机器学习笔记：计算学习理论 (Computational Learning Theory)

一、 核心框架：PAC 学习模型

PAC (Probably Approximately Correct，大概近似正确) 是计算学习理论的基础。

1. **定义理解**：
    - **近似正确 (AC)**：学习到的假设 h 的真误差 $\mathrm{error}_{\mathrm{true}}(h)$ 小于一个很小的常数 ϵ。
    - **大概 (P)**：学习器以至少 $1-\delta$ 的概率达到“近似正确”。
2. **PAC 可学习性**：如果一个概念类 C，学习器 L 可以在**多项式时间**内（关于 $1/\epsilon, 1/\delta, n, \mathrm{size}(c)$ 的多项式）输出满足条件的假设，则称该类是 PAC 可学习的。
3. **核心意义**：它不仅要求模型“学得准”，还要求“学得快（时间复杂度）”且“样本够（样本复杂度）”。

二、 两种误差与过拟合

1. **训练误差 ($\mathrm{error}_{\mathrm{train}}$)**：在已有训练集 D 上的错误率。
2. **真误差 ($\mathrm{error}_{\mathrm{true}}$)**：在全概率分布 $P(X)$ 下，假设 h 预测错误的概率。
3. **过拟合 (Overfitting)**：当 $\mathrm{error}_{\mathrm{true}}(h) > \mathrm{error}_{\mathrm{train}}(h)$ 时发生。
    - **过拟合量** = $\mathrm{error}_{\mathrm{true}}(h) - \mathrm{error}_{\mathrm{train}}(h)$。
    - **理论困境**：我们只能观察到训练误差，计算学习理论的目标就是**通过训练误差来界定真误差的范围。**

三、 有限假设空间 ($|H| < \infty$) 的样本复杂度

当假设空间 H 是有限的时候，我们关注**版本空间 (Version Space)**，即所有与训练集完全一致（$\mathrm{error}_{\mathrm{train}} = 0$）的假设集合。

1. **Haussler 定理**：
    - 如果训练样本数 m 足够多，版本空间中任何真误差 $> \epsilon$ 的假设被保留下来的概率会非常小。
    - **概率上限**：$|H|e^{-\epsilon m}$。
2. **样本复杂度公式 (必背)**：
   为了保证以 $1-\delta$ 的概率达到 PAC 学习，所需的样本数 m 满足：
   $m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$
    - **直观理解**：假设空间 $|H|$ 越大（模型越复杂），需要的样本量就越多，且呈对数关

系增长。

四、 不可知学习 (Agnostic Learning)

当目标概念 c 不在假设空间 H 中时，我们无法找到训练误差为 0 的假设。此时目标是寻找训练误差最小的 h。

1. **Hoeffding 不等式应用**：
   在这种场景下，真误差与训练误差的关系为：
   $\text{error}_{\text{true}}(h) \leq \text{error}_{\text{train}}(h) + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{2m}}$
2. **结论**：随着样本量 m 增加，训练误差和真误差的差距会以 $\sqrt{1/m}$ 的速度缩小。

五、 无限假设空间与 VC 维 (VC Dimension)

当假设空间 H 是无限的（如线性分类器、神经网络），$|H|$ 无法直接使用，需要引入 **VC 维** 来衡量模型复杂度。

1. **基本概念**：
   - **对分 (Dichotomy)**：对集合 S 的一种正负标签划分。
   - **打散 (Shattering)**：如果 H 能实现集合 S 上所有可能的划分（即 $2^{|S|}$ 种），则称 H 打散了 S。
   - **VC 维定义**：H 能够打散的**最大**有限子集的大小。
2. **线性分类器示例**：二维平面上的直线，最多能打散 3 个点（无法打散某些 4 点分布，如异或问题），故其 VC 维为 3。通常，n 维空间的线性超平面 VC 维为 $n+1$。
3. **基于 VC 维的样本复杂度**：
   $m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$
   - **关键点**：这证明了即便假设空间是无限的，只要 VC 维有限，模型就是 PAC 可学习的。

六、 典型案例分析

1. **布尔合取式 (Conjunctions)**：
   - n 个变量，$|H| = 3^n$（每个变量有：出现在式中、以否定形式出现、不出现三种状态）。
   - 样本复杂度随变量个数 n 线性增长。
2. **决策树 (Decision Trees)**：
   - n 个布尔特征，假设空间 $|H| = 2^{2^n}$，极其庞大。
   - **结论**：在没有先验假设的情况下，学习完全正确的决策树需要几乎所有的实例（2^n

），这就是“没有免费午餐”定理的体现。

复习建议 (记忆口诀)

- **PAC 两大要素**：ϵ 是精度，δ 是稳度。
- **有限空间看 $|H|$**：样本量 m 与 $\ln |H|$ 成正比。
- **无限空间看 VC 维**：只要能打散，复杂度就有上限。
- **一致性学习器**：只要训练误差为 0，样本够多，真误差一定小。
- **不可知学习**：目标不在 H 内，就找训练误差最小的那一个。