



模型评估与选择

Model Evaluation and Selection



俞俊、高飞、谭敏、余宙、匡振中

{yujun, gaofei, tanmin, yuz, zzkuang}@hdu.edu.cn

<http://mil.hdu.edu.cn>

学习目标

- 能够正确说出示例/样本、属性/特征、模型、假设空间等术语的含义；
- 实验过程中，能够正确使用留出法和交叉验证法；
- 实验过程中，能够正确划分训练集、验证集、测试集
- 给定预测结果和真实结果，能够正确计算以下性能度量：
 - 均方误差、错误率、精度、查全率、查准率、F1、ROC、AUC

CONTENTS

第一章 绪论

基本术语、泛化能力、归纳偏好

第二章 模型评估与选择

评估方法、性能指标

CONTENTS

第一章 绪论

基本术语、泛化能力、归纳偏好

第二章 模型评估与选择

评估方法、性能指标

机器学习的定义

- Tom Mitchell (1997)

- Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

[Mitchell, 1997] 给出了一个更形式化的定义：假设用 P 来评估计算机程序在某任务类 T 上的性能，若一个程序通过利用经验 E 在 T 中任务上获得了性能改善，则我们就说关于 T 和 P ，该程序对 E 进行了学习。

- 《机器学习》 周志华 著

- 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；
- 经验通常以“数据”（ data ）形式存在；
- 主要研究内容：在计算机上，从数据中产生“模型”（ model ）的算法，即“学习算法”（ learning algorithm ）

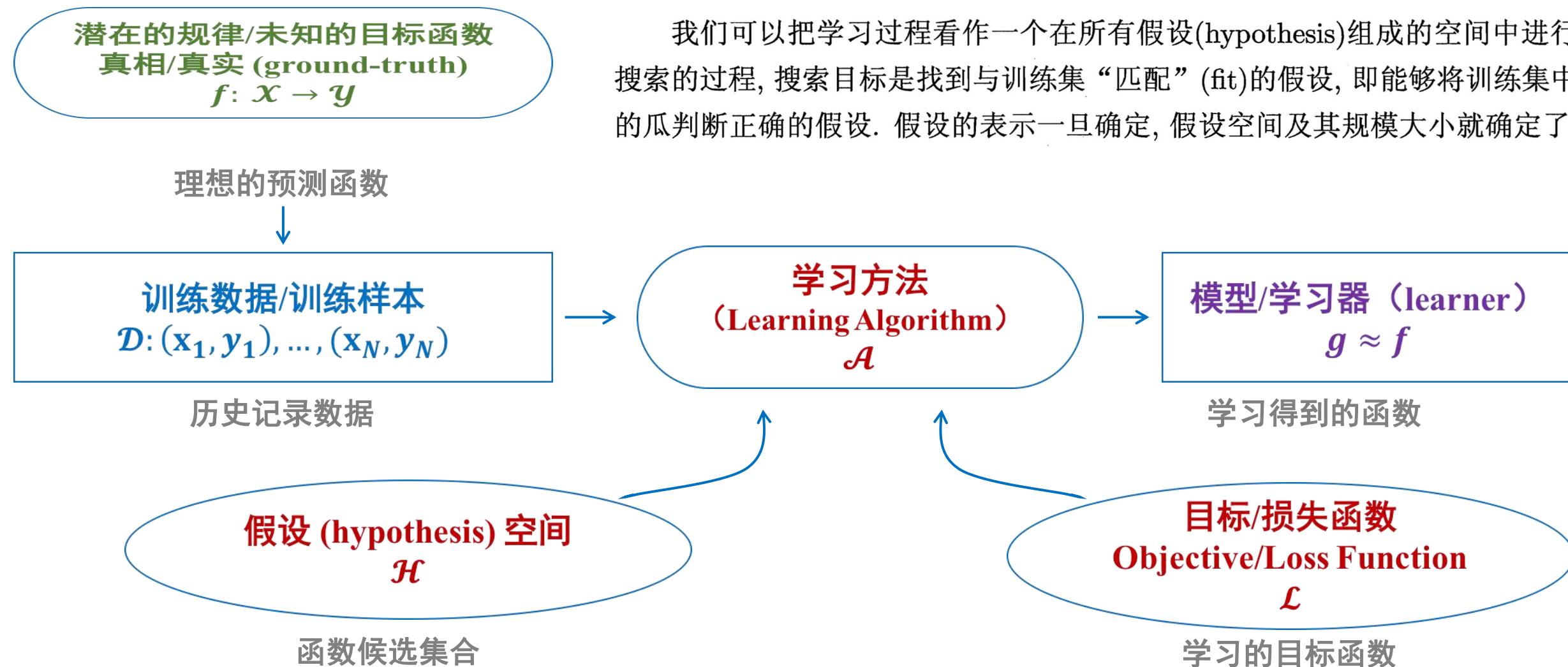
基本术语

- 数据集 (data set)
 - 示例 (instance)、样本 (sample)
 - 属性 (attribute)、特征 (feature)
 - 属性值 (attribute value)
 - 属性空间 (attribute space)、样本空间 (sample space)、输入空间
 - 特征向量 (feature vector)
-

一般地, 令 $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ 表示包含 m 个示例的数据集, 每个示例由 d 个属性描述(例如上面的西瓜数据使用了 3 个属性), 则每个示例 $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id})$ 是 d 维样本空间 \mathcal{X} 中的一个向量, $\mathbf{x}_i \in \mathcal{X}$, 其中 x_{ij} 是 \mathbf{x}_i 在第 j 个属性上的取值

d 称为样本 \mathbf{x}_i 的“维数” (dimensionality).

基本术语



假设空间

• 例：根据“色泽”、“根蒂”、“敲声”判断“好瓜”

- 特征无论取什么值都合适，用通配符“*”表示；没有“好瓜”，用 \emptyset 表示
- 各有3、2、2个属性值，则假设空间规模为 $4 \times 3 \times 3 + 1 = 37$

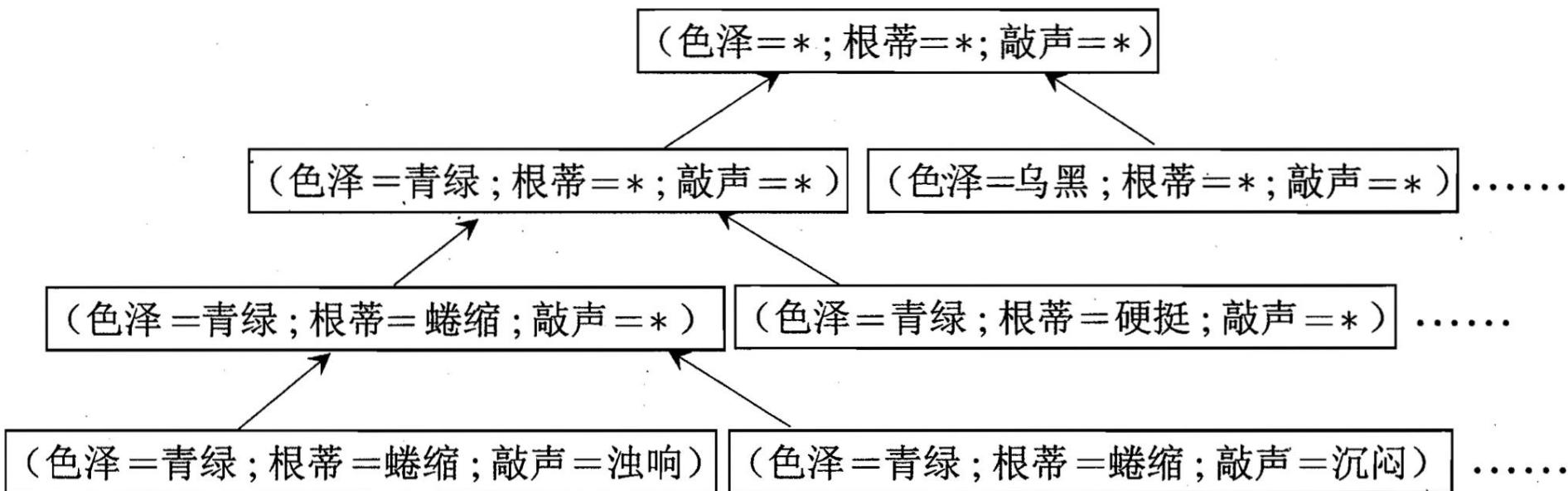


图 1.1 西瓜问题的假设空间

假设空间

- 问题：根据“色泽”、“根蒂”、“敲声”判断“好瓜”

- 特征各有3、3、3个属性值，则假设空间规模为？

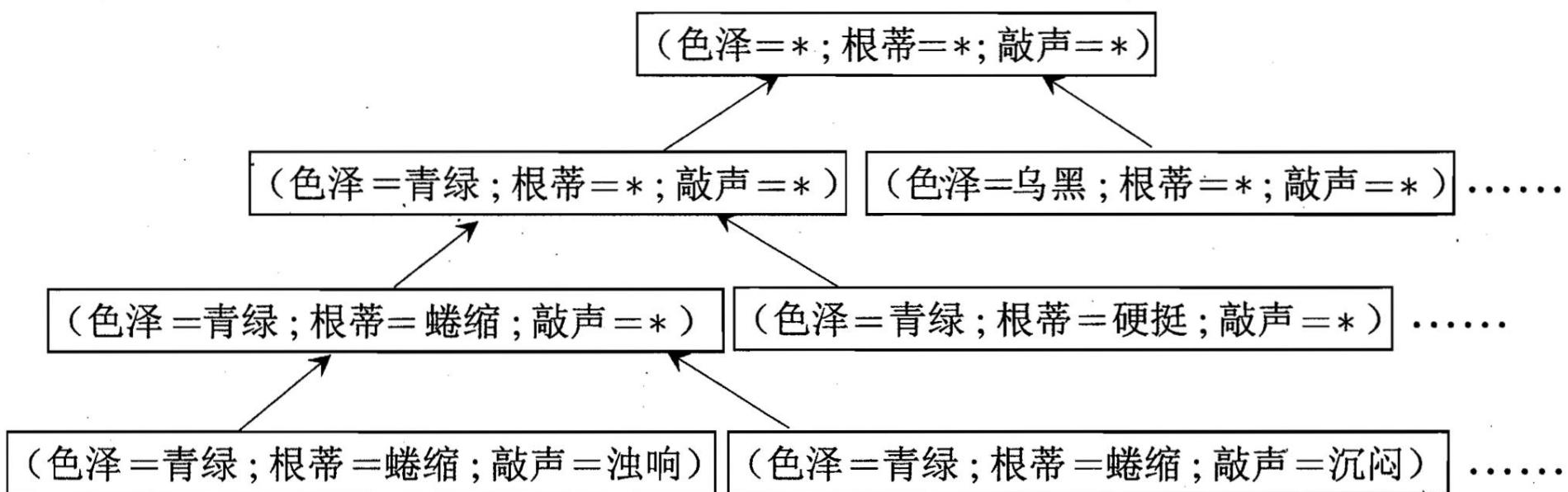


图 1.1 西瓜问题的假设空间

泛化能力：经验误差与过拟合

- 机器学习的目标是使学得的模型能很好地适用于“新样本”

- 相关术语

- 错误率 (error rate)

- 如果 m 个样本中 a 个样本分类错误，则 错误率 $E = a/m$

- 精度 (accuracy)

- 精度 = $1 - \text{错误率} = 1 - a/m$

- 误差 (error)

- 学习器的预测输出与样本的真实输出之间的差异

- 训练集上：训练误差 (training error)、经验误差 (empirical error)

- 测试集上：泛化误差 (generalization error)

泛化能力：经验误差与过拟合

• 误差、目标函数、损失函数 (Loss Function)

- 均方损失 (Squared Loss) : 分类问题

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$

- 绝对值损失 (Absolute Loss) : 分类问题

$$\ell(y, \hat{y}) = |y - \hat{y}|$$

- 二值损失 (Binary Loss) : 回归问题

$$\ell(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

泛化能力：经验误差与过拟合

• 优化 (Optimization)

- 理想情形：最小化期望误差 (Expected Error)

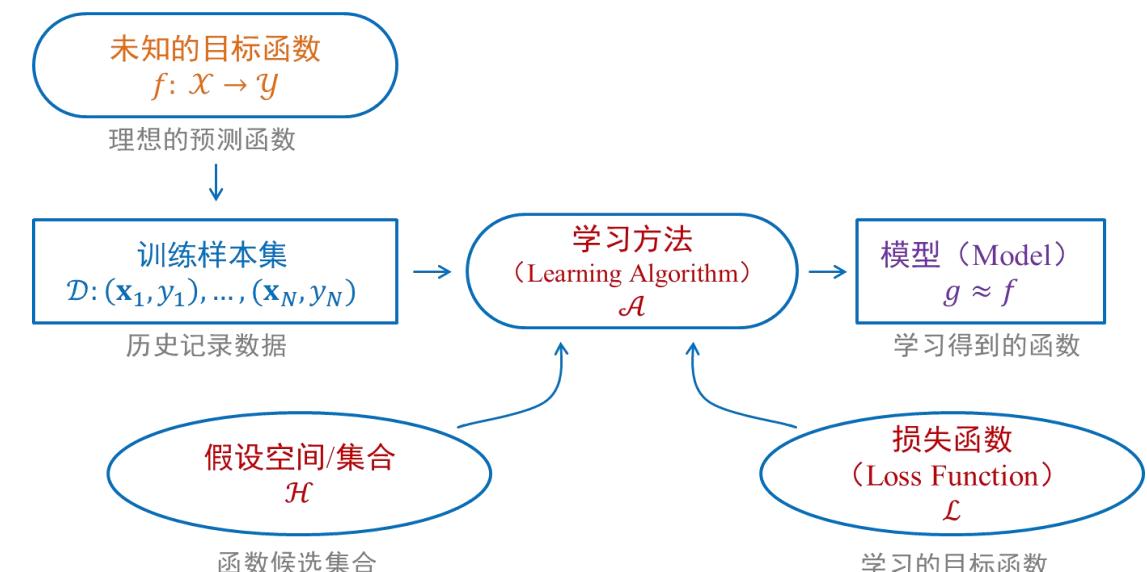
$$\epsilon \triangleq \mathbb{E}_{(x,y) \sim \mathcal{D}} [\ell(y, f(x))] = \sum_{(x,y)} \mathcal{D}(x, y) \ell(y, f(x))$$

- 学习过程，最小化训练误差 (Training Error)

$$\hat{\epsilon} \triangleq \frac{1}{N} \sum_{n=1}^N \ell(y_n, f(x_n))$$

• 问题

- 过拟合 (Overfitting)
- 欠拟合 (Underfitting)



泛化能力：经验误差与过拟合

有多种因素可能导致过拟合，其中最常见的情况是由于学习能力过于强大，以至于把训练样本所包含的不太一般的特性都学到了，而欠拟合则通常是由学习能力低下而造成的。

- 欠拟合 underfitting
- 过拟合 overfitting

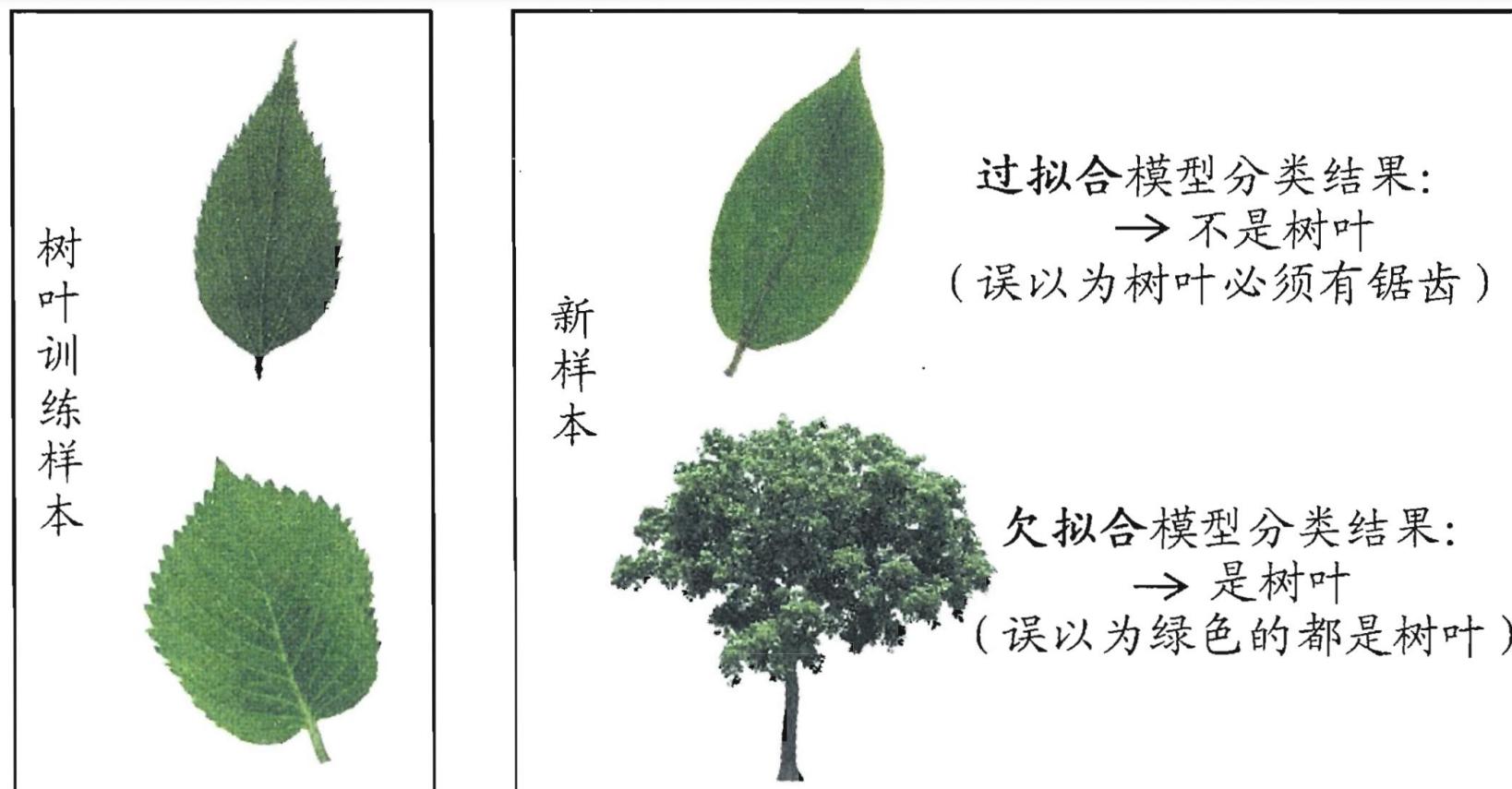
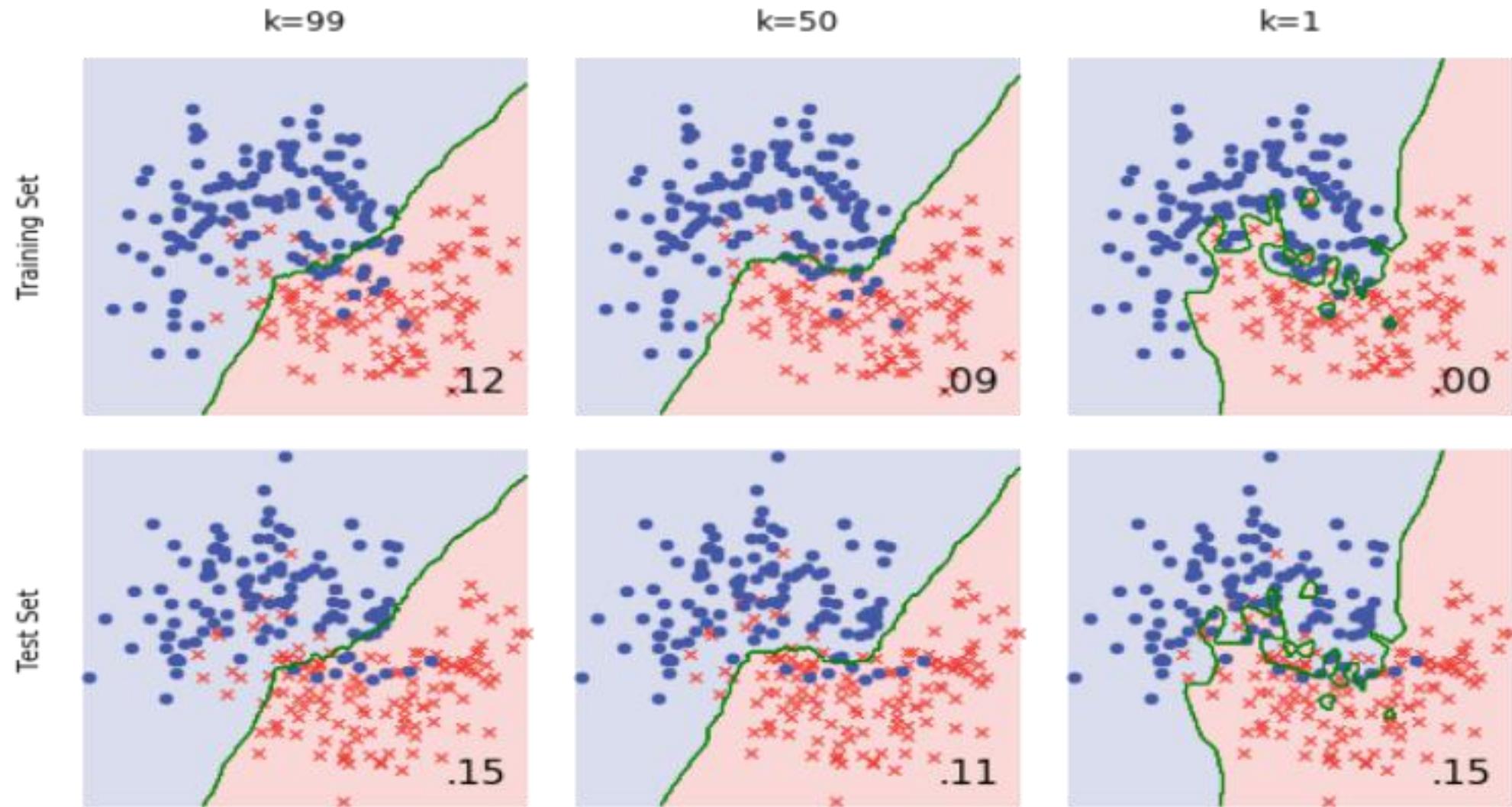


图 2.1 过拟合、欠拟合的直观类比

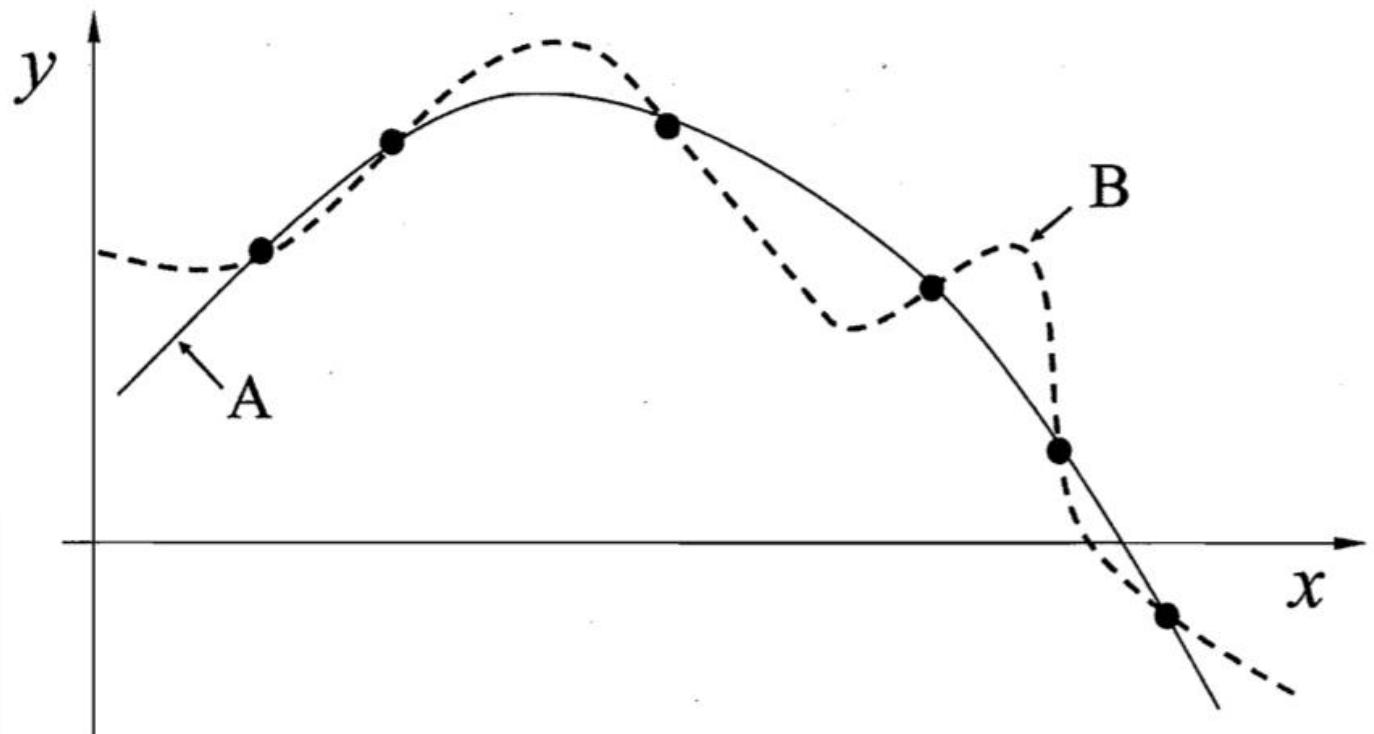
泛化能力：经验误差与过拟合

- 例：聚类，欠拟合 vs 过拟合



归纳偏好 (inductive bias)

算法在学习过程中对某种类型假设的偏好，称为“归纳偏好”(inductive bias)，或简称为“偏好”。



“奥卡姆剃刀” (Occam's razor):

“若多个假设与观察一致，

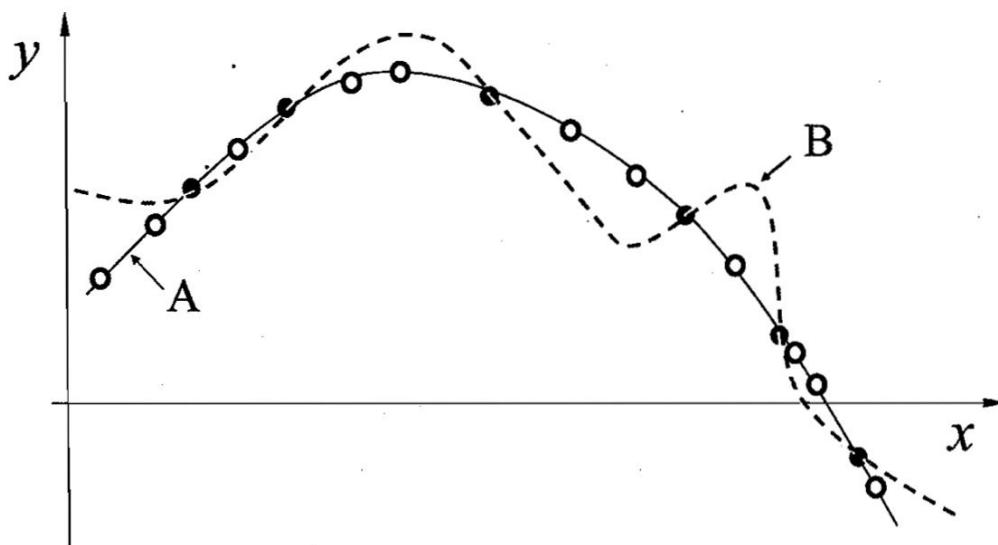
则选最简单的那个”

图 1.3 存在多条曲线与有限样本训练集一致

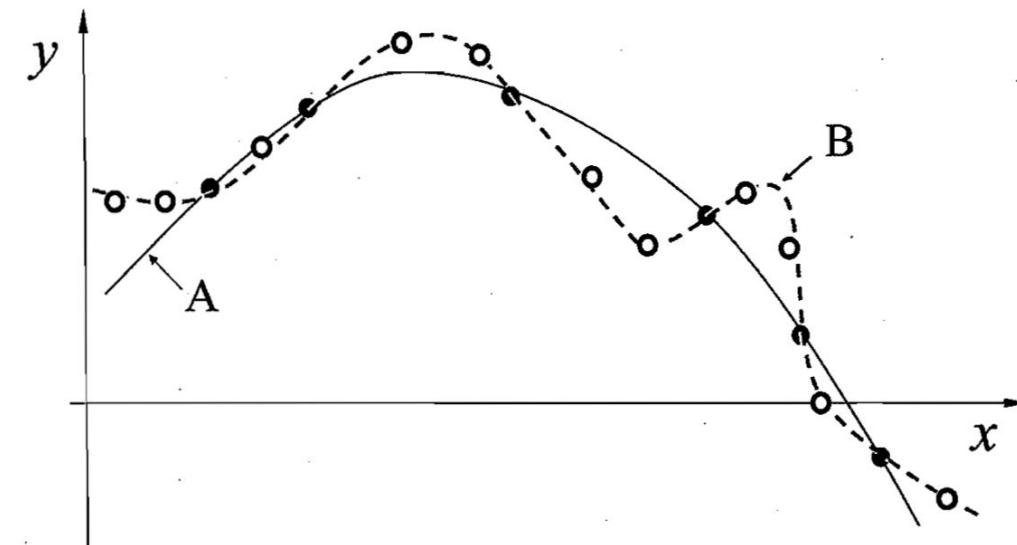
归纳偏好 (inductive bias)

“没有免费的午餐”定理 (No Free Lunch Theorem, 简称 NFL 定理)

- 对于一个学习算法 \mathcal{L}_a ，若它在某些问题上比 \mathcal{L}_b 好，则必定存在另一些问题，在那里 \mathcal{L}_b 比 \mathcal{L}_a 好。这个结论对任何算法均成立。



(a) A 优于 B



(b) B 优于 A

图 1.4 没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

CONTENTS

第一章 绪论

基本术语、泛化能力、归纳偏好

第二章 模型评估与选择

评估方法、性能指标

评估方法

• 测试误差 (testing error)

通常, 我们可通过实验测试来对学习器的泛化误差进行评估并进而做出选择. 为此, 需使用一个“测试集”(testing set)来测试学习器对新样本的判别能力, 然后以测试集上的“测试误差”(testing error)作为泛化误差的近似.

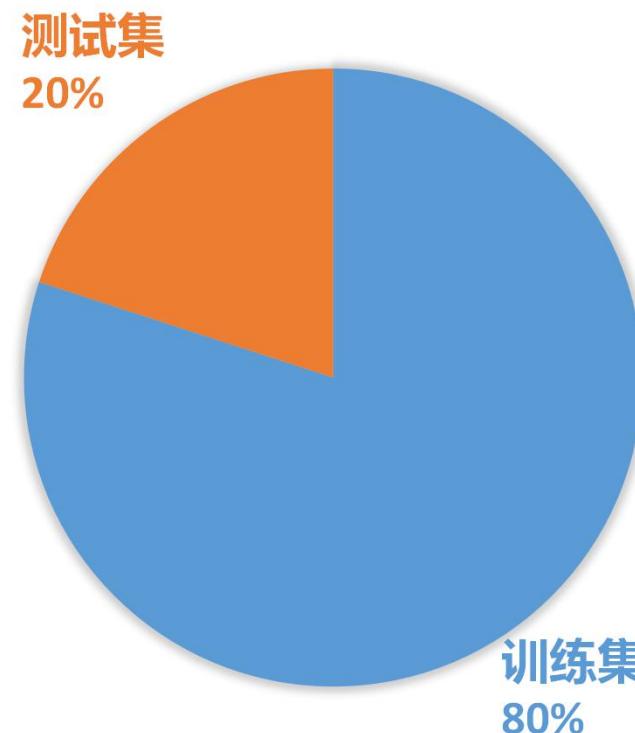
假设测试样本也是从样本真实分布中独立同分布采样而得. 但需注意的是, 测试集应该尽可能与训练集互斥, 即测试样本尽量不在训练集中出现、未在训练过程中使用过.



评估方法

• 1. 留出法

“留出法” (hold-out) 直接将数据集 D 划分为两个互斥的集合, 其中一个集合作为训练集 S , 另一个作为测试集 T , 即 $D = S \cup T$, $S \cap T = \emptyset$. 在 S 上训练出模型后, 用 T 来评估其测试误差, 作为对泛化误差的估计.



评估方法

- 1. 留出法

“留出法” (hold-out) 直接将数据集 D 划分为两个互斥的集合, 其中一个集合作为训练集 S , 另一个作为测试集 T , 即 $D = S \cup T$, $S \cap T = \emptyset$. 在 S 上训练出模型后, 用 T 来评估其测试误差, 作为对泛化误差的估计.

问题：如何划分训练/测试集比较合理？

评估方法

• 1. 留出法

• 分层采样 (stratified sampling)

- 目的：使得训练/测试集的划分保持数据分布的一致性，避免因数据划分过程引入额外的偏差而对结果产生影响。

	示例个数	正例个数	反例个数
数据集 D	1000	600	400
训练集 S	800	480	320
测试集 T	200	120	80

- 注意：单次使用留出法得到的估计结果往往不够稳定，因此一般要采用若干次随机划分，重复进行实验评估后取平均值作为留出法的评估结果。

评估方法

• 2. 交叉验证法 *k-fold cross validation*

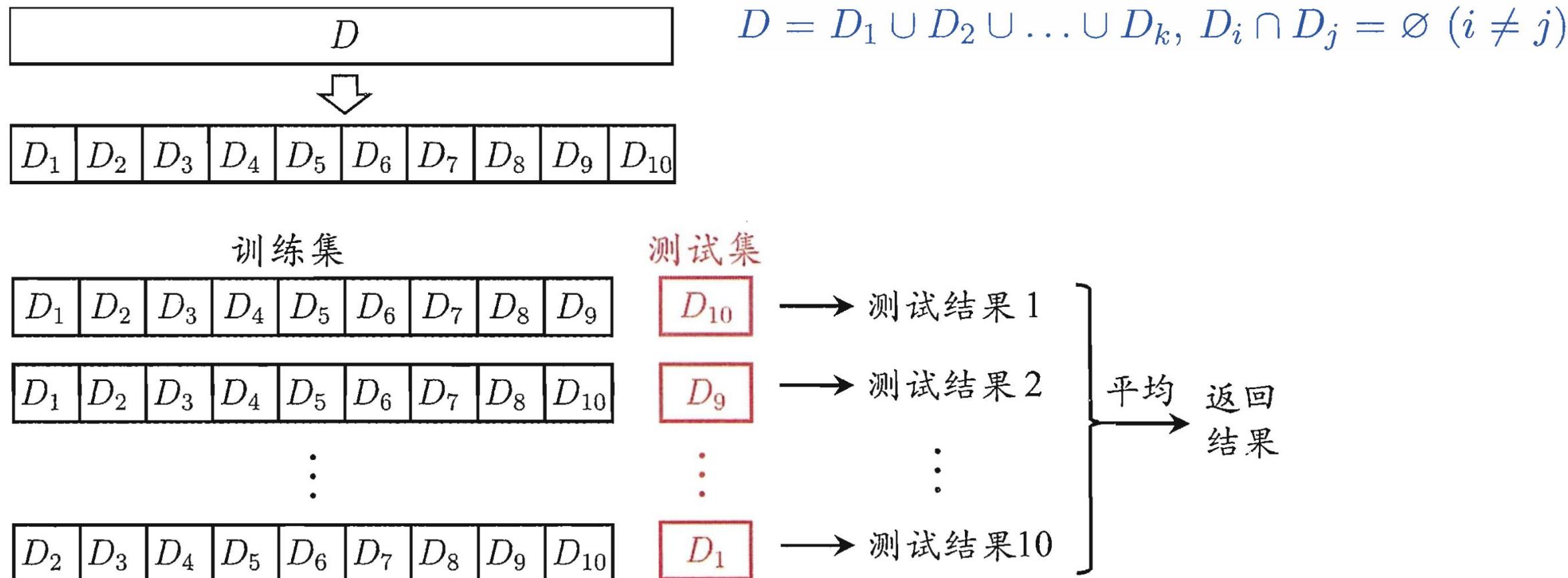


图 2.2 10 折交叉验证示意图

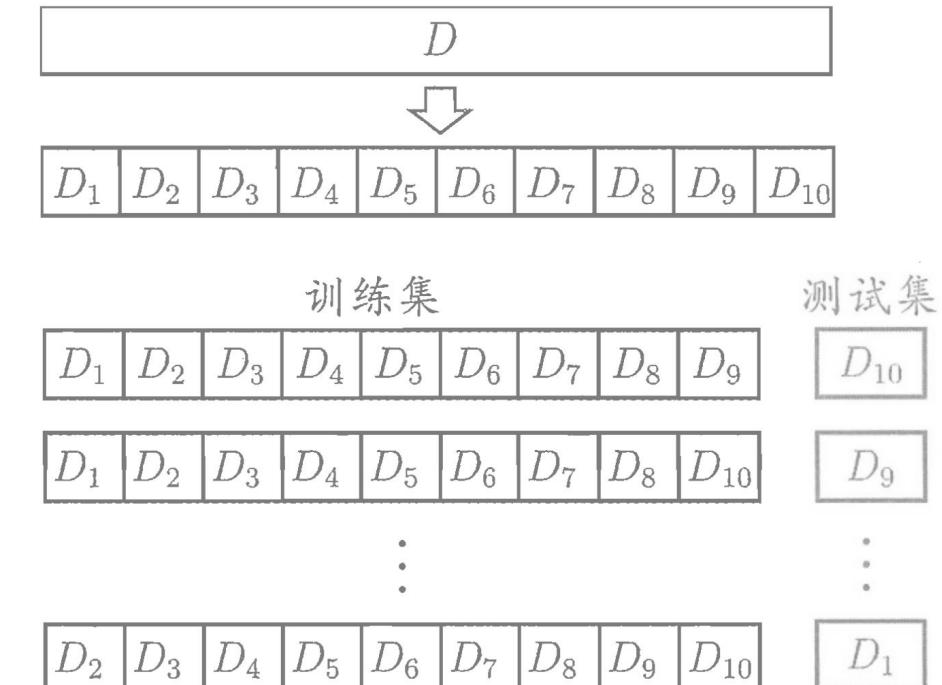
评估方法

• 2. 交叉验证法 **k-fold cross validation**

- 特例：留一法 (leave-one-out, LOO)

- $k = m$ 时
- 不受随机样本划分的影响
- 留一法的评估结果往往认为比较准确

• 缺陷 ?



评估方法

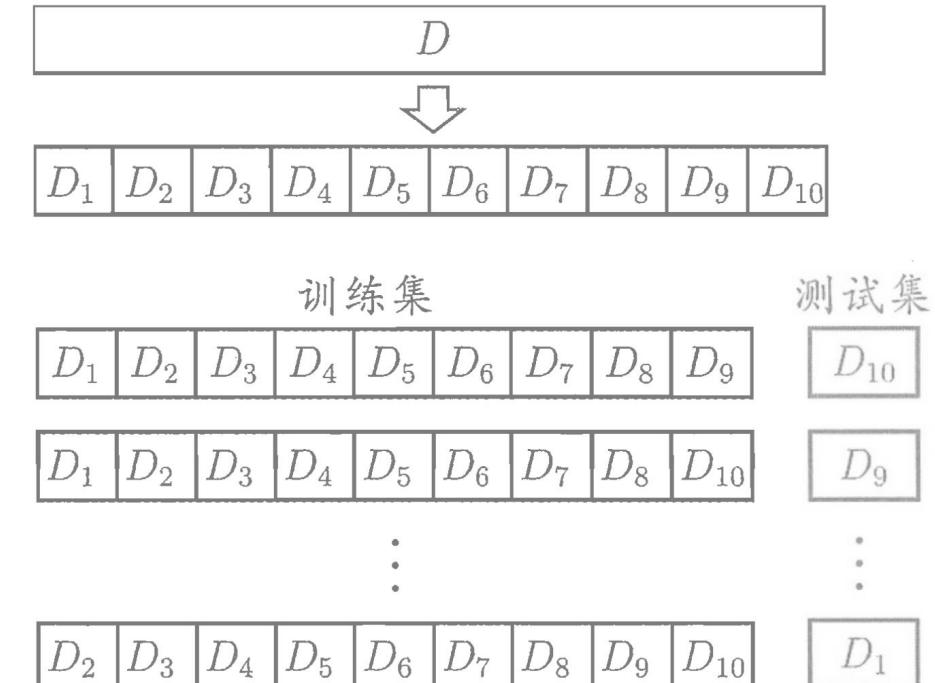
• 2. 交叉验证法 **k-fold cross validation**

- 特例：留一法（leave-one-out, LOO）

- $k = m$ 时
- 不受随机样本划分的影响
- 留一法的评估结果往往认为比较准确

- 缺陷：

- 当m比较大时，需要训练大量的模型，计算开销比较大
- 留一法的估计结果也未必永远比其他评价方法准确（“没有免费的午餐”定理）



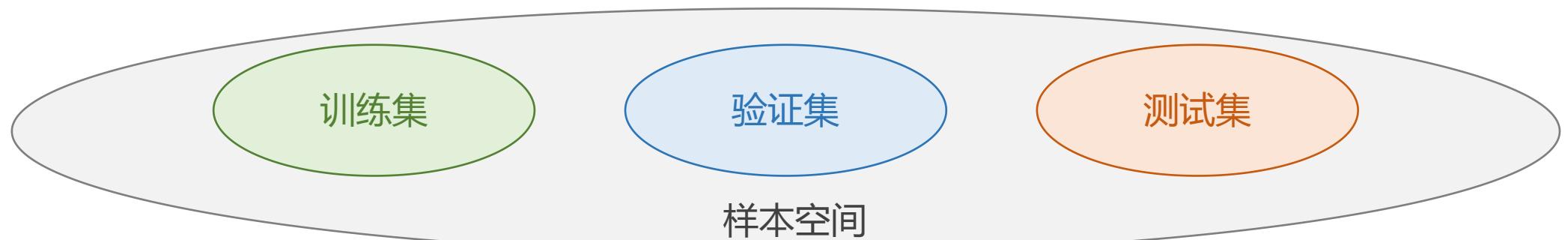
评估方法

- 3. 调参 (parameter tuning) 与最终模型

$$\ell(y, \hat{y}) = (y - \hat{y})^2 + \lambda |y - \hat{y}|$$

- 验证集 (validation set)

- 基于验证集上的性能进行模型选择与调参



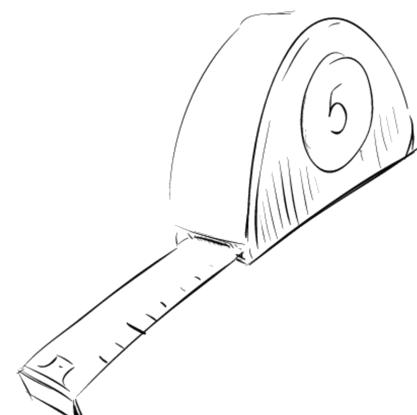
性能度量 (performance measure)

- **Notes**

- 性能度量反映了任务需求
- 对比不同模型的能力时，使用不同的性能度量往往会导致不同的评判结果

- **例如：**

- 房价预测：预测误差或预测精度
- 图像分类：精确度/错误率
- 深度学习模型压缩：模型参数量 或 模型文件大小
- 模型加速：模型运行时间 或 每秒可处理数据量



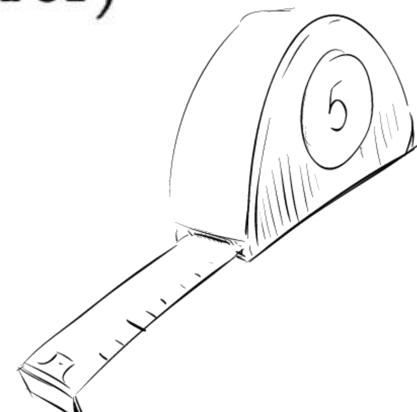
性能度量 (performance measure)

在预测任务中, 给定样例集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$, 其中 y_i 是示例 \mathbf{x}_i 的真实标记. 要评估学习器 f 的性能, 就要把学习器预测结果 $f(\mathbf{x})$ 与真实标记 y 进行比较.

- **回归问题：例如，房价预测？**

回归任务最常用的性能度量是“均方误差” (mean squared error)

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



性能度量 (performance measure)

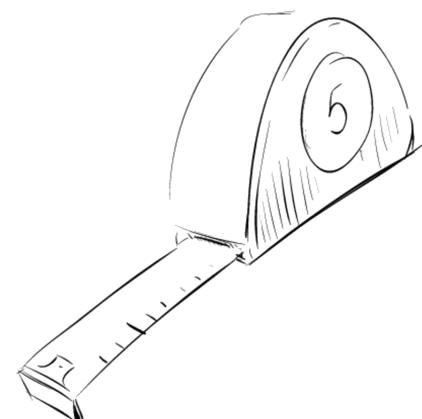
- 分类问题：例如，癌症辅助诊断（判断是否患癌症）

错误率定义为

$$E(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) \neq y_i) .$$

精度则定义为

$$\begin{aligned} \text{acc}(f; D) &= \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(x_i) = y_i) \\ &= 1 - E(f; D) . \end{aligned}$$



性能度量 (performance measure)

• 例：癌症辅助诊断（判断是否患癌症）

- 中国癌症发病率约为： $2/1000$



- 假设有1000名受测者，其中有2位癌症患者；

- 患有癌症病人：2人，正例
- 模型I：全部判断为没有癌症，错误率为？精度为？
- 模型II：2位患者预测正确，另有4位反例预测为患病，其余预测正确。错误率为？精度为？

	错误率	精度
模型I	$2/1000$	99.8%
模型II	$4/1000$	99.6%

模型I 好于 模型II？

性能度量 (performance measure)

• 查准率 (precision)、查全率 (recall)、F1

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例(true positive)、假正例(false positive)、真反例(true negative)、假反例(false negative)四种情形，令 TP 、 FP 、 TN 、 FN 分别表示其对应的样例数，则显然有 $TP + FP + TN + FN =$ 样例总数。分类结果的“混淆矩阵”(confusion matrix)如表 2.1 所示。

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)

性能度量 (performance measure)

- 查准率 (precision)、查全率 (recall)、F1

学习器 A, B, C

查准率 P 与查全率 R 分别定义为

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN}.$$

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)

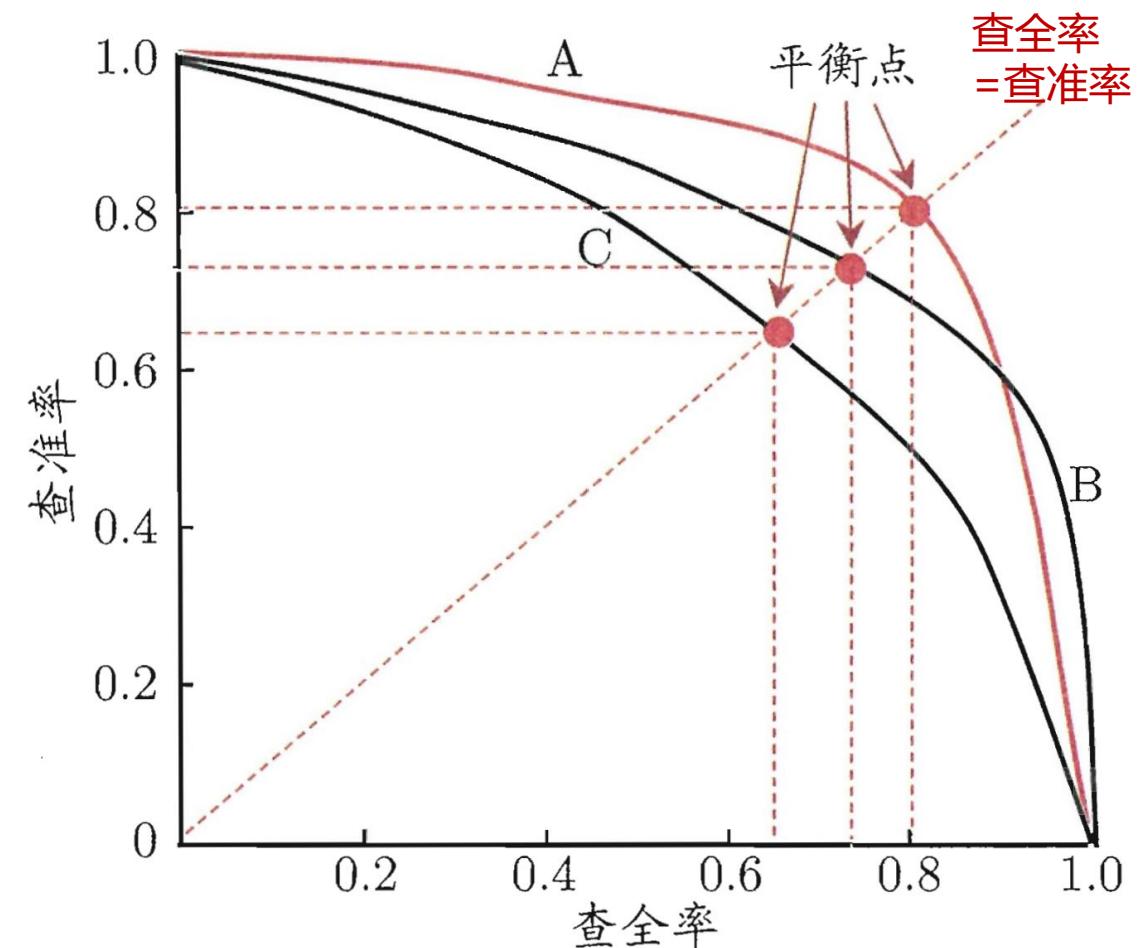


图 2.3 P-R曲线与平衡点示意图

性能度量 (performance measure)

- 查准率 (precision)、查全率 (recall)、F1

学习器 A, B, C

查准率 P 与查全率 R 分别定义为

$$P = \frac{TP}{TP + FP},$$

$$R = \frac{TP}{TP + FN}.$$

$F1$ 度量:

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

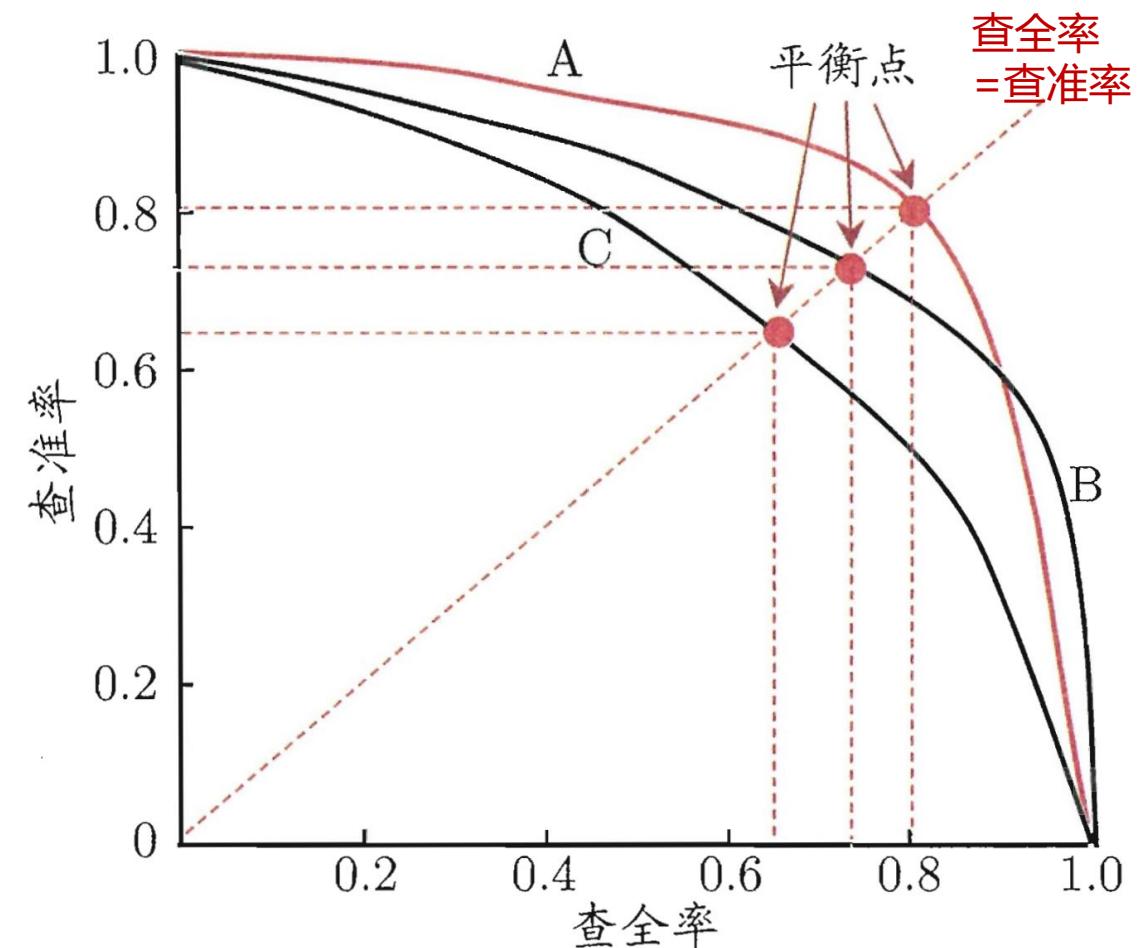


图 2.3 P-R 曲线与平衡点示意图

性能度量 (performance measure)

• 例：癌症辅助诊断（判断是否患癌症）

- 假设有1000名受测者，其中有2位癌症患者；

- 患有癌症病人：2人，正例
- 模型I：全部判断为没有癌症
- 模型II：2位患者预测正确，另有4位反例预测为患病，其余预测正确。

$$P = \frac{TP}{TP + FP} ,$$

$$R = \frac{TP}{TP + FN} .$$

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

	错误率	精度	查准率 P	查全率 R	F1
模型I	2 / 1000	99.8%			
模型II	4 / 1000	99.6%			

性能度量 (performance measure)

• ROC 与 AUC

“受试者工作特征” (Receiver Operating Characteristic)

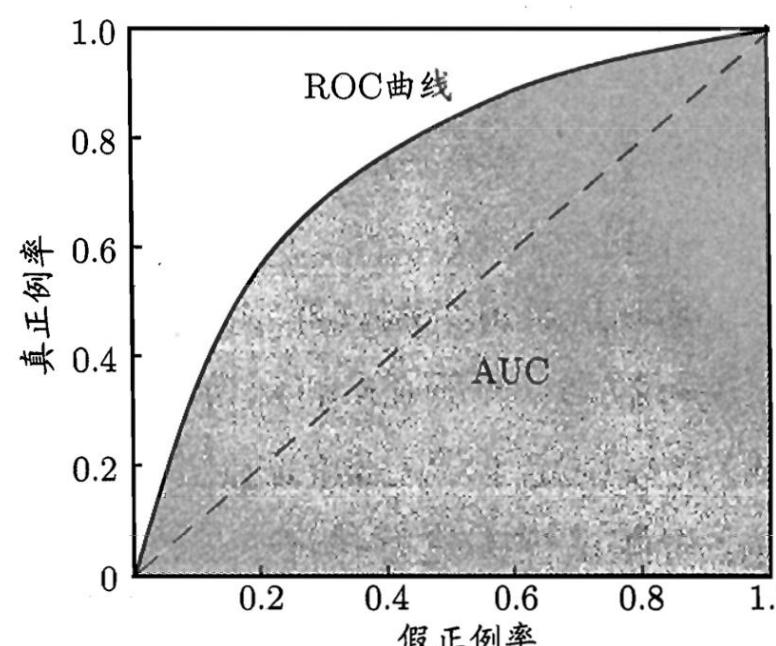
很多学习器是为测试样本产生一个实值或概率预测，然后将这个预测值与一个分类阈值(threshold)进行比较，若大于阈值则分为正类，否则为反类。

$$TPR = \frac{TP}{TP + FN},$$

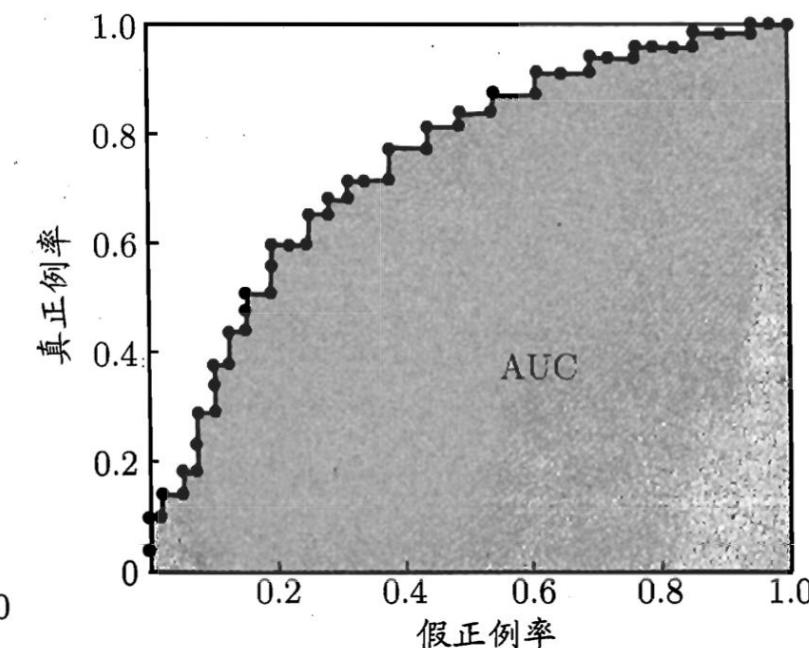
$$FPR = \frac{FP}{TN + FP}.$$

• AUC : ROC曲线下面积

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_i + y_{i+1})$$



(a) ROC 曲线与 AUC



(b) 基于有限样例绘制的 ROC 曲线与 AUC

图 2.4 ROC 曲线与 AUC 示意图

CONTENTS

第一章 绪论

基本术语、泛化能力、归纳偏好

第二章 模型评估与选择

评估方法、性能指标

小结

• 基本术语

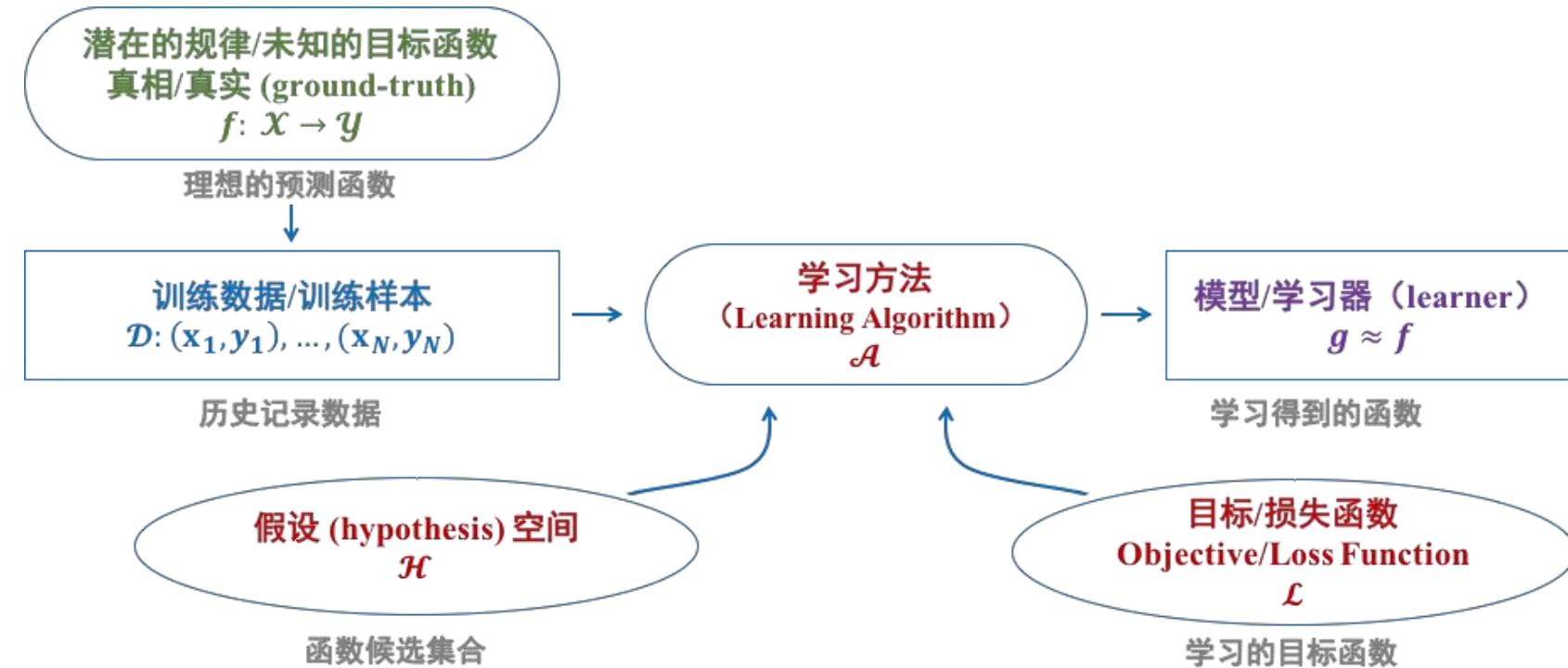
- 示例/样本
- 模型/学习器
- 假设空间
- 过拟合/欠拟合

• 评估方法

- 留出法、交叉验证法
- 调参、验证集

• 性能度量

- 均方误差、错误率、精度、查全率、查准率、F1、ROC、AUC



练习题

- 2.1 数据集包含 1000 个样本, 其中 500 个正例、500 个反例, 将其划分为包含 70% 样本的训练集和 30% 样本的测试集用于留出法评估, 试估算共有多少种划分方式.
- 2.2 数据集包含 100 个样本, 其中正、反例各一半, 假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别(训练样本数相同时进行随机猜测), 试给出用 10 折交叉验证法和留一法分别对错误率进行评估所得的结果.
- 2.4 试述真正例率(TPR)、假正例率(FPR)与查准率(P)、查全率(R)之间的联系.

END

模型评估与选择

Model Evaluation and Selection



俞俊、高飞、谭敏、余宙、匡振中

{yujun, gaofei, tanmin, yuz, zzkuang}@hdu.edu.cn

<http://mil.hdu.edu.cn>

版本空间（version space）

需注意的是，现实问题中我们常面临很大的假设空间，但学习过程是基于有限样本训练集进行的，因此，可能有多个假设与训练集一致，即存在着一个与训练集一致的“假设集合”，我们称之为“版本空间”（version space）。

表 1.1 西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

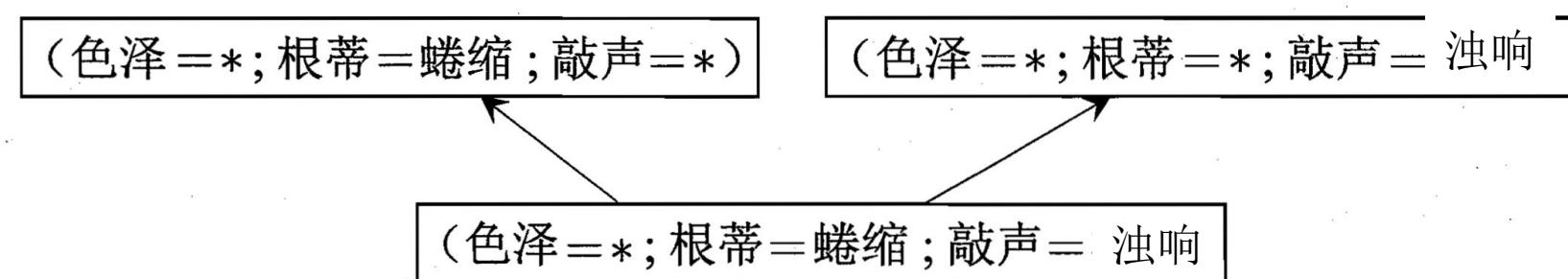


图 1.2 西瓜问题的版本空间

版本空间（version space）

习题

1.1 表 1.1 中若只包含编号为 1 和 4 的两个样例，试给出相应的版本空间。

表 1.1 西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
4	乌黑	稍蜷	沉闷	否

版本空间（version space）

习题

1.1 表 1.1 中若只包含编号为 1 和 4 的两个样例，试给出相应的版本空间。

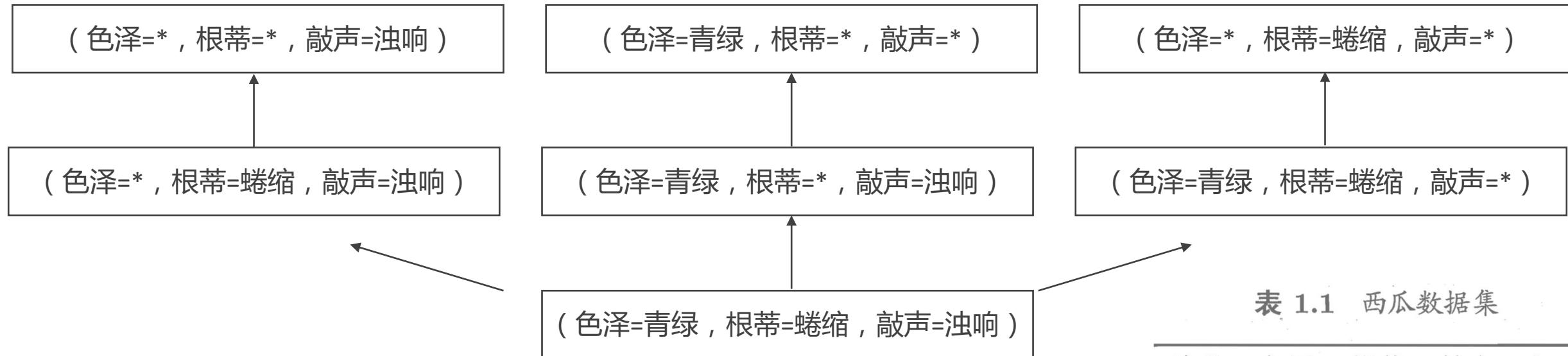


表 1.1 西瓜数据集

编号	色澤	根蒂	敲聲	好瓜
1	青绿	蜷缩	浊响	是