# 西瓜书笔记

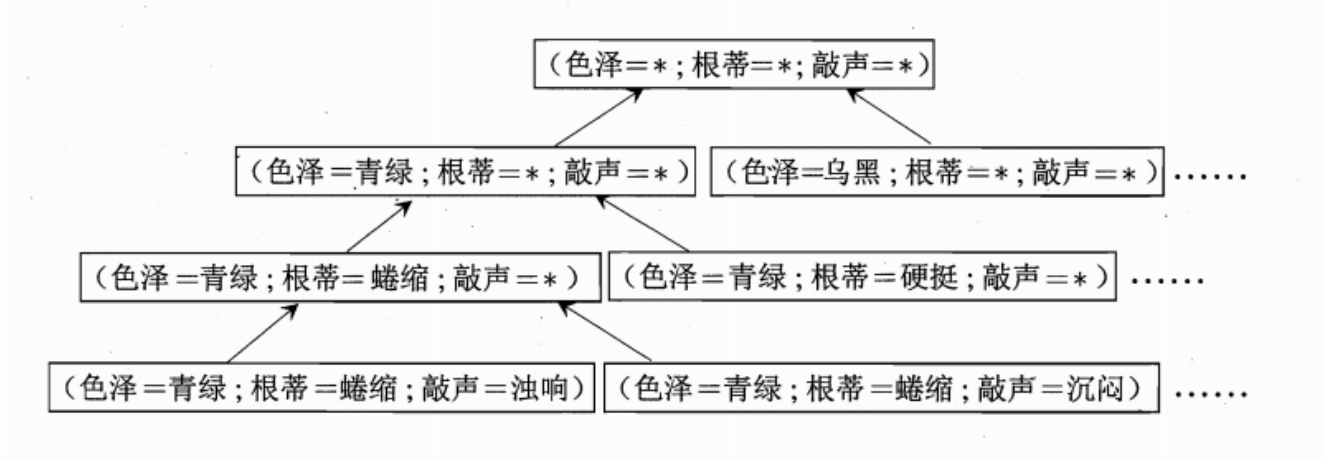


### 1.1 引言

- 机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能。
- 在计算机中，“经验”通常以数据形式存在，因此，机器学习主要研究在计算机上从数据中产生模型的算法，即“**学习算法**”
- 有了学习算法，我们把经验数据提供给它，它就能基于这些数据产生模型
- 在面对新的情况时，模型会给我们提供相应的判断

### 1.3 假设空间

- 归纳和演绎是科学推理的两大基本手段
- 我们可以把学习过程看作一个在所有假设组成的空间中进行搜索的过程，搜索目标是找到与训练集匹配的假设。
- 假设的表示一旦确定，假设空间及其规模大小就确定了
- 拿西瓜举例：若色泽、根蒂、敲声分别有3、2、2种可能取值，则我们面临的假设空间规模大小为4×3×3+1=37中，其中1代表空集




- 可以有许多策略对这个假设空间进行搜索，例如自顶向下，从一般到特殊，或是自底向上、从特殊到一般，搜索过程中可以不断删除与正例不一致的假设，和与反例一致的假设，最终将会获得与训练集一致的假设

### 1.4 归纳偏好

- 任何一个有效的机器学习算法本身必带有归纳偏好，否则它将被假设空间看似在训练集上“等效”的假设所迷惑，而无法产生确定的学习效果。
- 那么，有没有一般性的原则来引导算法确立的“正确的”偏好呢？
  “奥卡姆剃刀”：若有多个假设与观察一致，则选最简单的那个
- 在具体的现实问题中，算法的归纳偏好是否与问题本身匹配，大多数时候直接决定了算法能否取得好的性能
- “没有免费的午餐”定理:式（1.2）显示出，总误差竟然与学习算法无关！对于任意两个算法，我们都有：
  - $$\sum_f E_{ote}(\mathfrak{L}_a|X,f) = \sum_f E_{ote}(\mathfrak{L}_b|X,f)\ ,$$
- 所以，学习算法自身的归纳偏好与问题是否相配，往往会起到决定性的作用

### 2.1 概念

一些重要概念

- **错误率**:分类错误的样本数占样本总数的比例
- **精度**=1-错误率
- **误差**：实际预测输出与样本的真实输出之间的差异，其中，学习器在训练集上的误差称为“训练误差”或“经验误差” ,新样本上的误差称为“泛化误差”
- **过拟合**：把训练样本自身的一些特点当作了所有潜在样本具有的性质 从而导致泛化性能下降
- **欠拟合**：对训练样本的一般性质尚未学好

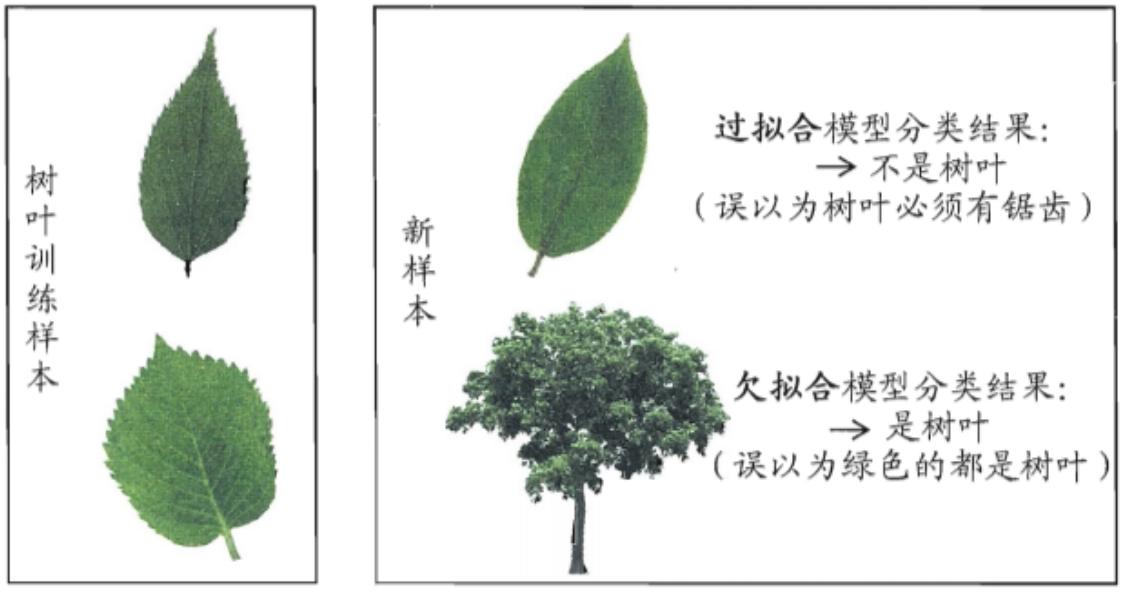



图 2.1 过拟合、欠拟合的直观类比

-

### 2.2 评估方法

- 通常，我们可以通过实验测试来对学习器的泛化误差进行评估，使用一个测试集，测试集上的“测试误差”作为**泛化误差**的近似

- 注意，测试机应该尽可能与训练集**互斥**

**常见处理数据集做法**

1. **留出法**

- 直接将数据集D划分为**两个互斥的集合**，分别为训练集和测试集，注意要尽可能保持数据分布的**一致性**
- 单次使用留出法得到的估计结果往往不够稳定可靠，在使用留出法时，一般要采用**若干次随机划分、重复进行实验后取平均值**

2. **交叉验证法**

- 先将数据集划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的那个子集作为测试集，从而可以进行k次训练和测试，最后求均值

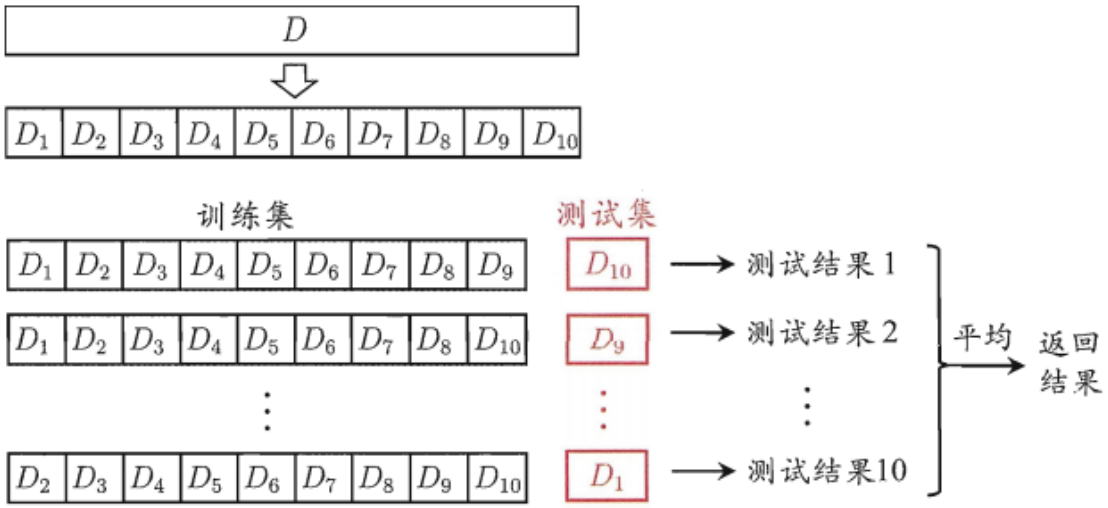



**图 2.2** 10 折交叉验证示意图

- 交叉验证法的一个特例：留一法
  m个样本划分为m个子集
- 留一法评估结果比较准确，但是计算量大

**3. 自助法（bootstapping）**

给定包含m个样本的数据集D，每次随机从D中挑选一个样本，将其拷贝放入D',再将此样本放回D中

- 显然，D中可能会有一部分样本在D'多次出现，另一部分样本不出现，样本在m次采样中始终不被采到的概率是：

$$\lim_{m\mapsto\infty}\left(1-\frac{1}{m}\right)^m \mapsto \frac{1}{e}\approx 0.368\ ,$$

  所以我们仍有约1/3的没在训练集中出现的样本用于测试，这样的结果，称为包外估计

- 自助法在数据集较小，难以有效划分训练/测试集**很有用**，然而，自助法产生的数据集改变了初始数据集的分布，这会引入**估计偏差。**

### 2.3 性能度量

对学习器的泛化性能进行评估，不仅需要上诉说的有效可行的实验估计方法，还需要衡量模型泛化能力的评价标准

- 回归任务最常用的性能度量是“均方误差”
- $$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2 \quad .$$

  **下面是分类任务中常用的性能度量**
  对于上文所说的错误率，对样例集D，分类错误率定义为：
  $$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right) \quad .$$