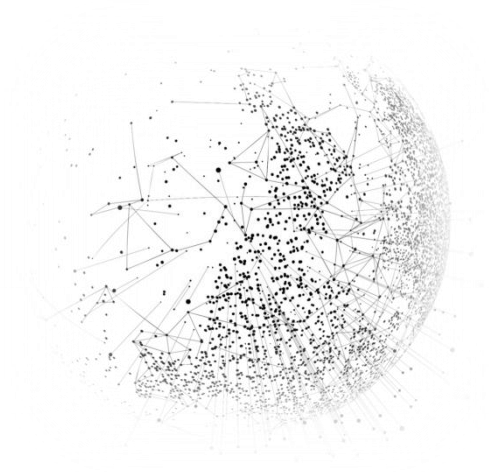
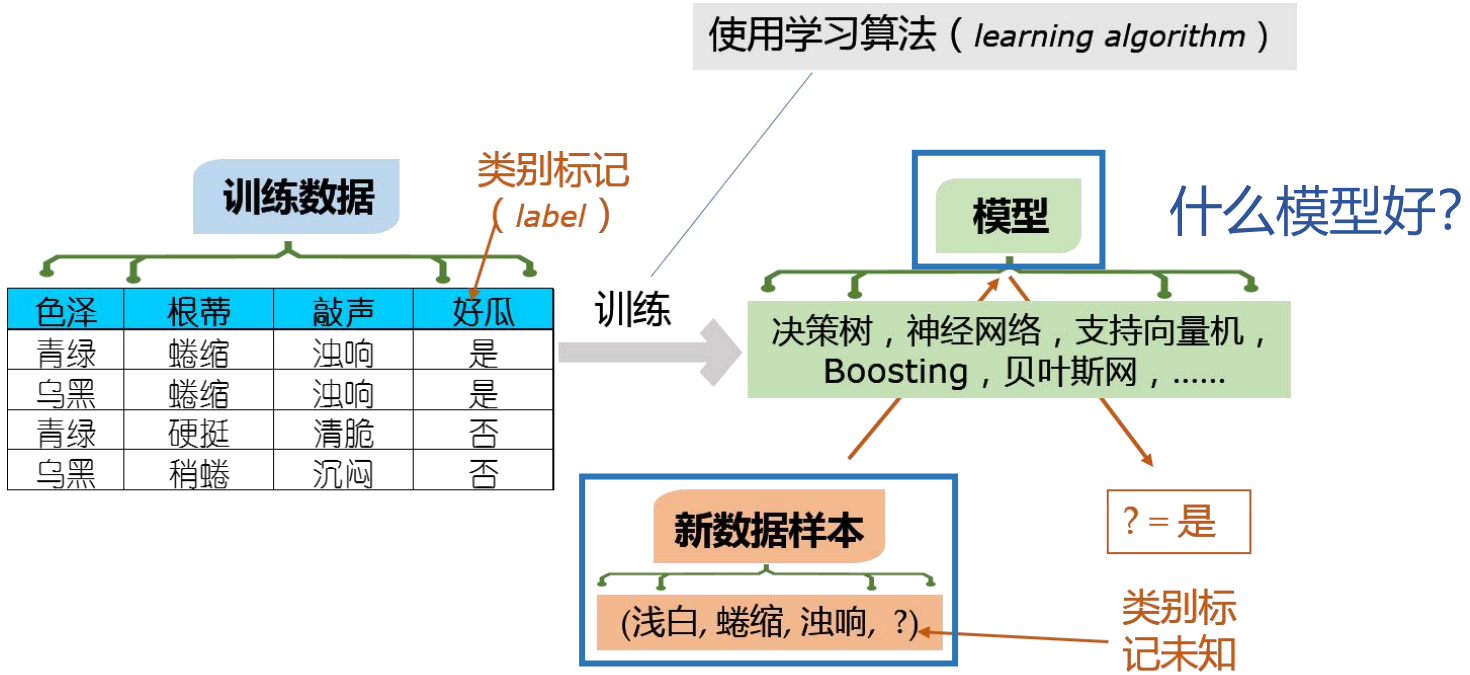


二、模型评估与选择

典型的机器学习过程



泛化能力强！

能很好地适用于 unseen instance

例如，**错误率**低、**精度**高

然而，我们手上没有 unseen instance，……

泛化误差 vs. 经验误差

泛化误差：在“未来”样本上的误差

经验误差：在训练集上的误差，亦称“训练误差”

- 泛化误差越小越好
- 经验误差是否越小越好?

NO!

因为会出现“过拟合”(overfitting)

过拟合 (Overfitting)

vs.

欠拟合 (Underfitting)

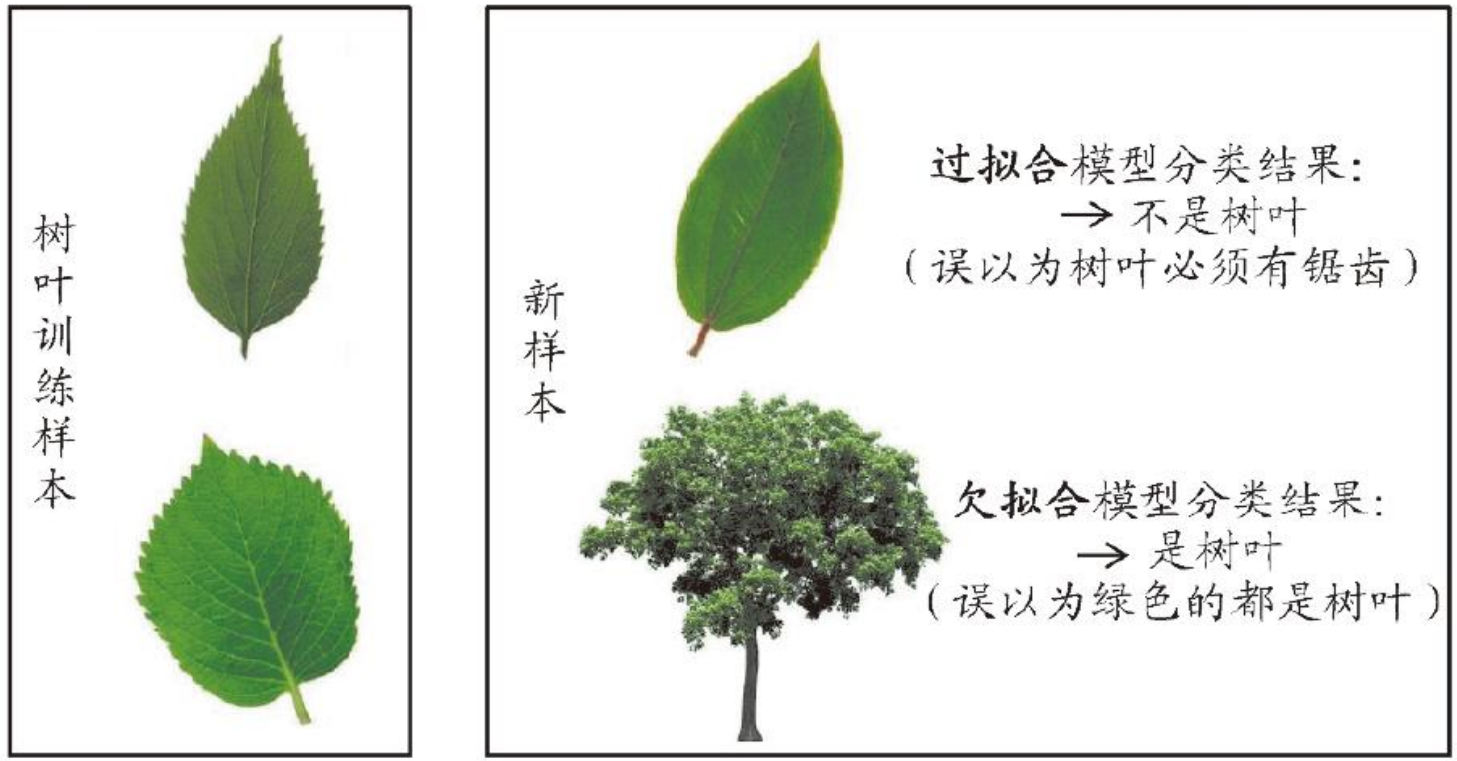


图 2.1 过拟合、欠拟合的直观类比

模型选择
(Model Selection)

三个关键问题:

- 如何获得测试结果? ⟹ 评估方法
- 如何评估性能优劣? ⟹ 性能度量
- 如何判断实质差别? ⟹ 比较检验

评估方法

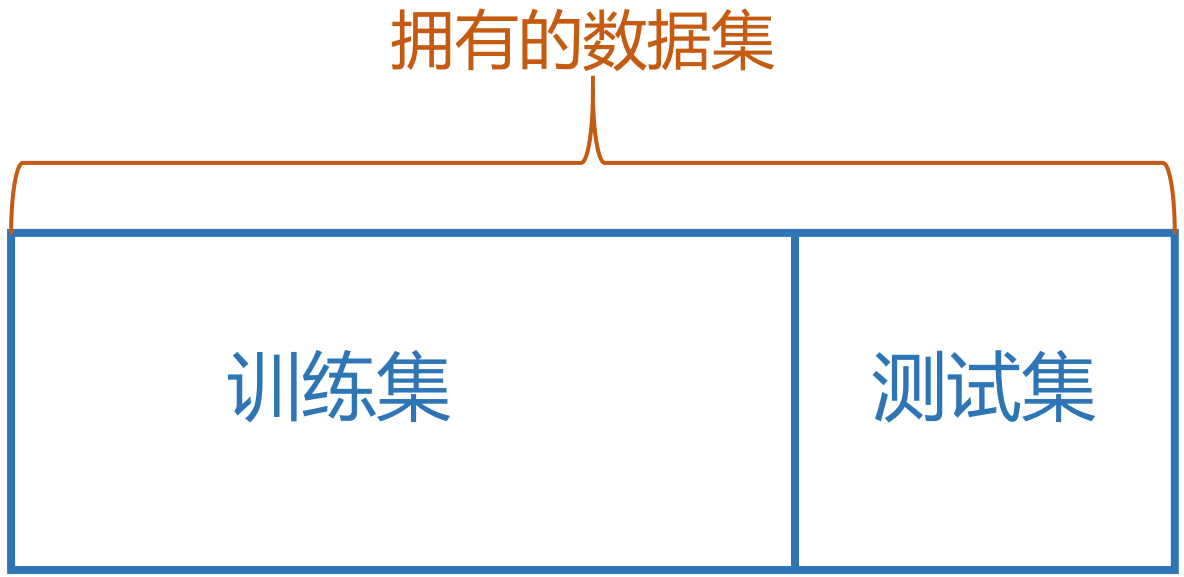
关键：怎么获得“测试集” (test set)？

测试集应该与训练集“互斥”

常见方法：

- 留出法 (hold-out)
- 交叉验证法 (cross validation)
- 自助法 (bootstrap)

留出法



注意：

- 保持数据分布一致性 （例如: 分层采样)
- 多次重复划分 (例如: 100次随机划分)
- 测试集不能太大、不能太小 (例如： 1/5~1/3)

k-折交叉验证法

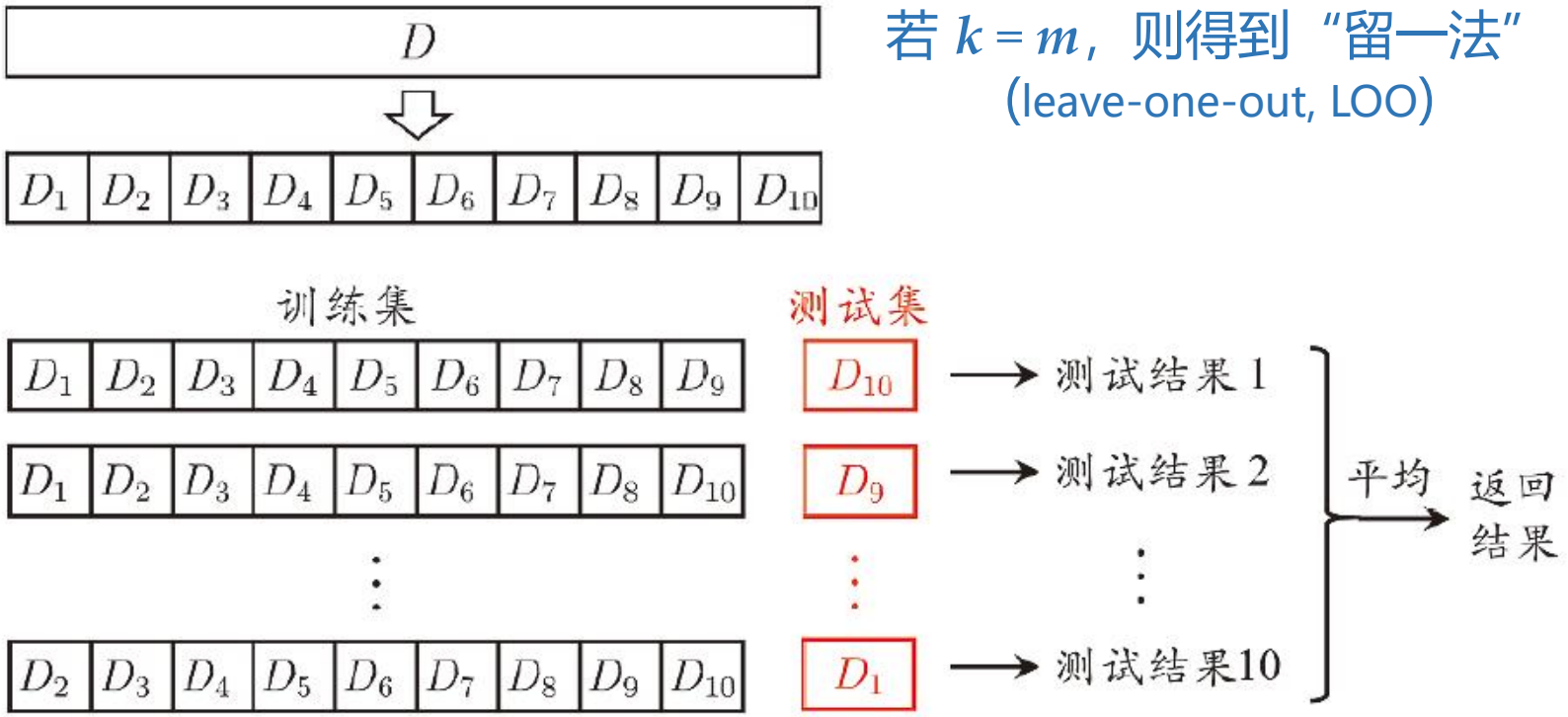


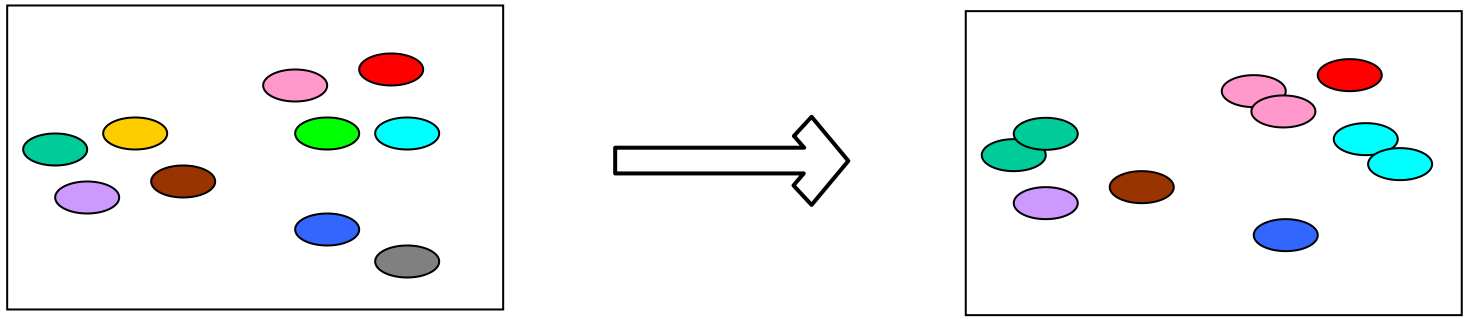
图 2.2　10 折交叉验证示意图

若 $k = m$，则得到“留一法”
(leave-one-out, LOO)

自助法

基于“自助采样” (bootstrap sampling)

亦称“有放回采样”、“可重复采样”



约有 36.8% 的样本不出现

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

“包外估计”(out-of-bag estimation)

- 训练集与原样本集同规模
- 数据分布有所改变

“调参”与最终模型

算法的参数：一般由人工设定，亦称“超参数”

模型的参数：一般由学习确定

调参过程相似：先产生若干模型，然后基于某种评估方法进行选择

参数调得好不好对性能往往对最终性能有关键影响

区别：训练集 vs. 测试集 vs. 验证集 (validation set)

算法参数选定后，要用“训练集+验证集”重新训练最终模型

模型选择
(Model Selection)

三个关键问题:

- 如何获得测试结果? ⇒ 评估方法
- 如何评估性能优劣? ⇒ 性能度量
- 如何判断实质差别? ⇒ 比较检验

性能度量

性能度量(performance measure)是衡量模型泛化能力的评价标准，反映了任务需求

使用不同的性能度量往往会导致不同的评判结果

什么样的模型是“好”的，不仅取决于算法和数据，还取决于任务需求

- 回归(regression) 任务常用均方误差：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2$$

错误率 vs. 精度

- 错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

- 精度：

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

查准率 vs. 查全率

表 2.1　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- 查准率：$$P = \frac{TP}{TP + FP}$$
- 查全率：$$R = \frac{TP}{TP + FN}$$

F1

F1 度量：

$$F1 = \frac{2 \times P \times R}{P + R}$$

$$= \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

$$\frac{1}{F1} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R} \right)$$

若对查准率/查全率有不同偏好：

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$$\frac{1}{F_\beta} = \frac{1}{1 + \beta^2} \cdot \left(\frac{1}{P} + \frac{\beta^2}{R} \right)$$

$\beta > 1$ 时查全率有更大影响；$\beta < 1$ 时查准率有更大影响

模型选择
(Model Selection)

三个关键问题:

- 如何获得测试结果? ⇒ 评估方法
- 如何评估性能优劣? ⇒ 性能度量
- 如何判断实质差别? ⇒ 比较检验

比较检验

在某种度量下取得评估结果后，是否可以直接比较以评判优劣？

NO！

因为：

- 测试性能不等于泛化性能
- 测试性能随着测试集的变化而变化
- 很多机器学习算法本身有一定的随机性

机器学习 …… “概率近似正确”

常用方法

统计假设检验 (hypothesis test) 为学习器性能比较提供了重要依据

- 两学习器比较
  - 交叉验证 t 检验 (基于成对 t 检验)
    - k 折交叉验证； 5x2交叉验证
  - McNemar 检验 (基于列联表，卡方检验)

统计显著性

- 多学习器比较
  - Friedman + Nemenyi
    - Friedman检验 (基于序值，F检验; 判断”是否都相同”)
    - Nemenyi 后续检验 (基于序值，进一步判断两两差别)