



山東財經大學

Shandong University of Finance and Economics

| 计算机科学与技术学院

School of Computer Science and Technology

MACHINE  
LEARNING

机器学习



---

# 第二章：模型评估 与选择

---

---

# 大纲

---

□ 过拟合与欠拟合

□ 性能度量

□ 评估方法

□ 比较检验

□ 偏差与方差

---



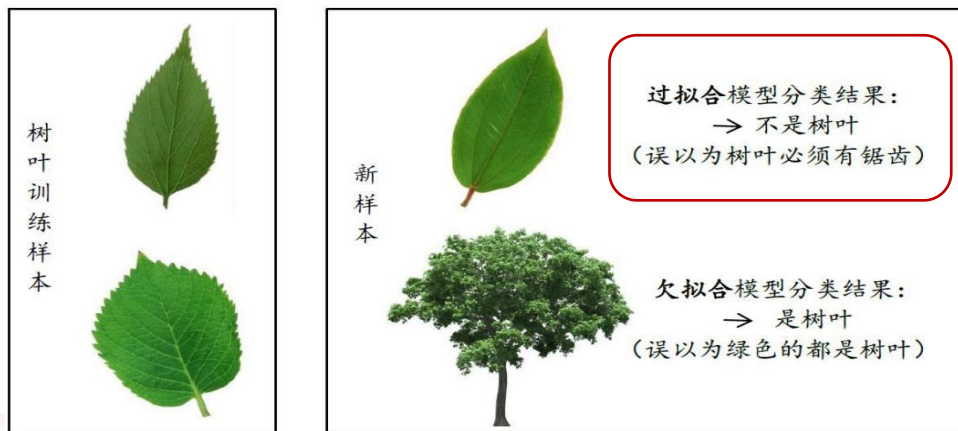
# 过拟合 与 欠拟合

## 机器学习中的问题

### □ 过拟合（过配）

- ✓ 学习器把训练样本学习的“太好”，将训练样本本身的特点当做所有样本的一般性质，以至于把训练样本所包含的不太一般的特性都学到了，导致泛化性能下降
- ✓ 过拟合，是机器学习面临的关键障碍
- ✓ 无法彻底避免，只能“缓解”或减小其风险
- ✓ 缓解方式：通过加正则项，优化目标函数

### □ 欠拟合（欠配）



过拟合、欠拟合的直观类比

# 过拟合 与 欠拟合

## 机器学习中的问题

### □ 过拟合（过配）

- ✓ 把训练样本本身特点 当做所有样本都会具有的一般性质
- ✓ 缓解方式：通过加正则项，优化目标函数

### □ 欠拟合（欠配）

- ✓ 学习能力低下而造成的，对训练样本的一般性质尚未学好
- ✓ 比较容易克服，解决方式
  - 决策树
    - ✓ 拓展分支
  - 神经网络
    - ✓ 增加训练轮数



过拟合、欠拟合的直观类比

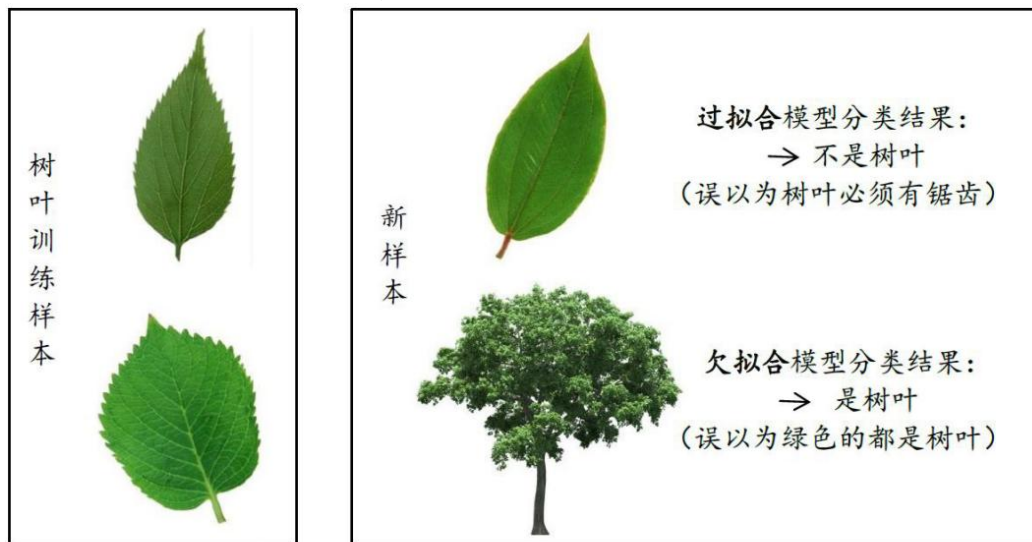
# 过拟合 与 欠拟合

## 过拟合

- 学习器把训练样本本身特点作为所有潜在样本都会具有的一般性质.

## 欠拟合

- 训练样本的一般性质, 尚未被学习器学好.



过拟合、欠拟合的直观类比

# 大纲

---

□ 过拟合与欠拟合

□ 性能度量

□ 评估方法

□ 比较检验

□ 偏差与方差

---

---

# 性能度量

□ **现实任务中**，有多种学习算法供选择，同一个算法有多种参数配置

## ➤ 理想的解决方案

- 评估候选模型的泛化误差，选择**泛化误差最小**的那个模型

## ➤ 实际上

- 然而，由于事先并不知道**新样本**的特征，无法直接获得泛化误差，**无法得到**泛化误差小的学习器。
- 我们只能从**训练样本**中尽可能学出适用于所有潜在样本的**普遍规律**，在遇到新样本时，才能做出正确的判别。



# 性能度量

□ **现实任务中**，有多种学习算法供选择，同一个算法有多种参数配置

## ➤ 实际上

- 然而，由于事先并不知道**新样本**的特征，无法直接获得泛化误差，**无法得到**泛化误差小的学习器。
- 我们只能从**训练样本**中尽可能学出适用于所有潜在样本的**普遍规律**，在遇到新样本时，才能做出正确的判别。
- 实际上：我们努力使**经验误差**最小化，获得在训练集上表现很好的学习器
- 同时，训练误差又由于**过拟合**现象的存在而不适合作为标准，那么，在现实中如何 **评估** 模型的 **泛化性能**？

# 性能度量

## □ 性能度量

- 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方法，往往会导致不同的评判结果

## □ 预测任务：回归、分类

- 给定样例集  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ ，评估学习器  $f$  的性能，即把预测结果  $f(x)$  和真实标记  $y$  进行比较.

# 性能度量

## □ 预测任务：回归、分类

- 给定样例集  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ ，评估学习器  $f$  的性能，即把预测结果  $f(x)$  和真实标记  $y$  进行比较。

## □ 回归任务

- **误差**：样本**真实输出**与**预测输出**之间的差异
  - **训练(经验)误差**：训练集上
  - **泛化误差**：新样本上
- 最常用的性能度量是“均方误差”：

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

# 性能度量

## □ 分类任务

错误率和精度、查准率和查全率、代价敏感错误率和代价曲线

### ● 错误率和精度，是最常用的两种性能度量

#### ➤ 错误率

- 分错样本 占 样本总数（ $m$ 个）的比例

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

#### ➤ 精度

- 分对样本 占 样本总数（ $m$ 个）的比例，即 1-错误率

$$\text{acc}(f; D) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(f(\mathbf{x}_i) = y_i) = 1 - E(f; D)$$

# 性能度量

## □ 分类任务：查准率、查全率

- 信息检索、Web搜索等场景中，经常需要衡量 **正例被预测出来的比率** 或者 **预测出来的正例中正确的比率**，此时查准率和查全率比错误率和精度更适合。
- 例如在信息检索中，我们会关心“检索出的信息中有多少是用户感兴趣的”，“用户感兴趣的信息中有多少被检索出来了”
- **查准率P（准确率）**：正例被预测出来的比率

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

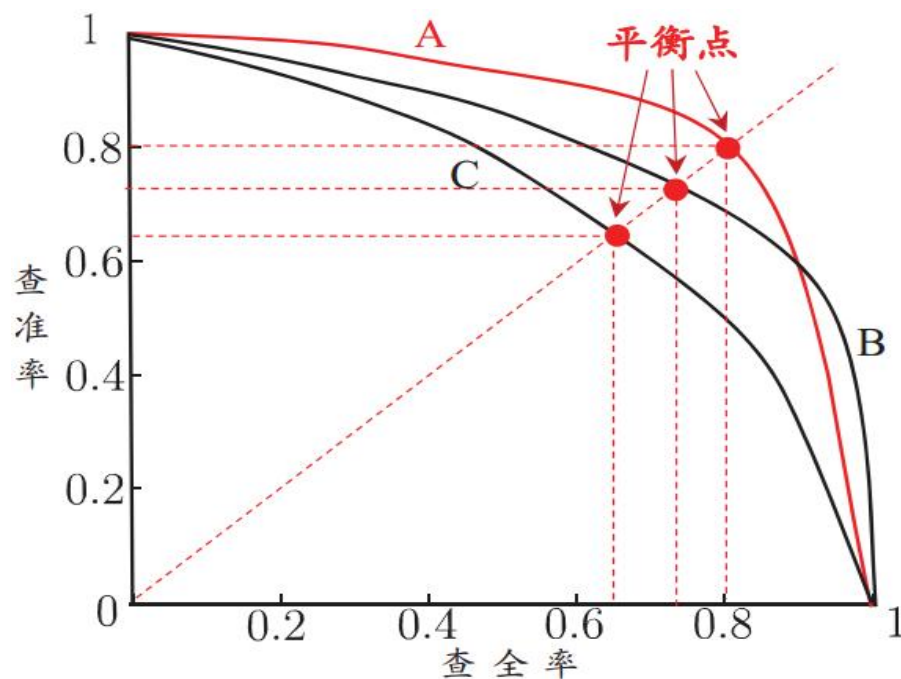
- **查全率R（召回率）**：预测出来的正例中正确的比率

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

# 性能度量

## □ 分类任务：查准率、查全率

根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”



**平衡点：**曲线上“查准率=查全率”时的取值，可用于度量P-R曲线有交叉的分类器性能高低

P-R曲线与平衡点示意图

# 性能度量

## □ 分类任务：代价敏感错误率和代价曲线

- 现实任务，不同类型的错误所造成的后果很可能不同，为了权衡不同类型错误所造成的不同损失，可为错误赋予非均等代价。
  - 错误地把患者诊断为健康人，使患者丧失了拯救生命的最佳时机；错误地把健康人诊断为患者，也许只是增加了进一步检查的麻烦
  - 门禁系统错误地把可通行人员拦在门外，将使得用户体验不佳；但错误地让陌生人进入，则会造成严重的安全事故。

# 性能度量

## □ 分类任务：代价敏感错误率和代价曲线

➤ 以二分类为例，可根据领域知识设定“代价矩阵”。

表 2.2 二分类代价矩阵

真实类别	预测类别	
	第 0 类	第 1 类
第 0 类	0	$cost_{01}$
第 1 类	$cost_{10}$	0

$cost_{ij}$  将第*i*类样本 预测为 第*j*类样本的代价。一般来说 $cost_{ii}=0$ ,  $cost_{01} > cost_{10}$   
损失程度越大,  $cost_{01}$  与  $cost_{10}$  值的差别越大。



# 大纲

---

□ 过拟合与欠拟合

□ 性能度量

□ 评估方法

□ 比较检验

□ 偏差与方差

---

The bottom of the slide features three horizontal bars of different colors: a light pink bar, a light red bar, and a light green bar.

# 评估方法

---

## □ 性能度量

- 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

## □ 模型选择

通过实验，测试学习器的泛化性能、时间开销、存储开销、可解释性等方面的因素，并进行评估，进而做出选择

# 评估方法

□ 数据集细分为：训练数据 = 训练集 + 测试集

□ 评估步骤

1. 使用“测试集”测试学习器对新样本的判别能力

➤ 例如

- 老师出了10道习题供同学们练习，考试时，老师又用同样的这10道题作为试题，考试成绩能否有效反映出同学们学得好不好呢？
- 希望得到泛化性能强的模型，好比是希望同学们对课程学得很好、获得了对所学知识“举一反三”的能力
- 训练样本相当于给同学们练习的习题，测试过程则相当于考试。
- 若测试样本被用作训练了，则得到的将是过于“乐观”的估计结果。

# 评估方法

## □ 评估步骤

1. 使用“测试集” 测试学习器对新样本的判别能力
  - ✓ 测试集，是从样本真实分布中独立采样获得
  - ✓ 测试集和训练集中的样本尽量互斥，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.
  - ✓ 通常将包含个  $m$  样本的数据集拆分成训练集 $S$  和 测试集 $T$ ：  
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.
2. 将测试集上的“测试误差” 作为泛化误差的近似，对模型进行评估

# 评估方法

---

## 常见的几种模型评估方法

- 留出法
- 交叉验证法
- 自助法

# 评估方法

## □ 留出法

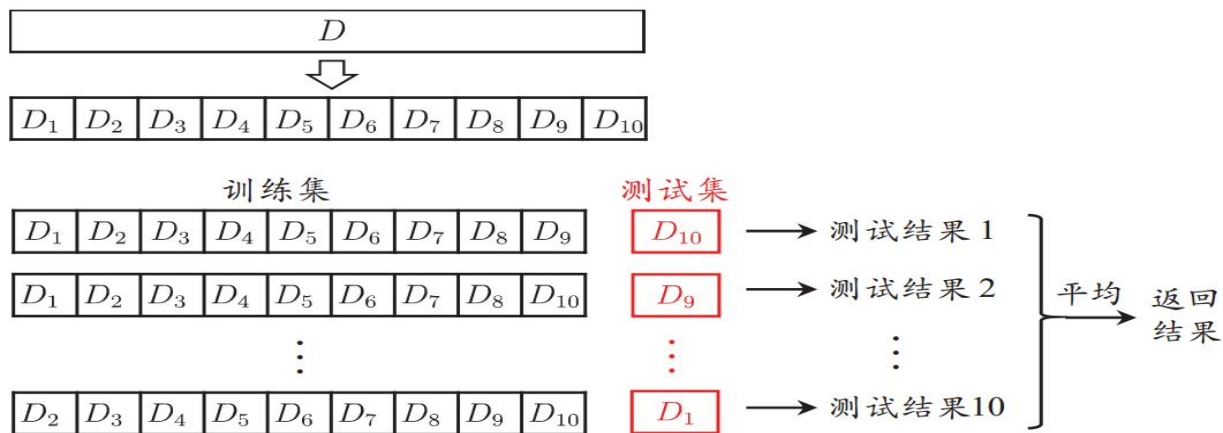
数据集 $D$ 划分为训练集 $S$  和测试集 $T$  两个互斥集合

- ✓ 训练/测试集划分要尽可能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，而对最终结果产生影响
  - 例如，在分类任务中，采用分层采样，即尽量保持样本的类别比例相似。
    - eg. 训练集 $S$  包含300个正例和300个反例；  
测试集 $T$  包含100个正例和100个反例。
- ✓ 缺点：  
只能得到一个评估值。

# 评估方法

## 交叉验证法

1. 将数据集分层采样，划分为  $k$  个大小相似的互斥子集。 $k$  常取 10 ( $k$ -fold cross validation)
  2. 每次用  $k-1$  个子集的并集作为训练集，余下的子集作为测试集
  3. 将第2步的计算过程重复  $p$  次，最终返回  $p$  个测试结果的均值.
- ✓ 可以得到  $p$  个评估值。
  - ✓ 评估结果的稳定性和保真性在很大程度上取决于  $k$  的取值。



10 折交叉验证示意图

# 评估方法

❑ 留出法

❑ 交叉验证法

缺陷

保留了一部分样本用于测试，实际评估模型所使用的训练集比 $D$ 小，引入了一些因训练样本规模不同，而导致的估计偏差

❑ 自助法

- 以自助采样法为基础[Efron and Tibshirani, 1993]
- 对含有 $m$ 个样本的数据集 $D$ 每次随机挑选一个样本拷贝至 $D'$ ，重复 $m$ 次，得到含 $m$ 个样本的**训练集** $D'$
- 初始数据集 $D$ 中约有36.8%的样本未出现在 $D'$ 中。
- $D \setminus D'$ 用作**测试集**。



# 评估方法

## □ 自助法

- 初始数据集 $D$ 中约有36.8%的样本未出现在 $D'$ 中。
- 实际模型与预期模型都使用 $m$ 个训练样本，约有1/3的样本没在训练集中出现，且用于测试。这样的测试结果称“包外估计”。
- 在数据集较小、难以有效划分 $S$ 和 $T$ 时很有用。
- ✓ 缺点：  
改变了初始数据集的分布，引入估计偏差。

# 大纲

---

□ 过拟合与欠拟合

□ 性能度量

□ 评估方法

□ 比较检验

□ 偏差与方差

---

The bottom of the slide features three horizontal bars of different colors: a light pink bar, a light red bar, and a light green bar.

# 比较检验

## □ 关于性能比较：

- 测试性能并不等于泛化性能
- 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，即便使用相同的参数设置在同一个测试集上多次运行，其结果也会有不同。

**直接选取相应评估方法在相应度量下比大小的方法，不可取！**

## □ 假设检验 为学习器性能比较提供了重要依据

- 基于假设检验的结果，我们可以推断出：  
若在测试集上，观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握有多大。

# 假设检验---单个学习器

单个学习器性能比较方法：二项检验、 $t$ -检验

## □ 二项检验

➤ 针对 一次 训练/测试，使用 留出法 估计 错误率

泛化错误率  $\epsilon$ ：学习器在一个样本上犯错的概率为 $\epsilon$

测试错误率  $\hat{\epsilon}$ ：在  $m$  个测试样本中，恰有  $\hat{\epsilon} \times m$  个被误分类

➤ 假定测试样本从样本总体分布中独立采样而来，我们可以使用“二项检验”对  $\epsilon \leq \epsilon_0$  进行假设检验，若测试错误率小于

$$\bar{\epsilon} = \max \epsilon \quad \text{s.t.} \quad \sum_{i=\epsilon_0 \times m + 1}^m \binom{m}{i} \epsilon^i (1 - \epsilon)^{m-i} < \alpha$$

则在  $\alpha$  的显著度下，假设  $\epsilon \leq \epsilon_0$  不能被拒绝

即也能以  $1-\alpha$  的置信度认为：模型的泛化错误率不大于  $\epsilon_0$

# 假设检验---单个学习器

## 单个学习器性能比较方法：二项检验、 $t$ -检验

对应的，面对多次重复留出法或者交叉验证法进行多次训练/测试时可使用“ $t$ 检验”。

假定得到了 $k$ 个测试错误率， $\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_k$ ，假设 $\epsilon = \epsilon_0$  对于显著度 $\alpha$ ，若  $[t_{-\alpha/2}, t_{\alpha/2}]$  位于临界范围  $|\mu - \epsilon_0|$  内，则假设不能被拒绝，即可认为泛化错误率 $\epsilon = \epsilon_0$ ，其置信度为 $1 - \alpha$ 。否则可拒绝该假设。

□ 二项检验

□  $t$ -检验( $t$ -test)



单个学习器泛化性能的假设进行检验

# 假设检验---Friedman检验

## 基于假设检验的不同学习器性能比较方法

- 交叉验证 $t$ 检验

- McNemar检验：衡量两学习器，分类结果的差别

→ 在一个数据集上，比较两个算法的性能

- Friedman检验

- Nemenyi后续检验

→ 在一组数据集上，对多个算法进行比较

# 大纲

---

□ 过拟合与欠拟合

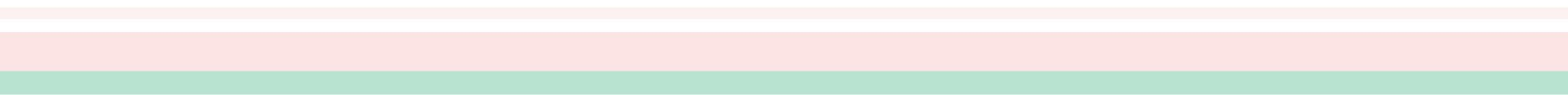
□ 性能度量

□ 评估方法

□ 比较检验

□ 偏差与方差

---

The bottom of the slide features three horizontal bars of different colors: a light pink bar, a light red bar, and a light green bar.

# 偏差与方差

## □ 偏差-方差分解

- 对学习算法期望的 泛化错误率 进行拆解，帮助解释泛化性能。
- 泛化误差 可分解为 偏差、方差与噪声 之和。
  - 偏差度量了学习算法期望预测与真实结果的偏离程度；
    - ✓ 刻画了学习算法本身的拟合能力；
    - ✓ 反映算法的效率
  - 方差度量了同样大小训练集的变动所导致的学习性能的变化；
    - ✓ 刻画了数据扰动所造成的影响；
    - ✓ 反映算法的稳定性
  - 噪声表达了当前任务上，任何学习算法所能达到的期望泛化误差的下界；
    - ✓ 刻画了学习问题本身的难度。



# 偏差与方差

## □ 偏差-方差分解

对测试样本 $x$ , 令 $y_D$ 为 $x$ 在数据集中的标记,  $y$ 为 $x$ 的真实标记,  $f(x; D)$  为训练集 $D$ 上学得模型 $f$ 在 $x$ 上的预测输出。

以回归任务为例, 学习算法的期望预期为:  $\bar{f}(x) = \mathbb{E}_D[f(x; D)]$

➤ 偏差: 期望输出与真实标记的差别

$$bias^2(x) = (\bar{f}(x) - y)^2$$

➤ 方差: 使用样本数目相同的不同训练集产生的方差

$$var(x) = \mathbb{E}_D \left[ (f(x; D) - \bar{f}(x))^2 \right]$$

➤ 噪声:

$$\epsilon^2 = \mathbb{E}_D \left[ (y_D - y)^2 \right]$$

# 偏差与方差

---

## □ 偏差-方差分解

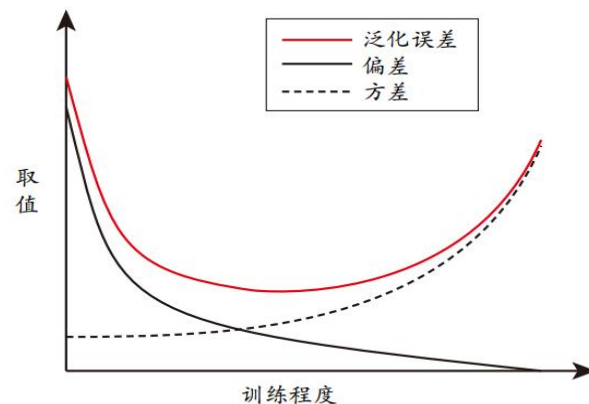
- 泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度所共同决定的。
- 给定学习任务，为了取得好的泛化性能，需要使偏差小(充分拟合数据)而且方差较小(减少数据扰动产生的影响)。

# 偏差与方差

## □ 偏差-方差分解

➤ 一般来说，偏差与方差是有冲突的，称为**偏差-方差窘境**。  
如右图所示，假如我们能控制算法的训练程度：

- **训练不足**时，学习器拟合能力不强，训练数据的扰动，不足以使学习器的拟合能力产生显著变化，此时**偏差主导**泛化错误率；
- 随着**训练程度加深**，学习器拟合能力逐渐增强，**方差**逐渐**主导**泛化错误率；
- **训练充足**后，学习器的拟合能力非常强，训练数据的轻微扰动，都会导致学习器的显著变化，若训练数据自身非全局特性被学到，则会发生**过拟合**。



# 阅读材料

---

- ❑ 自助采样法在机器学习中有重要用途, [Efron and Tibshirani, 1993]对此有详细讨论。
- ❑ ROC曲线在二十世纪八十年代后期被引入机器学习 [Spackman, 1989], AUC则是从九十年代中期起在机器学习领域广为使用 [Bradley, 1997]. [Hand and Till, 2001]将ROC曲线从二分类任务推广到多分类任务. [Fawcett, 2006]综述了ROC曲线的用途。
- ❑ [Drummond and Holte, 2006]发明了代价曲线. 代价敏感学习 [Elkan, 2001; Zhou and Liu, 2006]专门研究非均等代价下的学习。

# 阅读材料

---

- [Dietterich, 1998]指出了常规k折交叉验证法存在的风险,并提出了5\*2折交叉验证法. [Demsar, 2006]讨论了对多个算法进行比较检验的方法.
- [Geman et al., 1992]针对回归任务给出了偏差-方差-协方差分解,后来被简称为偏差-方差分解。但仅基于均方误差的回归任务中推导,对分类任务,由于0/1损失函数的跳变性,理论上推导出偏差-方差分解很困难。已有多种方法可通过试验队偏差和方差进行估计 [Kong and Dietterich, 1995; Kohavi and Wolpert, 1996; Breiman, 1996; Friedman, 1997; Domingos, 2000].

# 知识总结：第2章：模型评估与选择

## □ 训练误差与过拟合，评估方法，性能度量，比较检验，偏差与方差

1. **误差**：样本真实输出与预测输出之间的差异
  - ✓ 训练(经验)误差：训练集上
  - ✓ 泛化误差：新样本上
  - ✓ 机器学习的目标：得到泛化误差小的学习器。
2. **过拟合**：若学习器把训练样本学习的“太好”，将训练样本本身的特点 当做所有样本的一般性质，以至于把训练样本所包含的不太一般的特性都学到了，导致泛化性能下降
3. **欠拟合**：学习能力低下而造成的，对训练样本的一般性质尚未学好。训练数据较少时更容易发生欠拟合。

# 知识总结：第2章：模型评估与选择

## □ 训练误差与过拟合，评估方法，性能度量，比较检验，偏差与方差

### 1. 模型评估方法：

- ✓ 留出法：直接将数据集划分为两个互斥集合，作为训练集和测试集
- ✓ 交叉验证法： $k$  个大小相似的互斥子集（ $k$  最常取10）
- ✓ 自助法：以自助采样法为基础，对数据集 $D$ ，有放回采样 $m$ 次，得到训练集 $D'$ ， $D/D'$ 用做测试集。

### 2. 回归任务最常用的性能度量是“均方误差”：

### 3. 分类任务常用的性能度量：错误率和精度、查准率和查全率、ROC和AUC、代价敏感错误率和代价曲线