2. 课程大纲

1 大数据审计、机器学习绪论+Python 基础及环境配置

2 模型评估与选择

3 线性模型

4 决策树

5 神经网络

6 聚类

7 大数据审计概论

8 审计数据分析

9 报表识别

10 风险评估与审计报告

东北财经大学

DONGBEI UNIVERSITY OF FINANCE & ECONOMICS

第一章 (补)

大数据审计介绍

管理科学与工程学院

谢琛

大数据审计分析概念

大数据

- 大数据是指无法在一定时间内用常规计算机工具进行抓取、管理和处理的数据集合，而是需要在新模式下才具有更强决策能力、洞察能力、优化能力的海量、多样、高增长的信息资产。
- 大数据具有5V性质，即大规模（Volume）、高速增长（Velocity）、多样化（Variety）、低价值密度（Value）、真实性（Veracity）。

审计分析

- 审计分析是指通过对被审计对象相关资料的数据指标进行逻辑推理、因素分解和综合判断，以揭示资料的内在本质并了解其构成要素相互关系的审计方法。
- 常见的审计分析方法包括**比较分析**、**比率分析**、**账户分析**、**账龄分析**、**平衡分析**、**因素分析**等。

大数据审计分析

- 大数据审计分析是指在审计理论的指导下，将**大数据思维**融入到**审计分析**中，使用大数据技术对被审计对象**数据资料进行收集、清洗、整合**，形成符合要求的**审计数据**，并构建**分析模型**对审计**数据进行挖掘**，得到相应数据分析结果，以帮助审计人员了科学、准确做出审计判断的审计过程。

大数据审计分析相关概念

审计信息化：是指将**信息技术手段**应用于审计工作，全面改造审计业务流程、建立并完善审计工作方式、提高审计效率，最终取得良好审计质量的审计能力提升过程。

信息系统审计：是指通过收集和评价相关审计证据，以确定现有信息系统与相关资源能否有效地保护资产、维护数据完整、提供相关和可靠的信息，从而保证组织资源高效利用、组织目标顺利实现等方面做出合理判断的审计过程。

电子数据审计：是指利用计算机技术，对被审计单位的电子数据进行收集、整理和分析，从而发现**审计线索**、**获得审计证据**、得出**审计结论**的审计过程。

大数据审计：是指审计机构遵循**大数据思维**，采用大数据技术，利用数量巨大、来源广泛、格式多样的结构或非结构化数据， 对被审计对象开展全方位、多层次、立体化的审计工作。

大数据审计分析相关概念

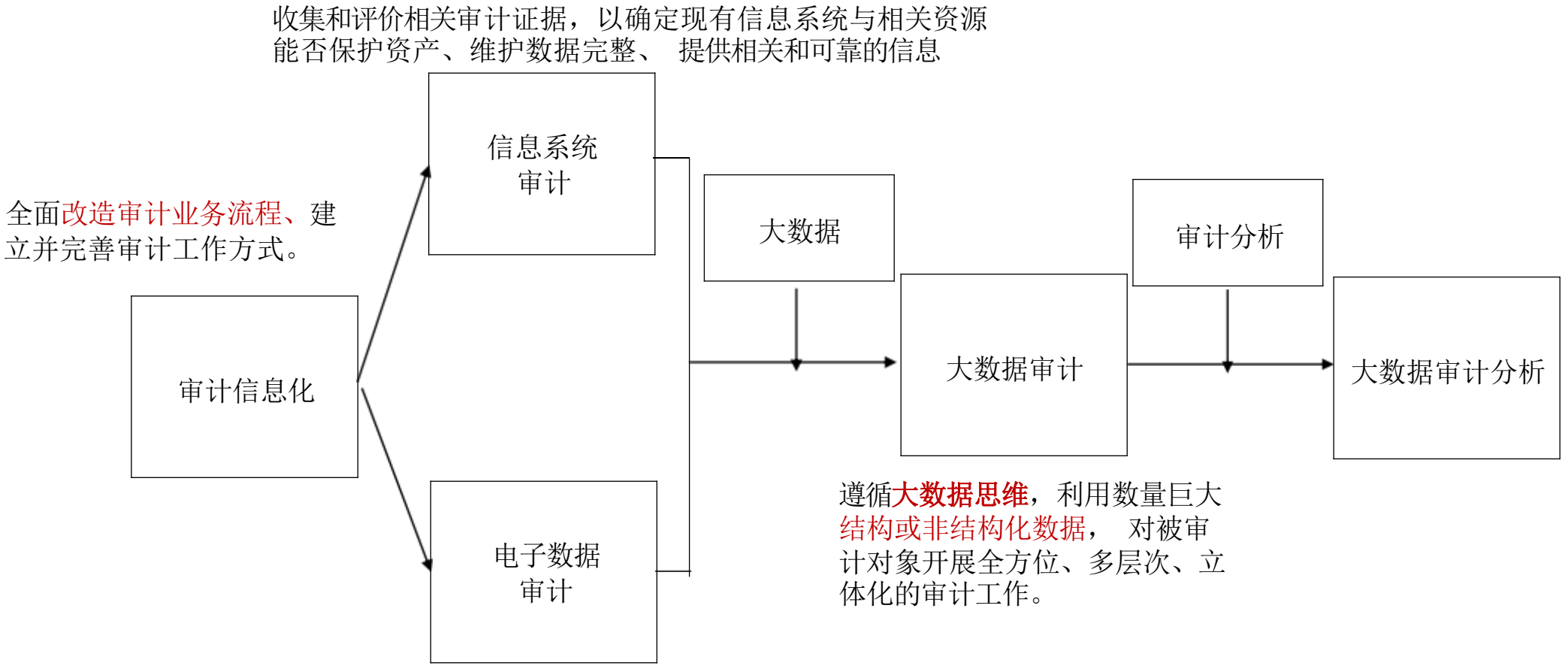


图1-1 各概念之间的逻辑关系

大数据审计分析

大数据审计分析是一种全新的审计工作方法，其内容贯穿数据获取——程序实施——证据分析，具体包括获取海量化的审计数据、实施自动化的审计程序、进行智能化的审计分析。

- 海量化的审计数据：保证了审计证据具有充分性；
- 自动化的审计程序：确保审计流程高效无误的得到执行；
- 智能化的审计分析：可以使审计证据具有恰当性，并最终帮助审计人员得出高质量的审计结论。

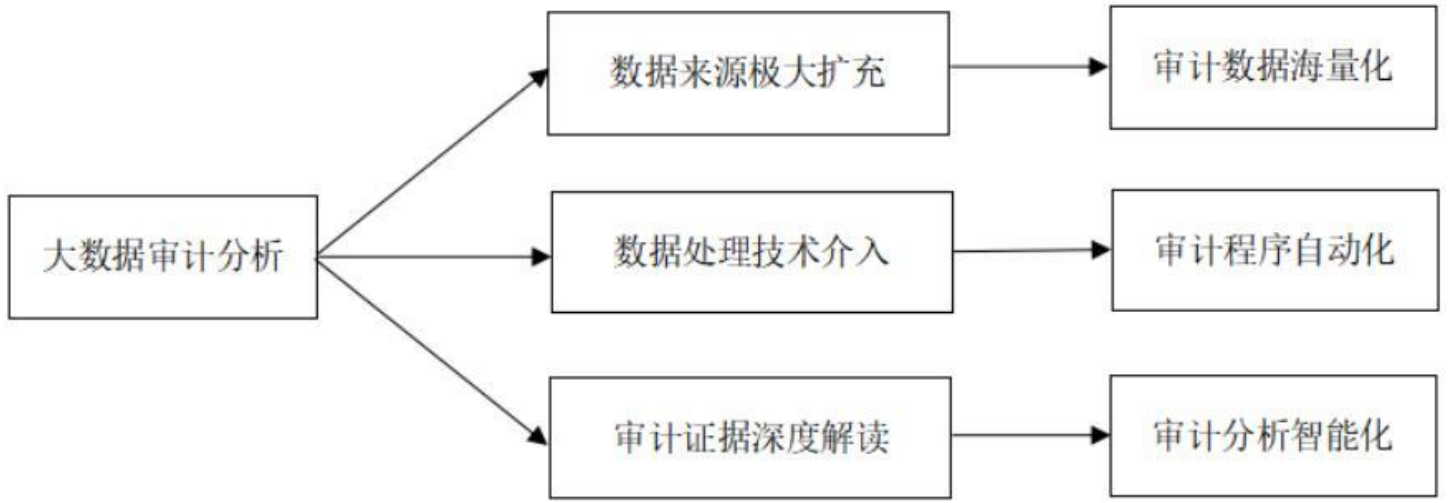


图 1-2 大数据审计分析的内容

东北财经大学

DONGBEI UNIVERSITY OF FINANCE & ECONOMICS

第二章

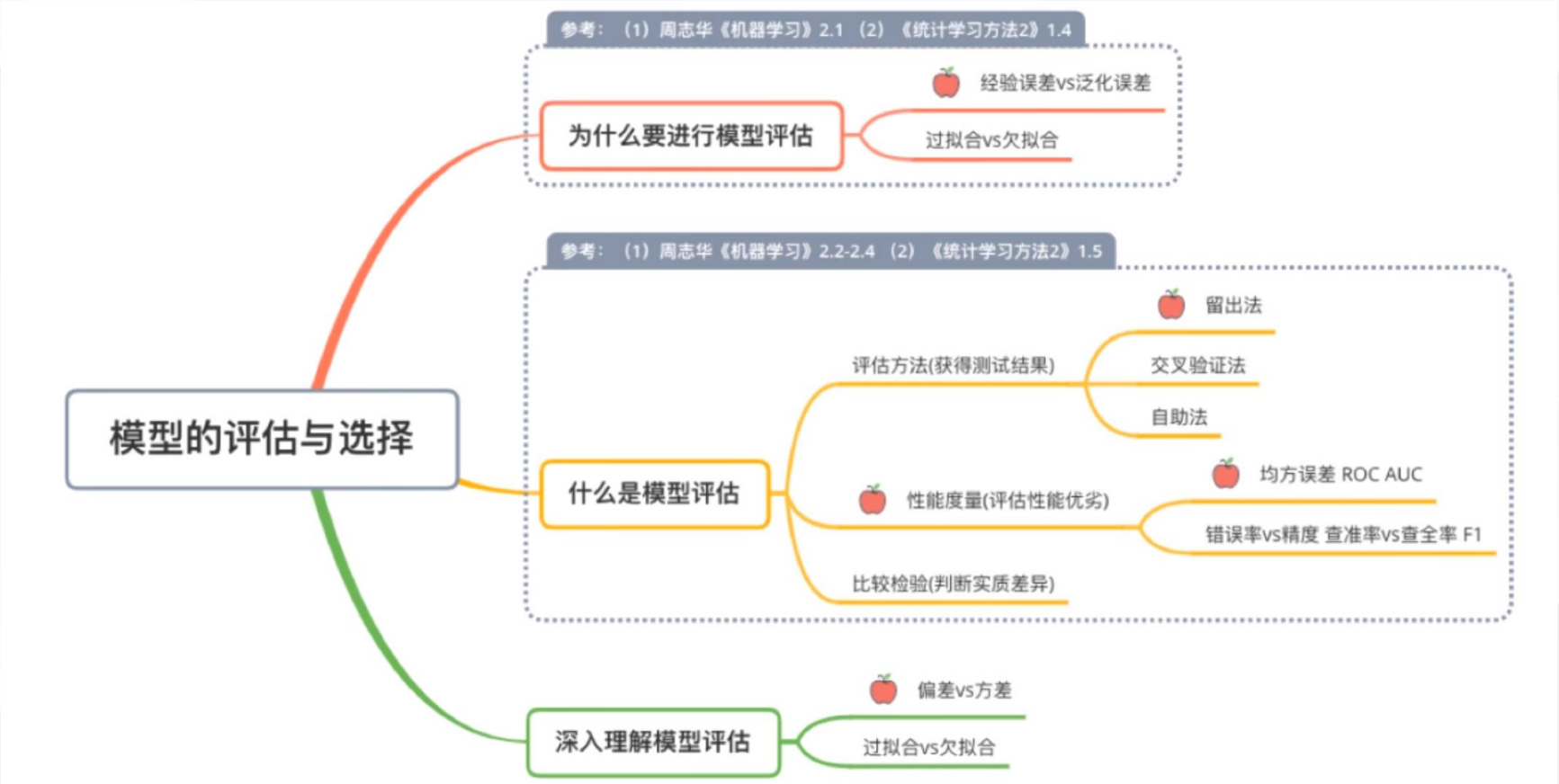
模型评估与选择

管理科学与工程学院

谢琛

1. 章节框架

- 概念与难点
- 误差与过拟合
- 评估方法
- 性能度量



1. 重难点

重难点

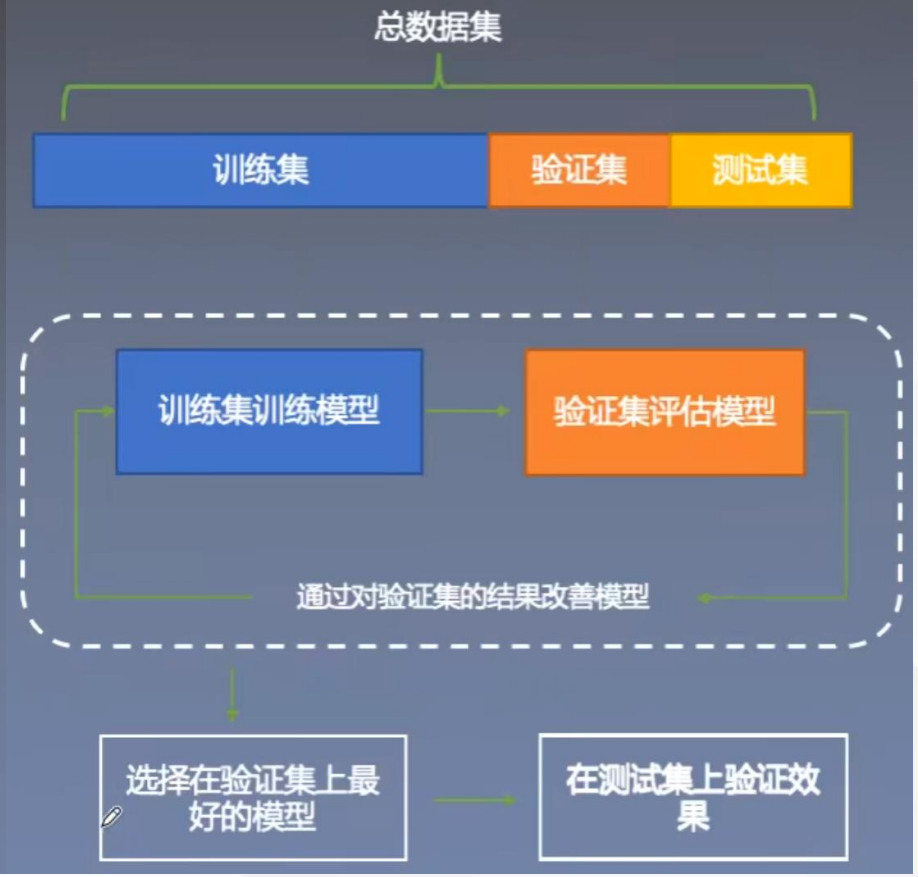
Difficult point

训练数据分层

- 训练集：用来训练模型，模型的迭代优化
- 验证集：调整超参数，优化模型
- 测试集：不参与训练流程，监测模型效果

经验误差 vs 泛化误差

- 经验误差：在训练集上面的误差-对应训练集数据
- 泛化误差：在“未来”样本上的误差-对应测试集数据
- 问题：验证集用来做什么？



2. 错误率 vs 精度 vs 误差

1. 错误率（error rate）：

- 分类错误的样本占样本总数的比例
- 样本总数：m
- 分类错误样本：a

错误率：$\mathrm{E} = \frac{a}{m}$

2. 精度（accuracy）：

- 分类正确的样本占样本总数的比例
- 样本总数：m
- 分类错误样本：a

精度：$1 - \frac{a}{m}$

3. 误差（error）：

- 实际预测输出与样本真实输出之间的差异
- 预测输出：$\hat{y}$
- 样本真实输出：y

误差：$\mathrm{err} = y - \hat{y}$

4. 训练误差（training error）：

- 在训练集上的误差
- 也称之为经验误差（empirical error）

5. 泛化误差（generalization error）：

- 在新样本上的误差
- 希望：泛化误差越小越好！

2. 误差与过拟合

4. 训练误差（training error）：

- 在训练集上的误差
- 也称之为经验误差（empirical error）

5. 泛化误差（generalization error）：

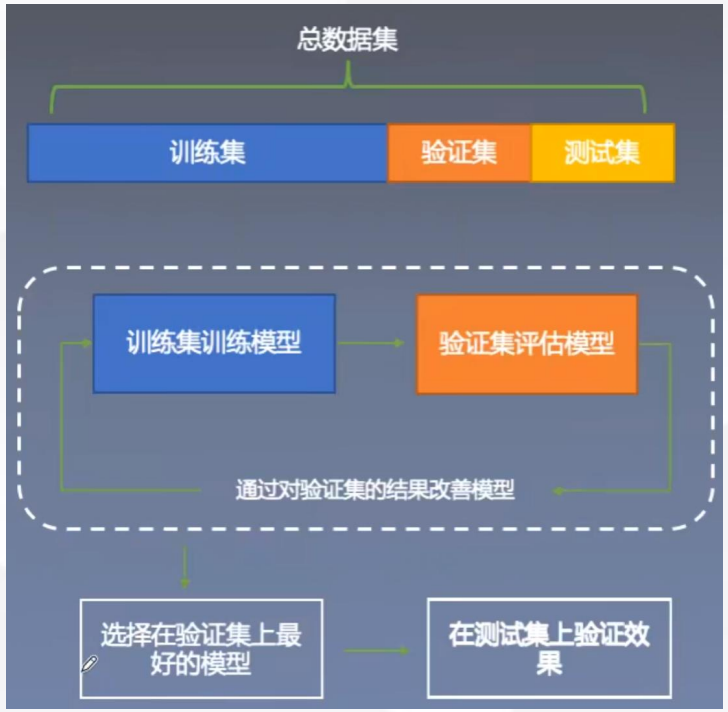
- 在新样本上的误差
- 希望：泛化误差越小越好!

目标：

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”

不希望：

把学习样本训练称精度100%，很可能把样本自身的特点当做所有潜在样本都具有的一般性质



由于事先并不知道新样本的特征，我们只能努力使经验误差最小化；

很多时候虽然能在训练集上做到分类错误率为零，但多数情况下这样的学习器并不好

2. 误差与过拟合

4. 训练误差（training error）：

- 在训练集上的误差
- 也称之为经验误差（empirical error）

5. 泛化误差（generalization error）：

- 在新样本上的误差
- 希望：泛化误差越小越好!

目标：

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”

不希望：

把学习样本训练称精度100%，很可能把样本自身的特点当做所有潜在样本都具有的一般性质

导致泛化能力下降

欠拟合（underfitting）：

- 学的程度不够，欠火候
- 照猫画虎

过拟合（overfitting）：

- 学的过度，学习能力过于强大
- 死记硬背

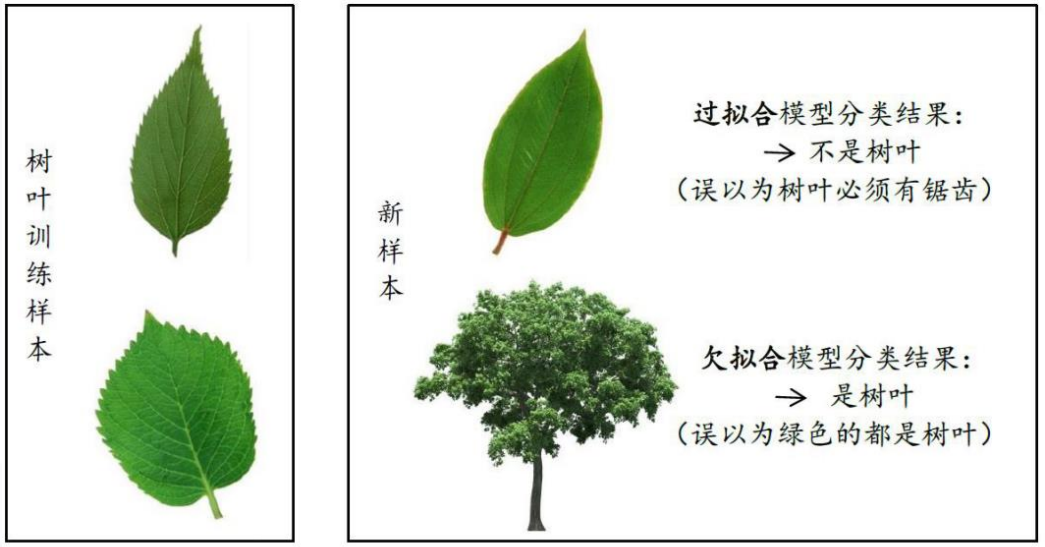
2. 误差与过拟合

欠拟合（underfitting）：

- 学的程度不够，欠火候
- 照猫画虎

过拟合（overfitting）：

- 学的过度，学习能力过于强大
- 死记硬背



过拟合、欠拟合的直观类比

目标：

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”

不希望：

把学习样本训练称精度100%，很可能把样本自身的特点当做所有潜在样本都具有的一般性质

过拟合：学习器把训练样本本身特点当做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.

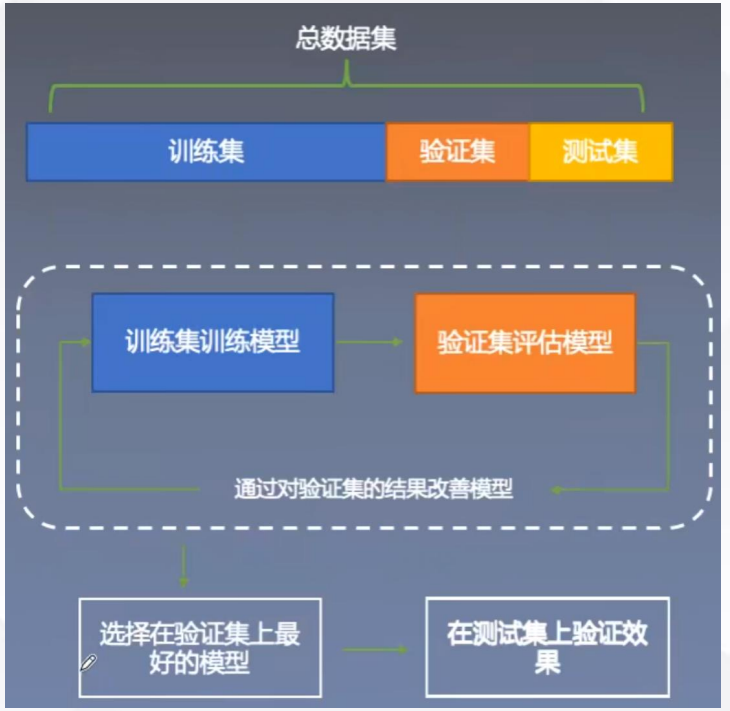
3. 评估方法

测试集（testing set）：

- 通过实验测试对模型的泛华误差进行评估
- 测试模型对新样本的辨别能力
- 从样本真实分布中同分布采样而得
- 测试样本尽量不要在训练样本中出现

希望：

- 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”
- 希望得到泛化能力强的模型 - 举一反三



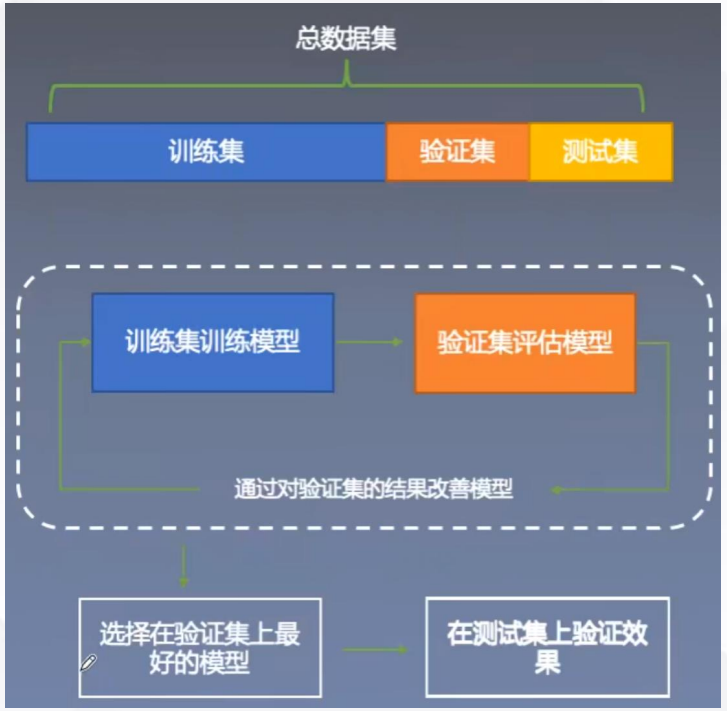
3. 评估方法

通常将包含个m样本的数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
拆分成训练集S和测试集T：

- 概念与难点
- 误差与过拟合
- 评估方法
- 性能度量

1. 留出法（hold-out）：

- 直接将数据集划分为两个互斥集合
- 训练/测试集划分要尽可能保持数据分布的一致性
- 数据切分-> 70/30, 80/20



3. 评估方法

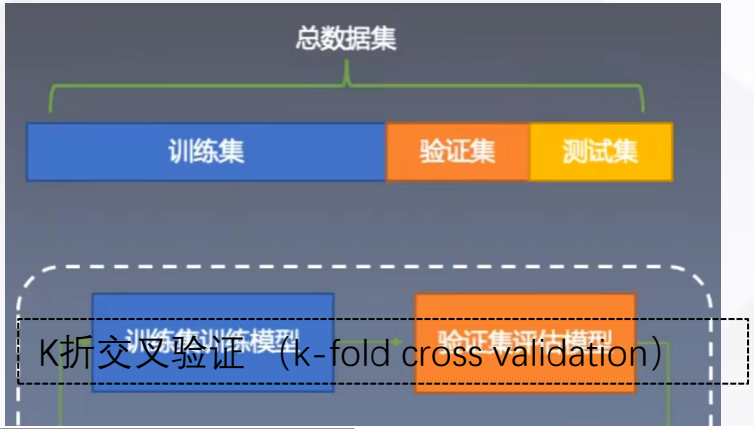
通常将包含个m样本的数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$拆分成训练集$S$和测试集$T$：

2. 交叉验证法（cross validation）：

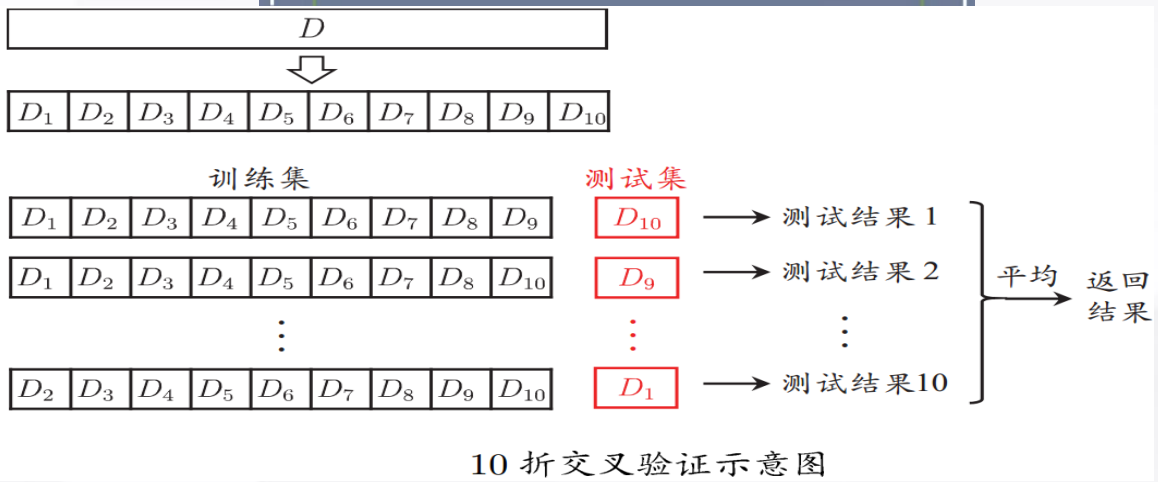
- 将数据集划分为k个大小相似的互斥子集
- 每次用k-1个子集的并集作为训练集
- 余下的子集作为测试集
- 最终返回k个测试结果的均值
- k最常用的取值是10

留一法（Leave-One-Out）：

- 数据集D包含m个样本，若k=m，则得到留一法
- 留一法的评估结果往往被认为比较准确
- 但面对数据集比较大时，训练的计算开销巨大



K折交叉验证（k-fold cross validation）



10 折交叉验证示意图

3. 评估方法

通常将包含个m样本的数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
拆分成训练集S和测试集T：

概念与难点
误差与过拟合
评估方法
性能度量

3. 自助法（bootstrapping）：

- 采样时每次从D中随机挑选一个样本，将它放入D'中，
  然后再将该样本放回D中
- 该样本在下次采样时扔有可能被采到
- 这个过程重复m次，得到包含m个样本的D'
- D中有一部分样本没被采样至D'（约为1/3，36.8%）
- 自主法在数据集较小、难以有效划分训练和测试集时很有用
- 自助法能从初始数据集中产生不同的训练集，对集成学习等方法有很大的好处
- 但自助法改变了原始数据集的分布，会引入偏差，在数据量足够时，留出法和交叉验证更常用

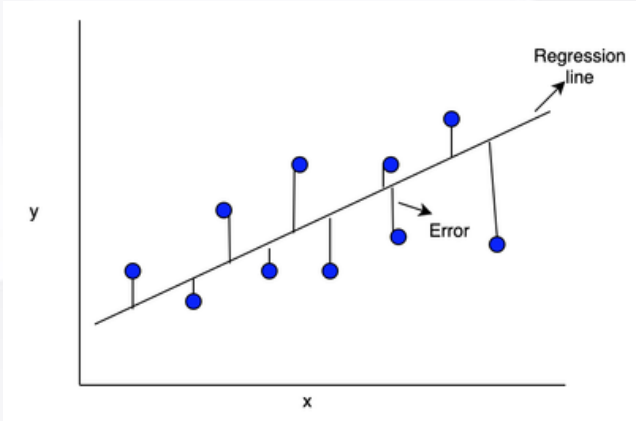
4. 性能度量（Performance Measure）

通常将包含个m样本的数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$

- 概念与难点
- 误差与过拟合
- 评估方法
- 性能度量

1. 均方误差（mean squared error）：

- 回归任务最常用的性能度量
- $\mathrm{MSE} = \frac{1}{\mathrm{m}}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$



2. 混淆矩阵（confusion matrix）：

- 分类任务最常用的性能度量
- TP、FP、TN、FN
- True Positive； False Negative
- False Positive； True Negative

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）

4. 性能度量（Performance Measure）

2. 混淆矩阵（confusion matrix）：

- 分类任务最常用的性能度量
- TP、FP、TN、FN
- True Positive； False Negative
- False Positive； True Negative

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）

1) 错误率（error rate）：

- 分类错误的样本占样本总数的比例
- 样本总数：m
- 分类错误样本：a

错误率：$\mathrm{E} = \frac{a}{m}$

2) 精度（accuracy）：

- 分类正确的样本占样本总数的比例
- 样本总数：m
- 分类错误样本：a

精度：$1 - \frac{a}{m}$

分类正确样本 $= TP + TN$

$m = TP + FN + FP + TN$

$$accuracy = \frac{TP+TN}{TP+FN+FP+TN}$$

4. 性能度量（Performance Measure）

2. 混淆矩阵（confusion matrix）：

- 分类任务最常用的性能度量
- TP、FP、TN、FN
- True Positive； False Negative
- False Positive； True Negative

分类结果混淆矩阵

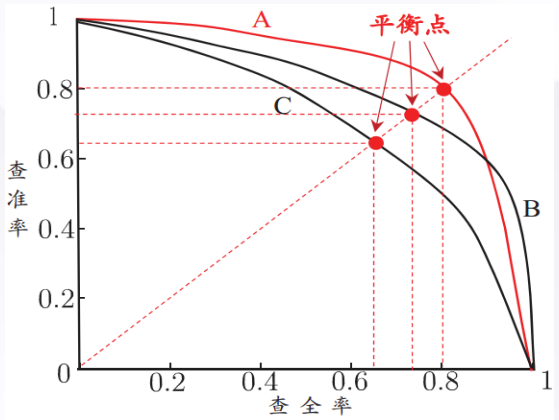
真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

3) 查准率（Precision）：

$$P = \frac{TP}{TP+FP}$$ （只关心预测为positive的结果）

4) 查全率（Recall）：

$$R = \frac{TP}{TP+FN}$$ （只关心实际为positive的结果）



P-R曲线与平衡点示意图

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

4. 性能度量（Performance Measure）

2. 混淆矩阵（confusion matrix）：

- 分类任务最常用的性能度量
- TP、FP、TN、FN
- True Positive； False Negative
- False Positive； True Negative

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）

3) 查准率（Precision）：

$$P = \frac{TP}{TP+FP}$$ （只关心预测为positive的结果）

4) 查全率（Recall）：

$$R = \frac{TP}{TP+FN}$$ （只关心实际为positive的结果）

5) F1 score：

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

4. 性能度量（Performance Measure）

2. 混淆矩阵（confusion matrix）：

- 分类任务最常用的性能度量
- TP、FP、TN、FN
- True Positive； False Negative
- False Positive； True Negative

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）

3) 查准率（Precision）：

$$\mathrm{P}=\frac{\mathrm{TP}}{\mathrm{TP+FP}}$$ （只关心预测为positive的结果）

4) 查全率（Recall）：

$$\mathrm{R}=\frac{\mathrm{TP}}{\mathrm{TP+FN}}$$ （只关心实际为positive的结果）

6) ROC (Receiver Operating Characteristic)：

纵轴（True Positive Rate）：$TPR=\frac{TP}{TP+FN}$

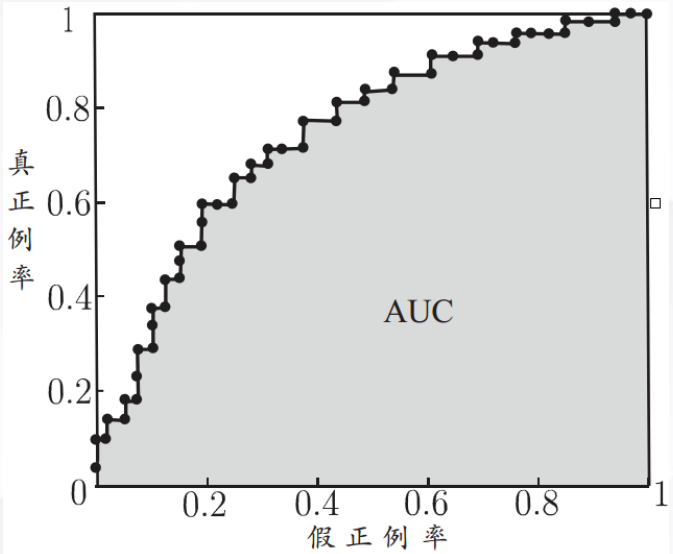
横轴（False Positive Rate）：$\mathrm{FPR}=\frac{FP}{TN+FP}$

4. 性能度量（Performance Measure）

6) ROC (Receiver Operating Characteristic)：

纵轴（True Positive Rate）：$TPR = \frac{TP}{TP+FN}$

横轴（False Positive Rate）：$\mathrm{FPR} = \frac{FP}{TN+FP}$



基于有限样例绘制的 ROC 曲线与 AUC

若某个学习器的ROC曲线被另一个学习器的曲线“包住”，则后者性能优于前者；

否则如果曲线交叉，可以根据ROC曲线下面积大小进行比较，也即AUC值.

分类结果混淆矩阵

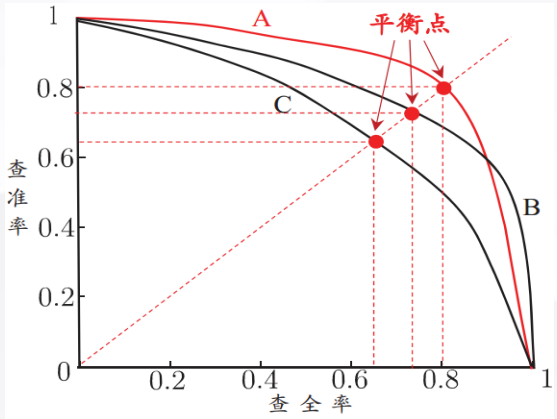
真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

3) 查准率（Precision）：

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$ （只关心预测为positive的结果）

4) 查全率（Recall）：

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$ （只关心实际为positive的结果）



P-R曲线与平衡点示意图