




机器学习入门与实践


朱晓旭 苏州大学计算机科学与技术学院

推荐图书





// 机器学习研究如何通过计算手段，利用经验来改善系统自身的性能。

- 前提：
 - 世界是有规律的
 - 数据是同分布的
 - 黑箱模型：
 - 缺乏坚实的理论基础
 - 人工智能需要一个坚实的理论基础，否则它的发展会有很大困难。
 - 丘成桐 2017 中国计算机大会 《现代几何学与计算机科学》
- 

机器学习三要素

- ▶ 一致性原则

- ▶ 同分布

- ▶ 样本空间划分

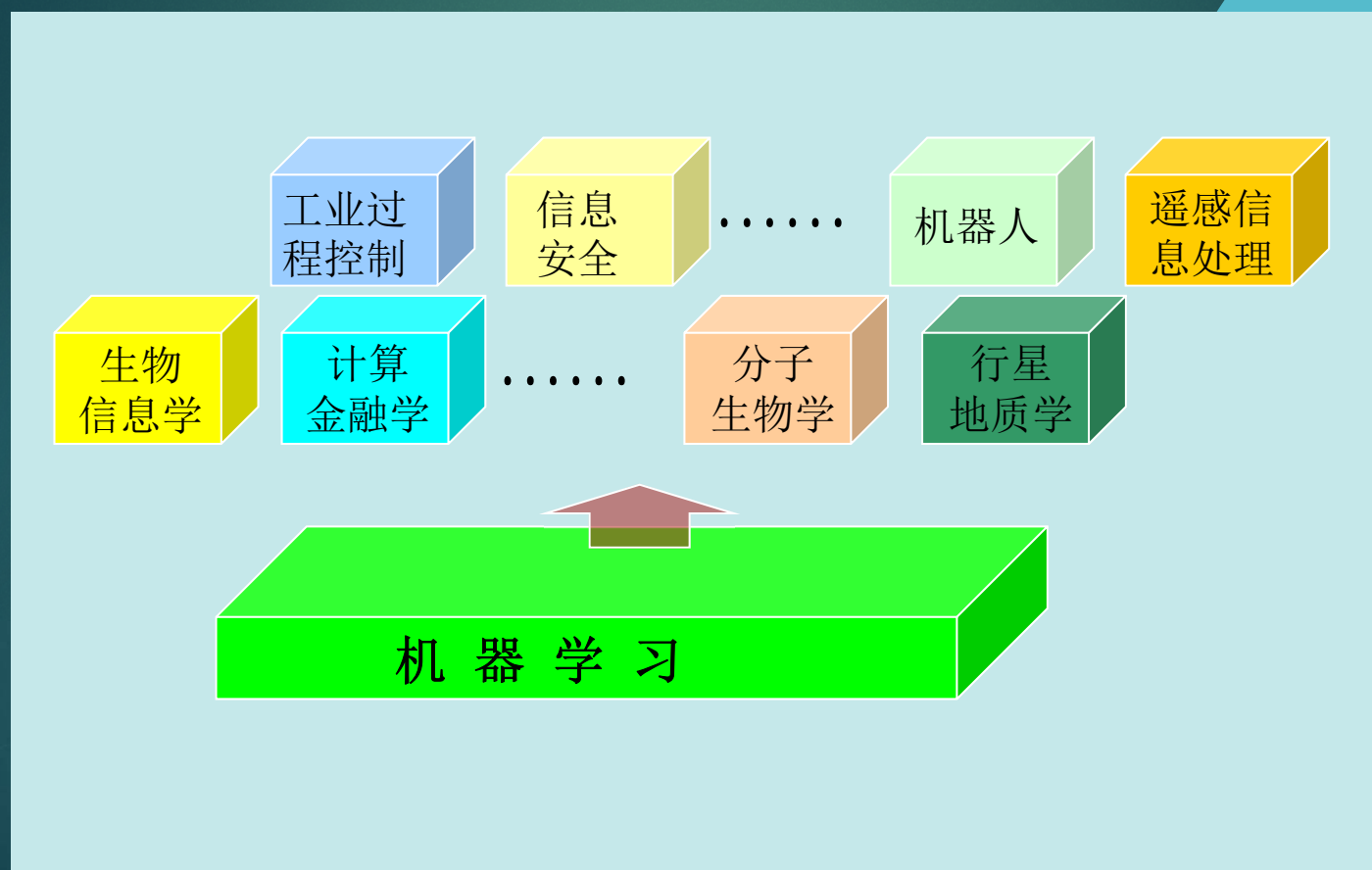
- ▶ 可区分

- ▶ 泛化能力

- ▶ 可扩展



机器学习应用无处不在



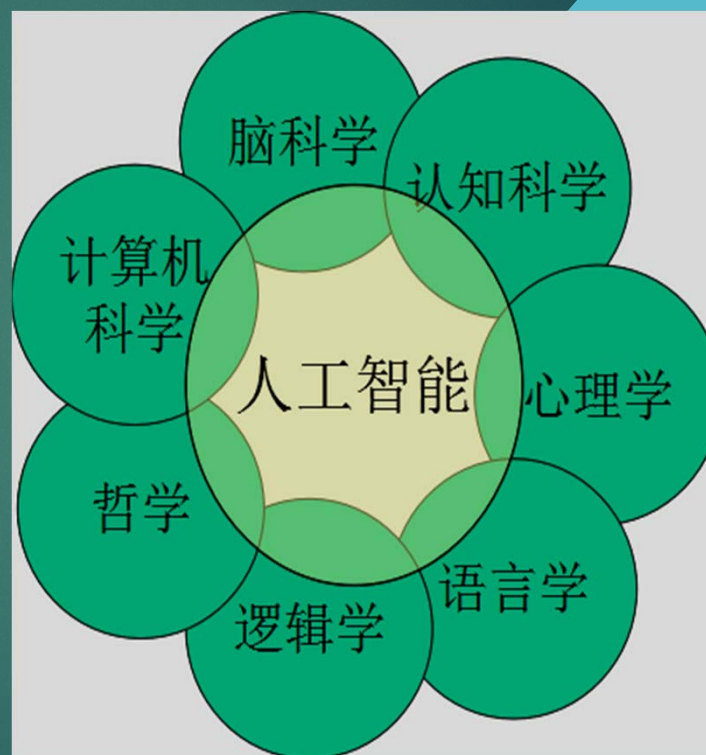
人工智能、机器学习和深度学习



人工智能是目标
机器学习是一种实现人工智能的方法
深度学习是一种实现机器学习的技术

交叉学科

- ▶ 1956
 - ▶ 第一次人工智能研讨会
- ▶ 1966
 - ▶ 盲目乐观
 - ▶ 进入低潮
- ▶ 1970
 - ▶ 专家系统
- ▶ 1986
 - ▶ 神经网络
- ▶ 2006
 - ▶ 基于统计的机器学习



机器学习分类

- ▶ 监督学习：从已知示例中泛化

- ▶ 回归

- ▶ 预测的变量是连续的

- ▶ 分类

- ▶ 预测的变量是离散的

- ▶ 无监督学习

- ▶ 聚类

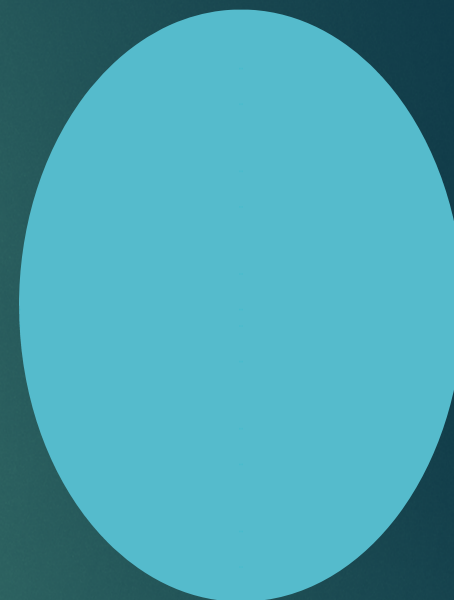
- ▶ 半监督学习

- ▶ 人工标记数据太贵

- ▶ 无监督有时不靠谱

- ▶ 强化学习

- ▶ 试错、奖励最大化



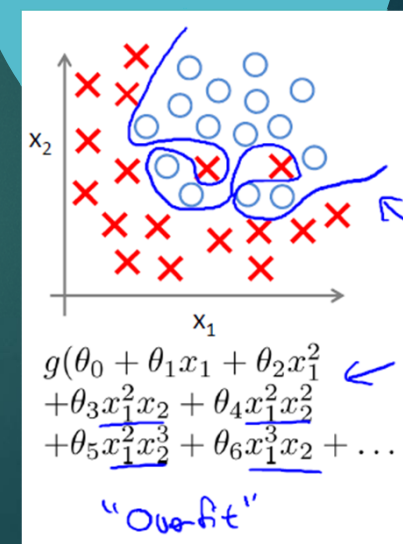
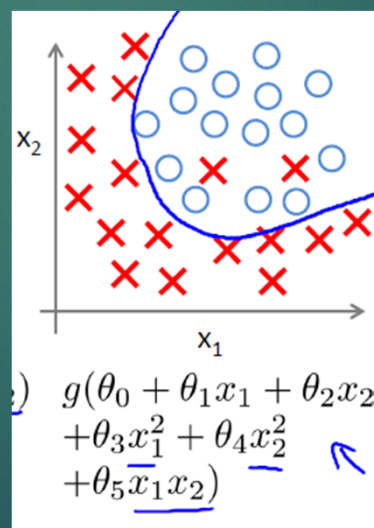
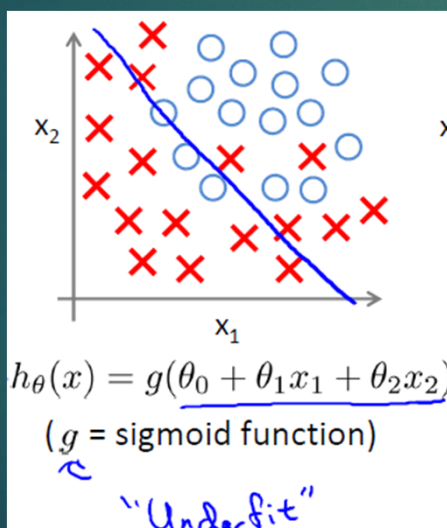
机器学习常见问题

▶ 欠拟合

▶ 模型过于简单，参数不够

▶ 过拟合

▶ 模型太复杂，参数过多，特征数目过多



性能度量

▶ 不同的应用度量指标有所不同、追求指标也不同

▶ 错误率

▶ 仅仅错误率有时不科学

▶ 查准率

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

▶ 查全率

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

▶ F1

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

真实情况	预测结果	
	y=1, 恶性	y=0, 良性
y=1, 恶性	TP (真正例)	FN (假反例)
y=0, 良性	FP (假正例)	TN (真反例)

Confusion Matrix

模型的评估

- ▶ 三个集合

- ▶ 训练集
- ▶ 测试集
- ▶ 开发集

- ▶ 2015年 ImageNet 国际计算机视觉挑战赛

- ▶ 错误率
 - ▶ 百度4.58%，微软是4.94%，谷歌4.8%
- ▶ 每周可以提交两次测试
- ▶ 百度用30个账号进行提交

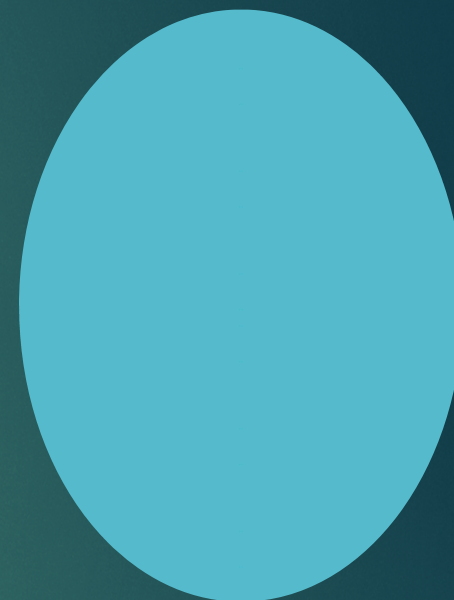
- ▶ 评估方法

- ▶ 留出法
- ▶ 交叉验证法



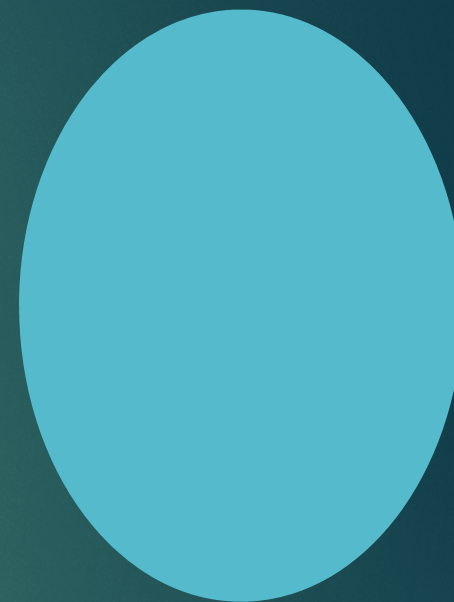
基于机器学习解决问题步骤

1. 数据预处理
2. 特征筛选
3. 选择适用本问题的学习算法
4. 训练模型（基于训练集、和开发集）
5. 测试模型（基于测试集）
6. 开放使用



经典机器学习算法

- ▶ 线性回归
- ▶ Kmeans
- ▶ 朴素贝叶斯
- ▶ 决策树
- ▶ KNN
- ▶ 支持向量机
- ▶ 最大熵
- ▶ 神经网络
 - ▶ 深度学习



分类

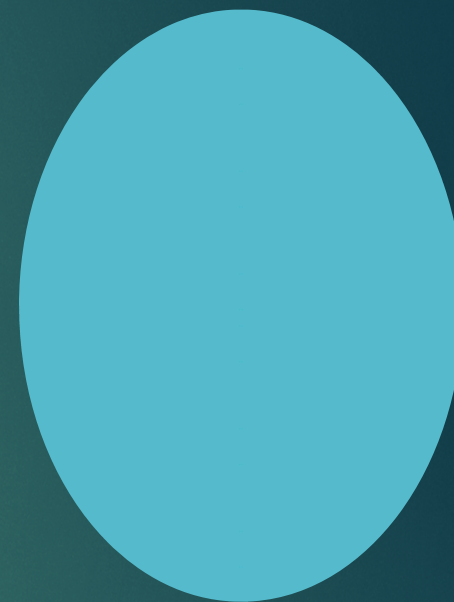
- ▶ 打标签

- ▶ 二分类（正负）

- ▶ 垃圾邮件
- ▶ 非垃圾邮件

- ▶ 多分类

- ▶ 体育新闻
- ▶ 娱乐星闻
- ▶ 社会新闻
- ▶



KNN (K-Nearest-Neighbors)

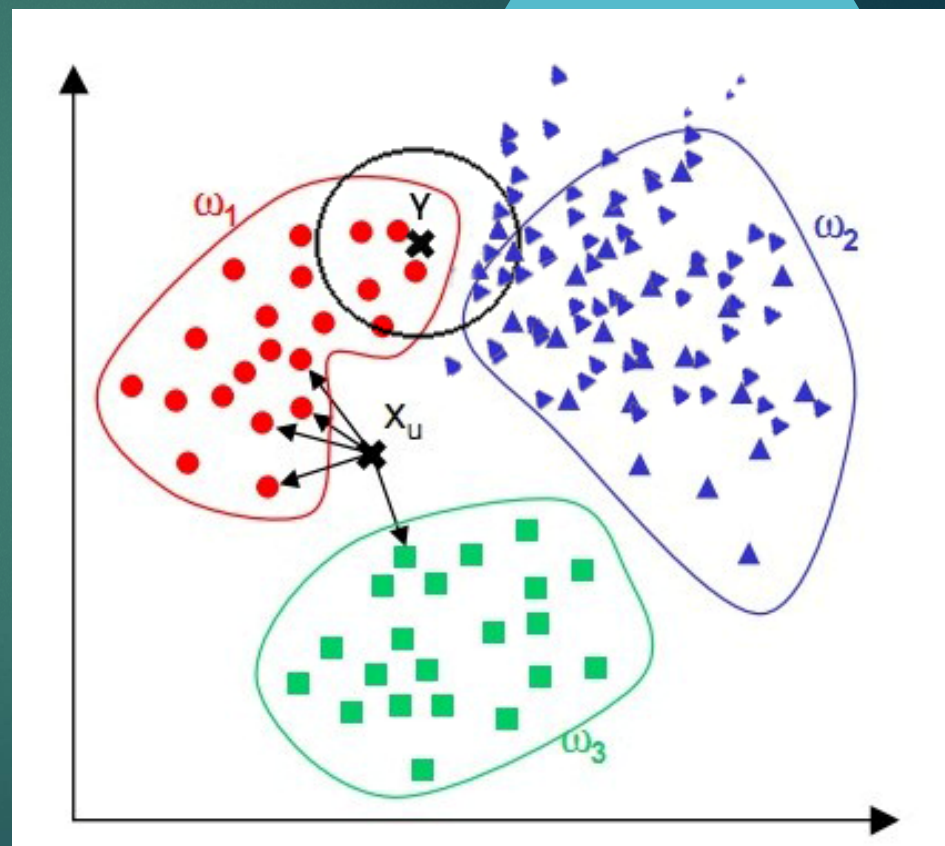
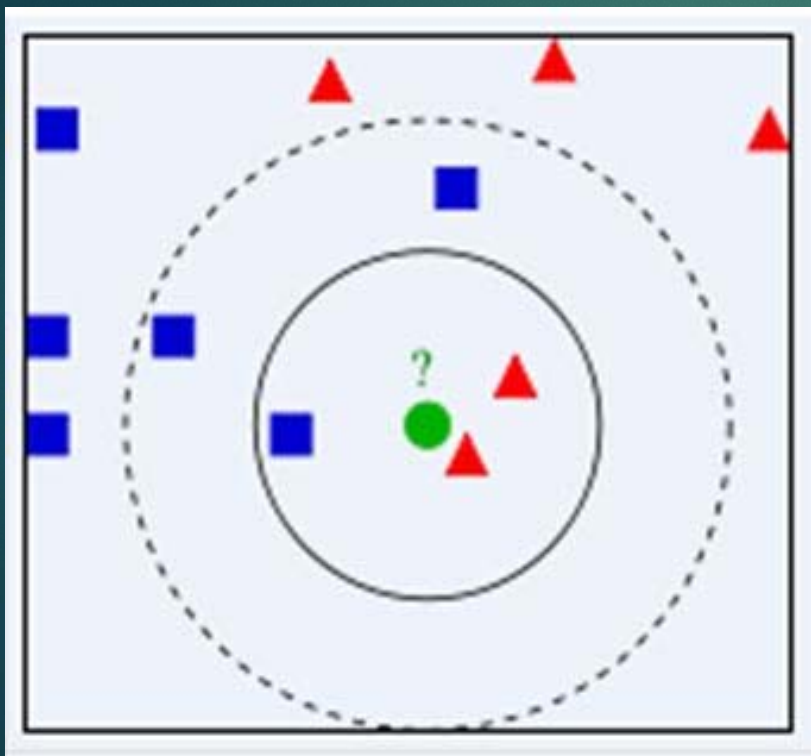
► K近邻分类：

- 为了判定未知样本的类别，以全部训练样本作为代表点，计算未知样本与所有训练样本的距离，并以近邻K个样本的**大多数类别**作为决策未知样本类别的依据



KNN的缺陷

- ▶ K的取值（通常是奇数）
- ▶ 噪声数据敏感



谢谢！

