第 16 章 三大学习原则

16.1 奥卡姆剃刀

定义 16.1.1 (奥卡姆剃刀（Occam's Razor）)

奥卡姆剃刀，又称“简约原则”，是一条哲学与科学方法论原则，主张

“如无必要，勿增实体。”

具体而言，当存在多个能够同等程度解释观测现象的假设时，应当优先选择假设数目最少（或结构最简单）的那一个。该原则由 14 世纪英格兰逻辑学家、方济会修士威廉 • 奥卡姆（William of Ockham）提出，故以其名命名。

在机器学习语境中，奥卡姆剃刀常被转述为：“在训练误差相近的模型中，选择复杂度较低者，以降低过拟合风险并提高泛化性能。”

注意　奥卡姆剃刀并非断言简单假设必然正确，而是一种启发式策略：在缺乏进一步证据前，优先采纳简洁解释，以减少不必要的复杂性与潜在错误。

♣

命题 16.1.1 (简单即更好：奥卡姆剃刀在机器学习中的阐释)

1. 简单的定义
   - 简单假设 h：由少量参数（比特）唯一确定，假设描述长度 $\ell(h)$ 较小。
   - 简单模型 $\mathcal{H}$：所含假设总数 $|\mathcal{H}|$ 较小，即增长函数 $m_{\mathcal{H}}(N)$ 更小。
   - 二者联系：
     - 若 h 由 ℓ 比特描述，则 $h \in \mathcal{H}$ 且 $|\mathcal{H}| \leq 2^{\ell}$；
     - 因此有 $\ell(h)\downarrow \Longrightarrow\ |\mathcal{H}|\downarrow \Longrightarrow\ m_{\mathcal{H}}(N)\downarrow$。
2. 哲学与实用理由
   - 数学保证：对于相同的训练误差，$|\mathcal{H}|$ 越小，$\Omega\big(\mathcal{H};N\big)$ 越小，泛化界越紧。
   - 哲学直觉：
     - 简单模型“完美”拟合数据的概率更低；
     - 一旦拟合成功，其统计显著性更高。
   - 实践建议：永远先尝试线性模型，再逐步增加复杂度，并自问“数据是否被过度建模？”

♠

例题 16.1 选择题：决策树桩的可分性概率

考虑 $\mathbb{R}^1$ 中的决策树桩作为假设集 $\mathcal{H}$，其生长函数 $m_{\mathcal{H}}(N)=2N$。对于 10 个独立同分布（i.i.d.）的样本，标签由公平硬币生成，数据 $\mathcal{D}$ 被 $\mathcal{H}$ 可分的概率为：

1) $\frac{1}{1024}$
2) $\frac{10}{1024}$
3) $\frac{20}{1024}$
4) $\frac{100}{1024}$

解答　正确选项为 $\boxed{3}$。决策树桩的生长函数 $m_{\mathcal{H}}(10)=20$，表示最多可实现 20 种标签分配。总标签

分配数为 $2^{10} = 1024$，因此可分概率为：

$$P(\text{可分}) = \frac{20}{1024}$$

■

16.2 采样偏差

定义 16.2.1 (采样偏差（Sampling Bias）)

采样偏差，又称选择偏差（Selection Bias），是指在抽样过程中，由于抽样方案或执行方式不当，导致样本不能代表总体，从而对总体参数产生系统性偏离的误差。

特征与成因

- 系统性：与随机抽样误差不同，采样偏差不会随样本量增加而减小，必须通过改进抽样设计或校正方法消除。
- 常见类型：
  - 抽样框偏差：抽样框未能覆盖全部总体单位；
  - 非随机抽样：便利抽样、自愿者抽样等导致样本自选择；
  - 分层失衡：各层样本比例与总体比例不符。

量化与检测

$$\text{采样偏差} = \mathbb{E}[\hat{\theta}_{\text{sample}}] - \theta_{\text{population}},$$

其中 $\hat{\theta}_{\text{sample}}$ 为样本统计量，$\theta_{\text{population}}$ 为总体真值。

减少策略

- 使用概率抽样（简单随机、分层、整群、系统抽样）；
- 扩大或校正抽样框；
- 事后加权或校准估计。

♣

命题 16.2.1 (应对采样偏差)

若训练数据存在采样偏差，则学到的模型也会继承同样的偏差。

实用经验法则

- 让训练/验证的数据分布尽可能贴近测试场景。
- 举例：若测试阶段的输入为“用户最近记录”（如 KDDCup 2011），则
  - 训练：应加重近期样本的权重；
  - 验证：使用“靠后”的用户记录。

♠

16.3 数据偷窥

定义 16.3.1 (数据偷窥（Data Snooping / Peeking）)

数据偷窥是指在完成正式分析或建模之前，过早地查看数据或初步结果，并据此调整分析流程、特征选择、模型结构或超参数的行为。该行为会导致统计结论出现系统性偏差，通常表现为：

- I 类错误膨胀：假阳性率被人为提高，使原本不显著的结论看似显著。
- 泛化性能虚高：训练-验证误差被低估，模型在实际部署时表现远低于预期。
- 可重复性受损：由于分析路径受偷窥结果影响，他人难以复现相同结论。

常见表现

- 根据初步显著性水平事后删减变量或重选阈值；
- 反复调参直到验证指标“足够好”；
- 在训练集上观察分布后人为构造特征。

避免策略

- 预先注册分析计划，锁定特征、模型及评价指标；
- 引入独立测试集，仅在最终一次性评估；
- 采用交叉验证或嵌套交叉验证，确保每一次调参都在训练折上完成；
- 对多次比较进行多重检验校正（如 Bonferroni、FDR）。

命题 16.3.1 (应对数据偷窥)

完全避免数据偷窥几乎不可能，除非极端坦诚。

极端坦诚做法　把测试数据锁进保险箱，永不提前查看。

相对可行的折中做法

- 保留验证集，但谨慎使用：任何基于验证集的调参都需记录并自我审查。
- 保持“盲视”：在建模阶段避免用任何数据结果做决策。
- 保持“怀疑”：对所有研究结果（包括自己的）都假设已受污染，并用“污染感”去解读。

核心思想　在“数据驱动建模（必要偷窥）”与“验证无偷窥”之间取得精细平衡。

16.4 三的威力

命题 16.4.1 (三大相关领域)

1. 数据挖掘（**Data Mining**）　利用海量数据发现其中有趣或有价值的模式与规律。
2. 人工智能（**AI**）　构建能够表现出智能行为的系统或算法，以解决复杂任务。
3. 统计学（**Statistics**）　基于数据对未知总体或过程进行推断、估计与假设检验。

相互关系

- 机器学习（ML）是连接三者的桥梁之一；
- 统计学为机器学习提供了大量有用工具；
- 数据挖掘与机器学习在实践中难以严格区分；
- 三者共同构成实现人工智能的一条可行路径。

命题 16.4.2 (三大理论边界)

1. **Hoeffding** 不等式（单假设）

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

适用于对 单个固定假设的误差验证或测试。

2. 多箱 **Hoeffding** 不等式（有限假设集）

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

其中 $M = |\mathcal{H}|$。适用于 验证（有限候选模型）场景。

3. **VC** 界（无限假设集）

$$\mathbb{P}[\exists h \in \mathcal{H}, |E_{\text{in}}(h) - E_{\text{out}}(h)| > \varepsilon] \leq 4\, m_{\mathcal{H}}(2N)\exp\left(-\tfrac{1}{8}\varepsilon^2 N\right),$$

以增长函数 $m_{\mathcal{H}}(N)$ 代替 M，适用于 训练阶段对任意规模假设集 $\mathcal{H}$ 的泛化保证。

命题 16.4.3 (三大线性模型)

1. 感知机算法 / **Pocket** 算法

假设函数：$h(x) = \text{sign}(s)$，其中 $s = w^\top x$

损失函数：0/1 损失（不可微）

优化特点：对线性可分数据保证收敛；对含噪数据可通过 Pocket 算法进行启发式改进

2. 线性回归

假设函数：$h(x) = s = w^\top x$

损失函数：平方损失（二次可微）

优化特点：存在解析解 $w = (X^\top X)^{-1}X^\top y$，计算简便，适用于小翻转噪声场景

3. 逻辑回归

假设函数：$h(x) = \sigma(s) = \dfrac{1}{1+e^{-s}}$，输出概率值

损失函数：交叉熵损失（对应最大似然估计）

优化特点：无闭式解，需采用迭代算法（如梯度下降、Newton 法等）

命题 16.4.4 (三大关键工具)

1. 特征变换（**Feature Transform**）

通过复杂映射 ϕ 将原始特征提升到高维空间，从而降低 E_{in}，但会增大 $d_{\text{VC}}(\mathcal{H})$。

2. 正则化（**Regularization**）

在损失函数中加入正则项 $\Omega(w)$，把有效 VC 维 $d_{\text{EFF}}(\mathcal{H}, \lambda)$ 控制到较小值，防止过拟合，代价是 E_{in} 可能略升。

3. 验证（**Validation**）

预留 K 个样本作验证集 $\mathcal{D}_{\text{val}}$，用 E_{val} 从有限候选 $\{g_1, \ldots, g_M\}$ 中选出最佳模型，减少选择自由度，换取可靠的泛化估计。

命题 16.4.5 (三大学习原则)

1. 奥卡姆剃刀（**Occam's Razor**）
   当不同假设的训练误差相近时，应优先选择结构更简单的模型。这一原则的核心是通过控制模型复杂度降低过拟合风险，在保证拟合能力的同时提升泛化性能。
2. 采样偏差（**Sampling Bias**）
   训练数据的分布必须与测试场景（真实世界）的分布保持一致。若违背此原则，模型学到的将只是训练数据的特定模式，而非问题的本质规律，最终导致泛化能力低下。
3. 数据偷窥（**Data Snooping**）
   建模过程中任何对测试数据或未公开结果的提前窥探，都会引入隐性偏差并污染最终结论。严谨的做法是将测试数据严格隔离，仅在模型确定后进行一次性评估，避免反复利用测试信息调整模型。

命题 16.4.6 (三大未来方向)

1. 更多变换（**More Transform**）
   通过核技巧、深度网络、特征嵌入等手段，把数据映射到更丰富的高维空间，以捕捉复杂结构。
2. 更多正则（**More Regularization**）
   采用稀疏约束、最大间隔、集成平均、提前停止等策略，抑制过拟合并提升泛化性能。
3. 更少标签（**Less Label**）
   借助无监督学习、半监督学习、自编码器、主动学习等方法，在标签稀缺的场景下仍能学得有效模型。

代表性技术与工具

- 核方法：SVM、核逻辑回归、RBF 网络、高斯核、二次规划。
- 集成与提升：Bagging、AdaBoost、随机森林、GBDT、OOB 误差。
- 稀疏与降维：PCA、矩阵分解、坐标下降、L1 正则。
- 深度学习：自编码器、深度神经网络、原型学习、k-means。
- 大间隔理论：软间隔、支持向量回归（SVR）、概率 SVM。

16.5 总结

笔记 [三大学习原则]

- 奥卡姆剃刀：如无必要，勿增实体——优先选择最简单的假设。
- 采样偏差：训练数据的采集方式必须与测试场景尽量匹配。
- 数据偷窥：任何对数据的“提前使用”都是污染。
- 三的威力：诸多概念（如上界、模型、工具、原则等）常以‘三’组合出现。