}(\hat{y},y)=(\hat{y}-y)^2\quad\text{度量“偏离”程度。}$$

学习启示　不同 err 会给出不同的学习指引——分类器通常优化 0/1 误差或其凸代理，而回归器则优化平方误差或其变体。♣

**例题 8.1 理想迷你目标：噪声与误差的相互作用**

给定

$$P(y=1\mid x)=0.2,\quad P(y=2\mid x)=0.7,\quad P(y=3\mid x)=0.1,$$

及两种逐点误差度量：

- 0/1 误差：$\text{err}(\hat{y},y)=\mathbb{I}[\hat{y}\neq y]$；
- 平方误差：$\text{err}(\hat{y},y)=(\hat{y}-y)^2$。

**解答**　计算各候选输出 $\hat{y}$ 的期望误差

$$\mathbb{E}_{y\sim P(y|x)}[\text{err}(\hat{y},y)]:$$

| $\hat{y}$ | 0/1 误差 | 平方误差 |
|---|---|---|
| 1 | 0.8 | 1.1 |
| 2 | 0.3* | 0.3 |
| 3 | 0.9 | 1.5 |
| 1.9 | 1.0 | 0.29* |

**结论**

- 0/1 误差下的理想迷你目标：$f(x) = \arg\max_{y\in\mathcal{Y}} P(y|x)$。
- 平方误差下的理想迷你目标：$f(x) = \sum_{y\in\mathcal{Y}} y \cdot P(y|x)$。

证明 设给定输入 $x$ 的条件分布为 $P(y \mid x)$，$y \in \mathcal{Y}$。

（**1**）**0/1** 误差要最小化期望 0/1 误差

$$\mathbb{E}_{y\sim P(y|x)}\big[\mathbb{I}[\hat{y} \neq y]\big] = 1 - P(\hat{y} \mid x).$$

显然，当且仅当选择使 $P(y \mid x)$ 最大的标签作为预测值时上式最小，故

$$f(x) = \arg\max_{y\in\mathcal{Y}} P(y \mid x).$$

（**2**）平方误差要最小化期望平方误差

$$\mathbb{E}_{y\sim P(y|x)}\big[(\hat{y} - y)^2\big] = \sum_{y\in\mathcal{Y}} (y - \hat{y})^2 P(y \mid x).$$

对 $\hat{y}$ 求导并令导数为零：

$$\frac{\partial}{\partial \hat{y}} \sum_{y} (y - \hat{y})^2 P(y \mid x) = -2 \sum_{y} (y - \hat{y}) P(y \mid x) = 0.$$

解得

$$\hat{y} = \sum_{y\in\mathcal{Y}} y\, P(y \mid x),$$

即条件期望 $\mathbb{E}[y \mid x]$，因此

$$f(x) = \sum_{y\in\mathcal{Y}} y\, P(y \mid x).$$

■

## 8.3 算法级误差度量

**命题 8.3.1 (指纹验证中的误差度量选择)**

设指纹系统的输出为 $g \in \{-1, +1\}$，真实标签为 $y \in \{-1, +1\}$，其中

$$y = +1 : \text{真实用户}, \qquad y = -1 : \text{入侵者}.$$

定义两种错误类型：

- 假接受（**False Accept**）：$g = +1$，$y = -1$；
- 假拒绝（**False Reject**）：$g = -1$，$y = +1$。

若采用 0/1 误差

$$\mathrm{err}(g, y) = \mathbb{I}[g \neq y],$$

则假接受与假拒绝被同等惩罚；实际应用需引入代价敏感误差

$$\mathrm{err}_{\mathrm{FP}}(g, y) = \begin{cases} C_{\mathrm{FA}}, & \text{假接受}, \\ C_{\mathrm{FR}}, & \text{假拒绝}, \\ 0, & \text{正确决策}. \end{cases}$$

因此，指纹验证系统的误差度量必须根据业务需求选择 $C_{\mathrm{FA}}$ 与 $C_{\mathrm{FR}}$，而 0/1 误差通常不足以反映真实风险。

**命题 8.3.2 (指纹验证：超市 vs. CIA)**

下表给出两场景下假接受（FA）与假拒绝（FR）的代价矩阵，其中数值代表单位损失。

| $g$ | 超市折扣系统 | | CIA 门禁系统 | |
|---|---|---|---|---|
| | $y=+1$（顾客） | $y=-1$（入侵者） | $y=+1$（员工） | $y=-1$（入侵者） |
| $+1$ | 0 | 1（小额折扣） | 0 | 1 000（严重泄密） |
| $-1$ | 10（顾客流失） | 0 | 1（员工不悦） | 0 |

结论　误差度量必须场景化：

- 超市：$C_{\mathrm{FR}} \gg C_{\mathrm{FA}}$，应优先降低假拒绝；
- CIA：$C_{\mathrm{FA}} \gg C_{\mathrm{FR}}$，应优先降低假接受。

## 8.4 加权分类

**命题 8.4.1 (加权分类：CIA 代价矩阵)**

给定二元分类问题，设类别标签 $y \in \{-1, +1\}$，模型输出 $h(x) \in \{-1, +1\}$。对 CIA 门禁场景，定义逐点代价（权重）如下：

$$\mathrm{Cost}\big(h(x), y\big) = \begin{cases} 1, & \text{若} y = +1,\ h(x) \neq +1 \quad (\text{假拒绝}), \\ 1000, & \text{若} y = -1,\ h(x) \neq -1 \quad (\text{假接受}), \\ 0, & \text{其余情况}. \end{cases}$$

则加权误差度量为

$$E_{\text{out}}(h) = \mathbb{E}_{(x,y)\sim P}\Big[\text{Cost}\big(h(x), y\big)\Big],$$

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N}\text{Cost}\big(h(x_n), y_n\big).$$

此即为加权分类：对不同 $(x, y)$ 赋予不同权重，以反映业务风险。

**命题 8.4.2 (加权分类中最小化 $E_{\text{in}}$)**

给定加权误差

$$E_{\text{in}}^{\text{w}}(h) = \sum_{n=1}^{N} w_n\,\mathbb{I}\big[y_n \neq h(x_n)\big], \quad \text{其中} \quad w_n = \begin{cases} 1, & y_n = +1, \\ 1000, & y_n = -1. \end{cases}$$

朴素想法

- 若数据线性可分，PLA 不受权重影响，仍能收敛。
- Pocket 算法可把“替换规则”改为：
  若 $E_{\text{in}}^{\text{w}}(w_{t+1}) < E_{\text{in}}^{\text{w}}(w)$，则用 $w_{t+1}$ 替换 $w$。
- 经权重修改的 Pocket 算法在加权 $E_{\text{in}}^{\text{w}}$ 上依旧保持类似收敛保证。

等价构造把每个负样本复制 1000 次，得到新数据集

$$\mathcal{D}' : \underbrace{(x_1, +1)}_{1\,\text{次}},\ \underbrace{(x_2, -1), \ldots, (x_2, -1)}_{1000\,\text{次}},\ \ldots,\ \underbrace{(x_N, +1)}_{1\,\text{次}}.$$

则

$$E_{\text{in}}^{\text{w}}(h)\,\text{(原问题)} = E_{\text{in}}(h)\,\text{(复制后问题)},$$

两者最小化等价。

---

**算法 8.4.1:** Weighted Pocket Algorithm（加权口袋算法）

---

**输入:** 数据集 $\mathcal{D} = \{(x_n, y_n, w_n)\}_{n=1}^{N}$，其中 $w_n = 1$ 若 $y_n = +1$，$w_n = 1000$ 若 $y_n = -1$

**输出:** 权重向量 $w_{\text{best}}$

初始化 $w_0 \leftarrow \mathbf{0}$，$w_{\text{best}} \leftarrow w_0$;

**repeat**

    // 加权 PLA

    以概率正比于 $w_n$ 随机选取样本 $(x_n, y_n)$;

    若 $y_n\,(w_t^\top x_n) \le 0$，则 $w_{t+1} \leftarrow w_t + y_n x_n$;

    否则 $w_{t+1} \leftarrow w_t$;

    // 加权口袋替换规则

    **if** $E_{in}^{\text{w}}(w_{t+1}) < E_{in}^{\text{w}}(w_{best})$ **then**

        $w_{\text{best}} \leftarrow w_{t+1}$;

**until** 达到预设迭代次数;

**return** $w_{best}$;

---

注[系统化路径（Reduction）] 利用“虚拟复制”思想，可将上述加权策略推广到多数其他学习算法！

## 8.5 总结

笔记 [噪声与误差]

- 噪声与概率目标：可用条件概率 $P(y\,|\,x)$ 取代确定性目标函数 $f(x)$。
- 误差度量：决定“理想”目标函数的选取。
- 算法级误差度量：由用户定义，需兼顾“合理性”与“友好性”。
- 加权分类：通过虚拟“样本复制”即可轻松实现。