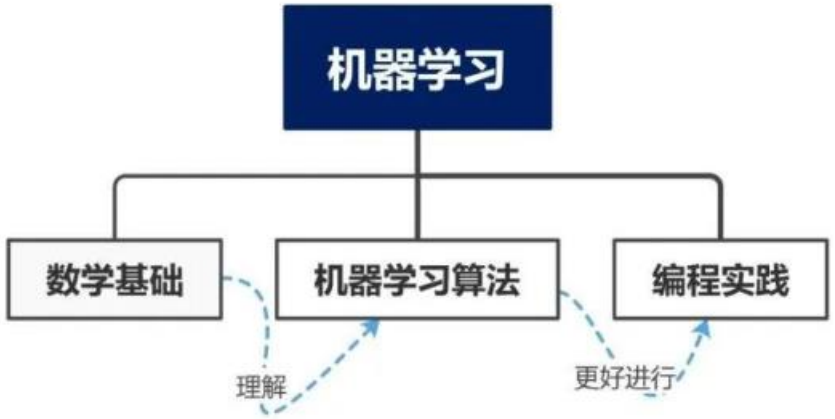
机器学习绪论

目录

基本概念

1. 什么是机器学习？



- **关键字**：机器？学习？
- **学习**：人类学习 vs 机器学习

x_1	y_1
3	4.5
1	2.5
7	8.5
2	3.5
4	？

找规律

$y_1 = x_1 + 1.5$

5.5

x_2	y_2
3	10.5
6	37.5
8	65.5
1	2.5
2	？

找规律

$y_2 = x_2^2 + 1.5$

5.5

x_1	x_2	y
2	7	52.8
6	9	96.7
5	3	21.2
1	2	6.0
4	5	？

找规律

$y = x_1^{3/2} + x_2^{2+} 1$

37.2

- **机器学习**：从数据中自动学出规律；通常依赖大量数据，数据量越大，总结的规律越准确。
- 目的就是从大量的数据中，找到类似 y = a1*x1 + a2*x2 + a3*x3 + … + b这样的规律。

基本概念

2. 监督学习与无监督学习

- **监督学习：**处理的数据是带有标签（label）的

	萼片长度	萼片宽度	花瓣长度	花瓣宽度	分类	
X	5.1	3.5	1.4	0.2	setosa	y
	4.9	3	1.4	0.2	setosa	
	7	3.2	4.7	1.4	versicolor	
	6.4	3.2	4.5	1.5	versicolor	
	6.3	3.3	6	2.5	virginica	
	5.8	2.7	5.1	1.9	virginica	
	6.5	3	5.8	2.2	?	
	6.2	2.9	4.3	1.3	?	

基本概念

2. 监督学习与无监督学习

- **无监督学习**：处理的数据是没有标签的，由于不带标签，核心任务是分析数据本身的结构

X

萼片长度	萼片宽度	花瓣长度	花瓣宽度
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
7	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.3	3.3	6	2.5
5.8	2.7	5.1	1.9
6.5	3	5.8	2.2
6.2	2.9	4.3	1.3

分组？

客户	下单数量	平均单价	单次下单最大数额
A	136	4536	93268
B	50	226	656
C	3	10	32
D	121	5304	70326
E	46	325	513
F	5	16	26

X

分组？

{AD: 金牌客户}
{BE: 银牌客户}
{CF: 铜牌客户}

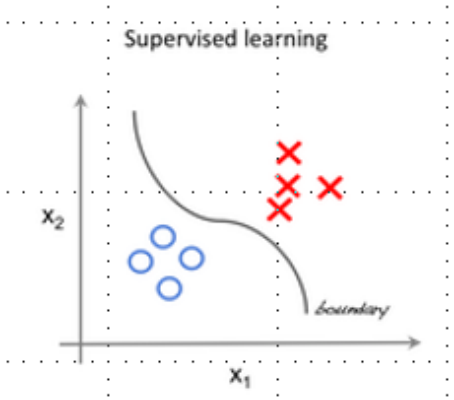
基本概念

2. 监督学习与无监督学习

- **监督学习**：处理的数据是带有标签（label）的
- **无监督学习**：处理的数据是没有标签的，由于不带标签，核心任务是分析数据本身的结构

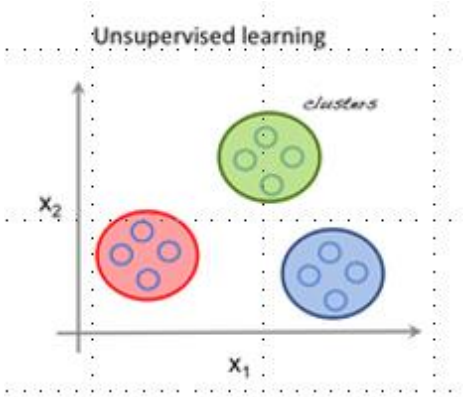
监督学习

学习D=(X, y)中X->y的映射关系



无监督学习

学习D=(X)中X的特征或规律



基本概念

2. 回归和分类问题

- **回归问题：**预测具体的数值如股票价格、身高大小等连续型定量结果

x_1	x_2	y
2	7	52.8
6	9	96.7
5	3	21.2
1	2	6.0
4	5	

$$y = x_1^{3/2} + x_2^{2} + 1$$

基本概念

2. 回归和分类问题

- **分类问题：**预测特定类别如文档主题、信用好坏等离散型定性结果

萼片长度	萼片宽度	花瓣长度	花瓣宽度	分类
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
6.5	3	5.8	2.2	?
6.2	2.9	4.3	1.3	?

$$y = \{\text{setosa, versicolor, virginica}\}$$

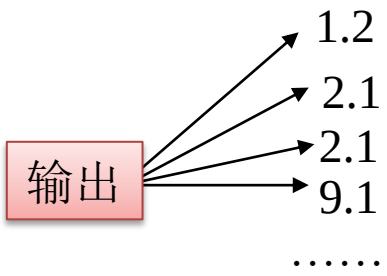
基本概念

2. 回归和分类问题

- **回归问题**：预测具体的数值如股票价格、身高大小等连续型定量结果
- **分类问题**：预测特定类别如文档主题、信用好坏等离散型定性结果

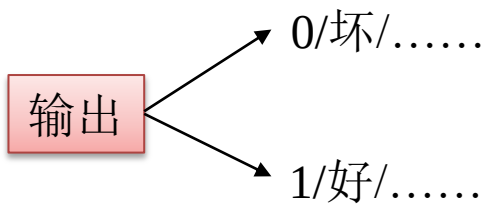
回归问题

输出连续型定量结果



分类问题

输出离散型定性结果



基本概念

2. 回归和分类问题

- **问题：以下不属于回归问题的是？**

  A. 根据历史股票数据预测未来股价

  B. 根据历史用户下单记录预测销售量

  C. 根据历史检测记录预测未来感染人数

  D. 观察图像判断图像中是否有狗

- **问题：以下不属于分类问题的是？**

  A. 根据还款记录预测信用好坏

  B. 根据历史分数线预测今年是否可以考上大学

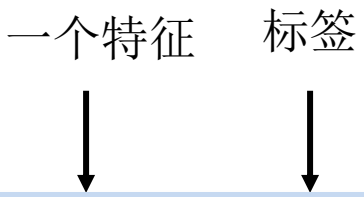
  C. 根据采购记录预测未来14天内应该购买多少库存

  D. 根据采购记录预测明天是否采购某一商品

基本概念

3. 样本、特征与标签

- **样本：**数据集中的一个元素
- **特征：**每个样本具有的属性
- **标签：**需要去预测的变量



萼片长度	萼片宽度	花瓣长度	花瓣宽度	分类
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
6.5	3	5.8	2.2	?
6.2	2.9	4.3	1.3	?

一个样本

基本概念

4. 训练数据和测试数据

- **数据集的描述：**对于三分类问题，数据集表示为$D = \{(x_1, y_1), (x_2,y_2), \ldots, (x_n, y_n)\}$, $x_i \in R^d$, $y_i \in \{0, 1, 2\}$
- **训练数据：**用来训练模型的数据
- **测试数据:** 用来验证模型好坏的数据

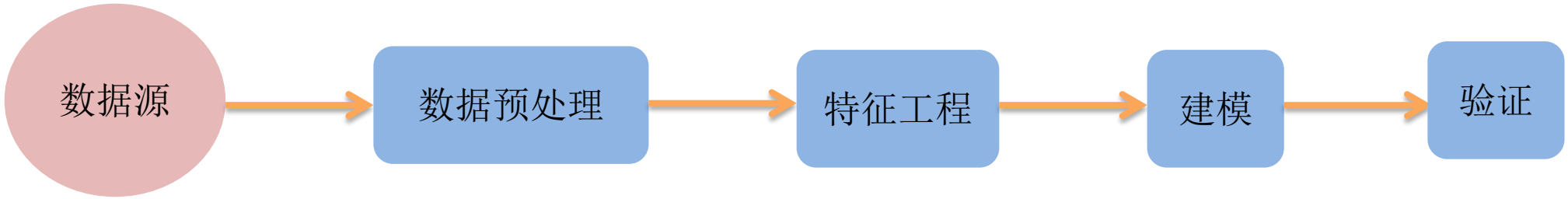
基本概念

4. 机器学习算法

基本概念

5. 机器学习的建模流程

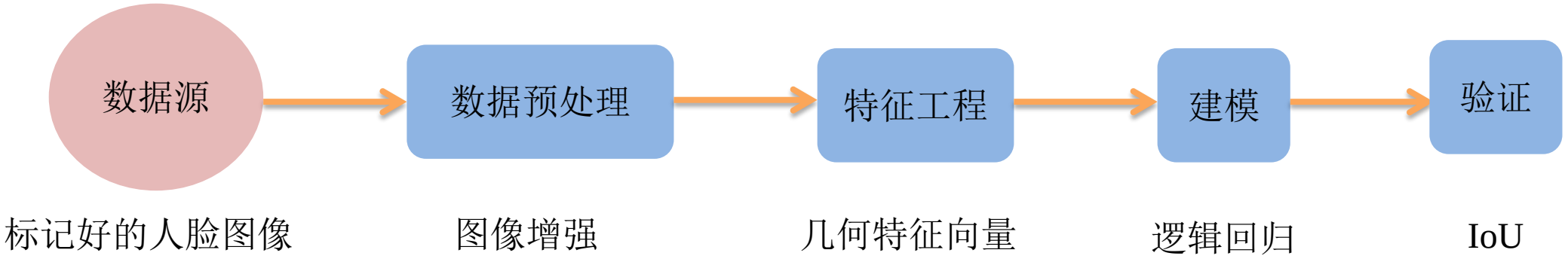
- **数据预处理：**去噪声、处理缺失值、表格清洗、数据增强
- **特征工程：**创建特征向量，特征是指很强的信号
- **建模:** 建立映射关系
- **验证:** 查看模型的效果



基本概念

5. 机器学习的建模流程——人脸识别

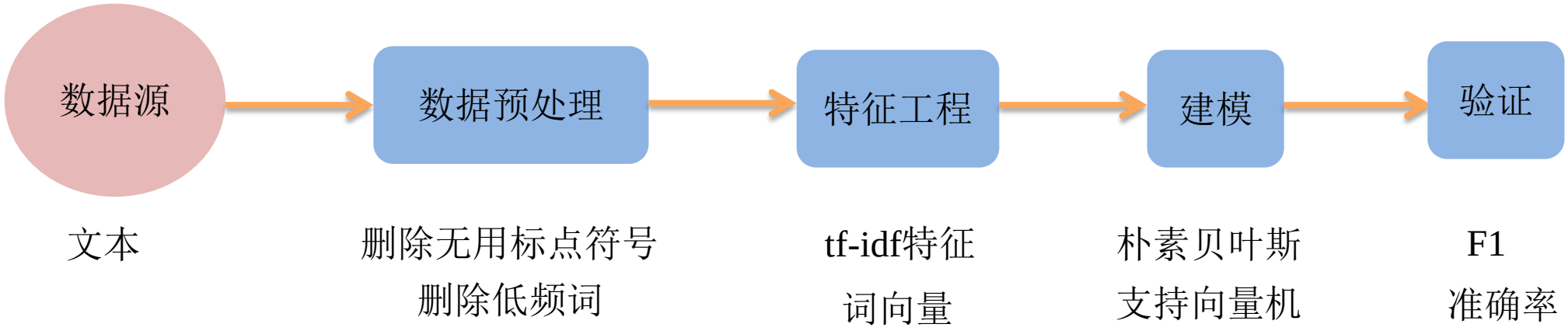
- **数据预处理：**去噪声、处理缺失值、表格清洗、数据增强
- **特征工程：**创建特征向量，特征是指很强的信号
- **建模:** 建立映射关系
- **验证:** 查看模型的效果



基本概念

5. 机器学习的建模流程——情感分析

- **数据预处理：**去噪声、处理缺失值、表格清洗、数据增强
- **特征工程：**创建特征向量，特征是指很强的信号
- **建模:** 建立映射关系
- **验证:** 查看模型的效果



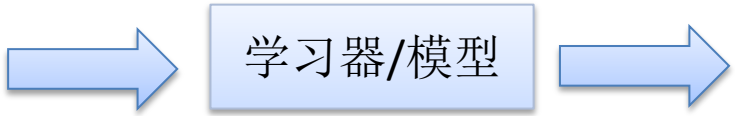
目录

经验误差与过拟合

1. 经验误差

- **误差**：学习器的实际预测输出与样本的真实输出之间的差异
- **经验误差**：在训练集上的误差为训练误差/经验误差
- **泛化误差**：在新样本上的误差为泛化误差

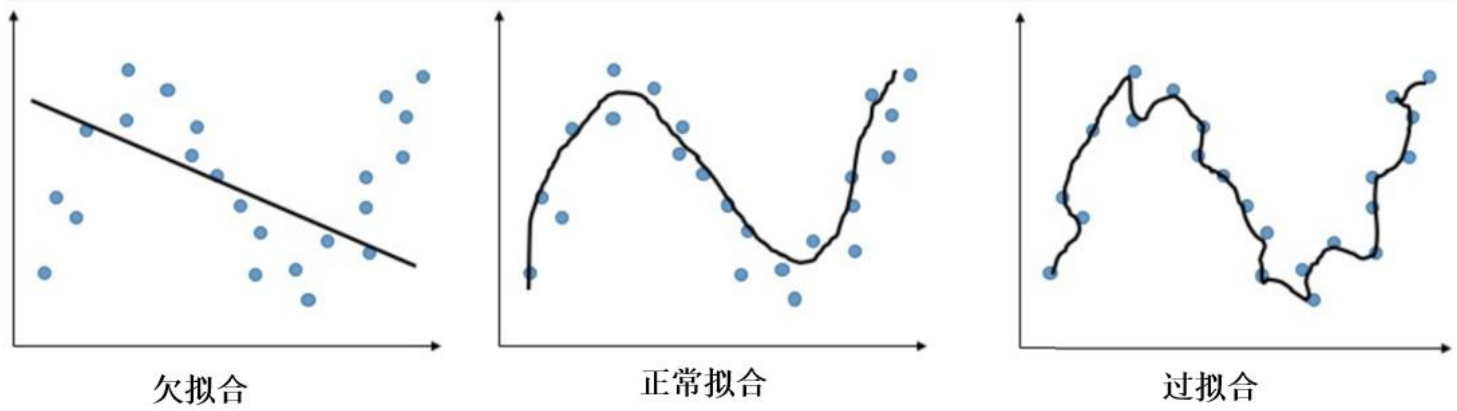
编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍卷	沉闷	否



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	是
4	乌黑	稍卷	沉闷	否

经验误差与过拟合

- **目标：**得到泛化误差小的模型/学习器
- **过拟合：**当学习器把训练样本学的“太好”了的时候，很可能已经把训练样本自身的一些特点当作了所有潜在样本都会具有的一般性质，这样就会导致泛化性能下降，这种现象称为过拟合。具体表现就是最终模型在训练集上效果好；在测试集上效果差。模型泛化能力弱。
- **欠拟合：**欠拟合是指对训练样本的一般性质尚未学好。在训练集及测试集上的表现都不好。



Thank you!