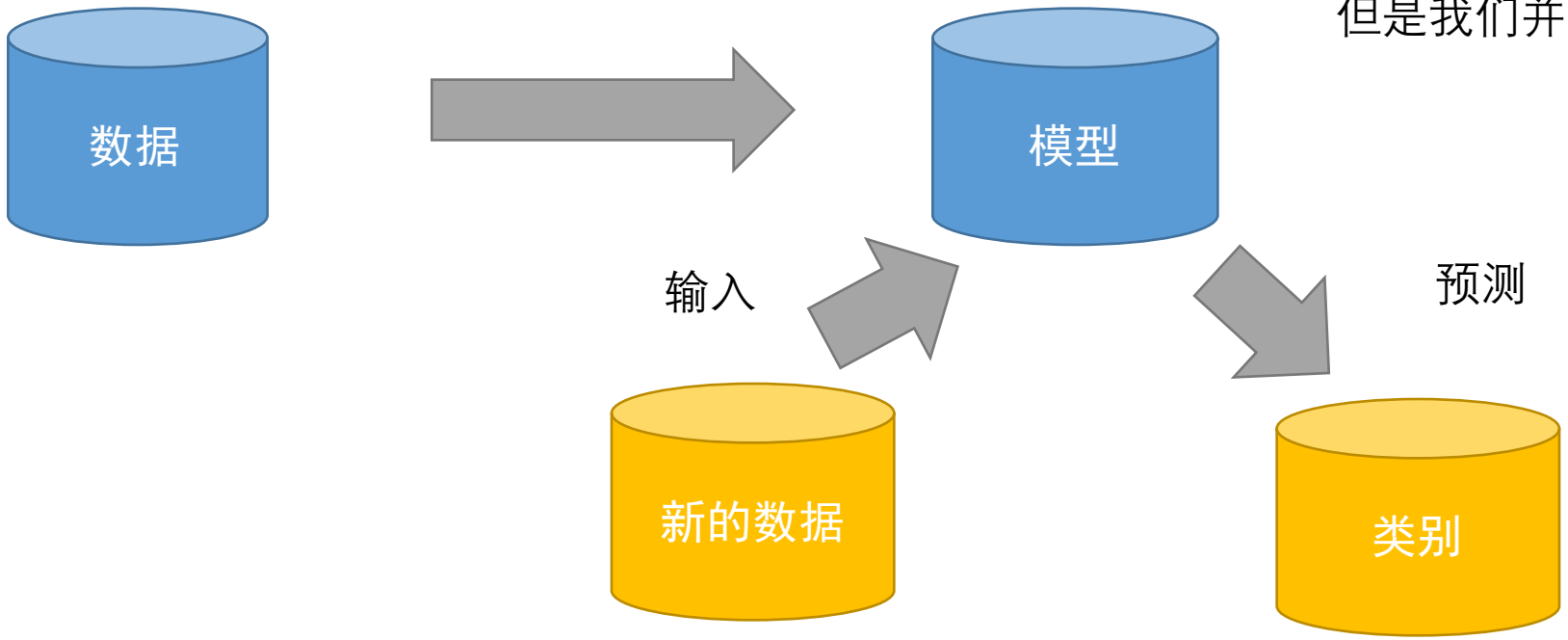
模型评估

Model evaluation

模型评估与选择

一般的机器学习过程



什么是好的额模型？
回答：泛化能力强的模型，也就是可以很好处理未知数据模型
但是我们并不可能拥有所有数据

泛化误差与经验误差

- 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，又称“训练误差”
- 所以，泛化误差越小越好
- 经验误差并不是越小越好，因为会发生过拟合的问题（overfitting）

过拟合与欠拟合


树叶训练样本


新样本
过拟合模型分类结果：
→ 不是树叶
（误以为树叶必须有锯齿）
欠拟合模型分类结果：
→ 是树叶
（误以为绿色的都是树叶）

模型选择–2个关键问题

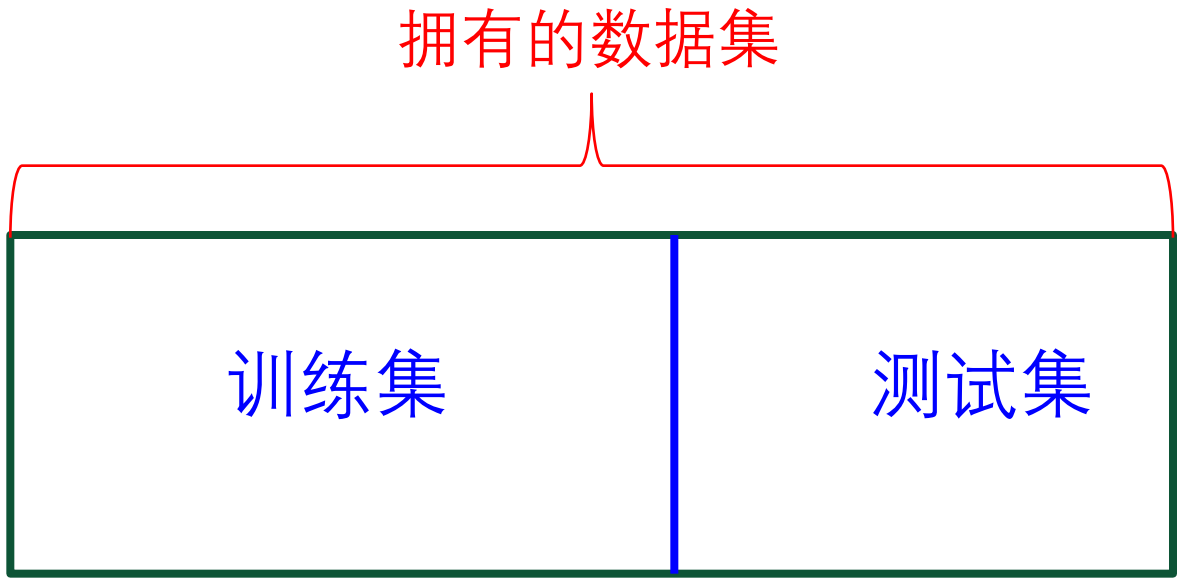
- 如何获得测试集？
- 如何评估测试集？

怎么获得测试集

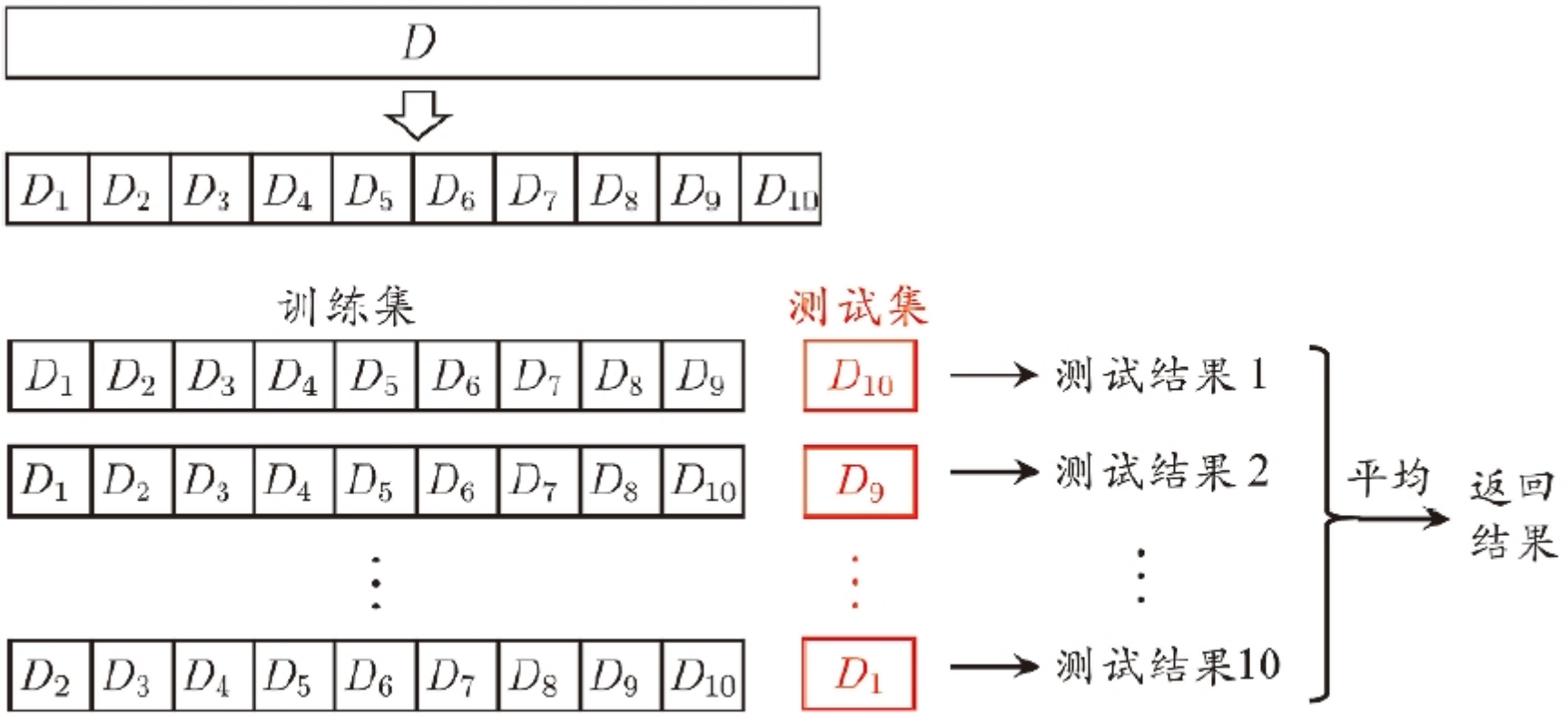
- 留出法
- K-折交叉验证

留出法

- 测试集一般为数据的1/5 ~ 1/3
- 保持数据分布的一致性



K-折交叉验证法


D
D1 D2 D3 D4 D5 D6 D7 D8 D9 D10
训练集
测试集
D1 D2 D3 D4 D5 D6 D7 D8 D9
D10
测试结果 1
D1 D2 D3 D4 D5 D6 D7 D8 D10
D9
测试结果 2
D2 D3 D4 D5 D6 D7 D8 D9 D10
D1
测试结果10
平均
返回结果

评价模型

- 使用不同的评价指标会导致不同的评判结果
- 什么样的模型是“好”的，不仅仅取决于算法和数据，还取决于任务需求
- 回归任务常用均方误差：

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

评价模型

- 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：

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

评价模型

- 准确率：P = TP/（TP + FP）
- 召回率：R = TP/（TP + FN）

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)

评价模型

- 在图片过滤系统中，使用违禁品检测模型每天对滤线上的图片进行过滤。每天大约有1万张图片被模型认为是包含违禁品的图片，违禁品检测模型的准确率是93%，召回率是97%，这意味着什么呢？
- 准确率：这1万张图片中有9300张图片是真的违禁品
- 召回率：9300张图片是每天所有包含违禁品图片的97%

评价模型

- F-score：综合准确率与召回率的指标

$$F1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times TP}{\text{样例总数} + TP - TN}$$

- 更加通用的形式
  - **β** >1时召回率更具影响力，**β** <1时准确率更具影响力

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