第七章 机器学习

- 概述
- 分类

概述

●学习是人类获取知识的重要途径和智能的重要标志

●机器学习（Machine learning）：使计算机能模拟人的学习行为，自动地通过学习来获取知识和技能，不断改善性能，实现自我完善。

（1）学习机理

人类获取知识、技能和抽象概念的天赋能力。

（2）学习方法

机器学习方法的构造是在对生物学习机理进行简化的基础上，用计算的方法进行再现。

（3）学习系统

能够在一定程度上实现机器学习的系统。

机器学习发展简史

- 机器学习属于人工智能的范畴。1956年达特矛斯会议是人工智能诞生的标志，提出“学习或者智能的任何其他特性的每一个方面都应能被精确地加以描述，使得机器可以对其进行模拟。”

1.神经元模型研究阶段 (1950年代中到60年代初期)

- 机器学习方法通过监督（有教师指导的）学习来实现神经元间连接权的自适应调整，产生线性的模式分类和联想记忆能力。

2.符号概念获取研究阶段（1960年代初到1970年代）

- 其特点是使用符号而不是数值表示来研究学习问题，其目标是用学习来表达高级知识的符号描述

机器学习发展简史

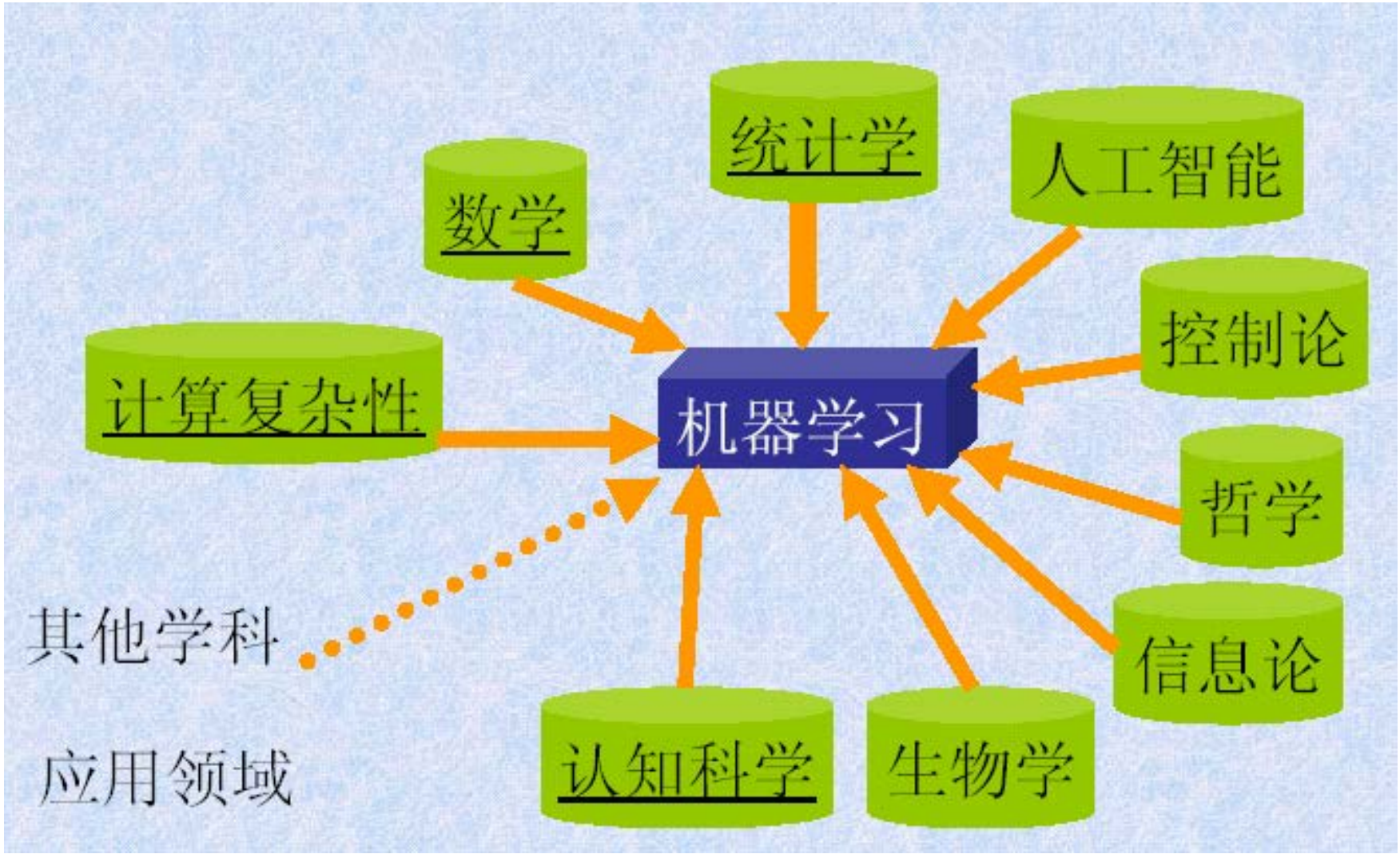
3. 基于知识的学习系统研究阶段（1970年代中期到1980年代中期）

- 有关学习方法: 示例学习、示教学习、 观察和发现学习、类比学习、基于解释的学习。

4. 机器学习的最新阶段始于1986年。

- 是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。
- 机器学习理论主要是设计和分析一些让计算机可以自动“学习”的算法。
- 机器学习已成为新的边缘学科并成为高校课程。

机器学习是一门边缘学科



机器学习的主要研究领域

- 目前最主要的应用领域有：专家系统、认知模拟、规划和问题求解、数据挖掘、网络信息服务、图像识别、故障诊断、自然语言理解、机器人和博弈等领域。
- 大部分的应用研究领域基本上集中于以下两个范畴：分类和问题求解。

机器学习的基本概念

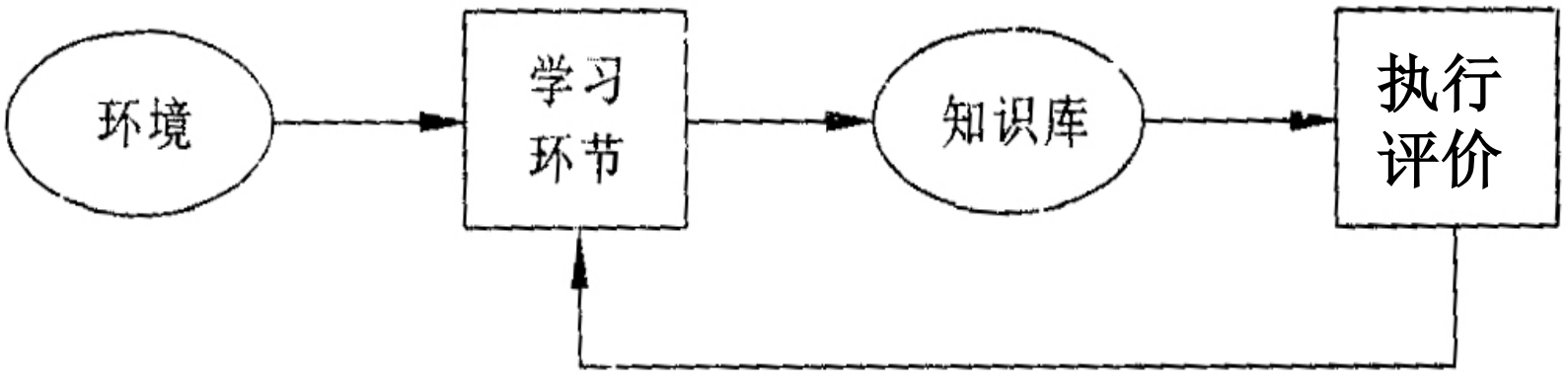
- 学习系统结构：环境、学习、知识库、执行与评价

① 环境向系统的学习部件提供某些信息，

② 学习环节对环境提供的信息进行整理、分析归纳或类比，形成知识，放入/修改知识库。

③ 知识库存储经过加工后的信息（即知识）。

④ 执行环节根据知识库完成任务，同时把获得的信息反馈给评价环节，对知识进行评价，进一步改善执行环节的行为。



箭头	• 表示信息的流向 • 根据反馈信息决定是否要从环境中索取进一步的信息进行学习,以修改、完善知识库中的知识
环境	• 外部信息的来源 • 为系统的学习提供有关信息
学习	• 系统的学习机构 • 对信息进行分析、综合、类比、归纳，获得知识
知识库	• 存储由学习得到的知识 • 存储时进行适当的组织,既便于应用又便于维护
执行	• 处理系统面临的现实问题 • 应用学习到的知识求解问题
评价	• 验证、评价执行环节的效果

机器学习的分类（根据反馈的不同）

■监督学习：主要特点是要在训练模型时提供给学习系统训练样本以及样本对应的类别标签，因此又称为有导师学习。

✓典型的监督学习方法：决策树、支持向量机（SVM）、监督式神经网络等分类算法和线性回归等回归算法。

■无监督学习：主要特点是训练时只提供给学习系统训练样本，而没有样本对应的类别标签信息。

✓典型的无监督学习方法：聚类学习、自组织神经网络学习。

■强化学习：主要特点是通过试错来发现最优行为策略而不是带有标签的样本学习。

机器学习的主要应用

- ✓图像处理\识别（人脸识别、图片分类）
- ✓自然语言处理
- ✓网络安全（垃圾邮件检测、恶意程序\流量检测）
- ✓自动驾驶
- ✓机器人
- ✓医疗拟合预测
- ✓神经网络
- ✓金融高频交易
- ✓互联网数据挖掘/关联推荐

。 。 。 。 。 。

分类问题

- 物以类聚
  - ✓人类认识自然界的一个重要途径
  - ✓根据经验知识，将具有共同特征的事物归纳为相同的类别
  - ✓一个类别 → 一个概念




→ “树”

分类方法之一: 手工方法

- Web发展的初期，Yahoo使用人工分类方法来组织Yahoo目录，类似工作还有：开放式分类目录搜索系统(ODP)、PubMed等
  - 优点：
    - 如果是专家来分类精度会非常高
    - 如果问题规模和分类团队规模都很小的时候，能够保持分类结果的一致性
  - 缺点：
    - 代价昂贵
    - 难以进行规模扩展
- 因此，需要自动分类方法

分类方法之二: (人工撰写)规则的方法

- Google Alerts的例子是基于规则分类的
- 存在一些IDE开发环境来高效撰写非常复杂的规则 (如 Verity)
- 通常情况下都是布尔表达式组合 (如Google Alerts)
  - 优点：
    - 如果规则经过专家长时间的精心调优，精度会非常高
    - 可解释性好
  - 缺点：
    - 建立和维护基于规则的分类系统非常繁琐
    - 开销大

分类方法之三: 统计/概率方法

- 分类被定义为一个有监督的学习问题，包括：

(i) 训练(training)：通过有监督的学习，得到分类函数，然后将其应用实际的分类任务

(ii) 测试/应用/分类(test)：应用于对新样本的分类

- 优点：
  - 速度快，扩展性强，效果好
  - 不需要专家
- 缺点：
  - 需要手工构建训练集(但是普通人即可)
  - 有些方法解释性差

常用的分类方法

- 朴素贝叶斯分类器（Naïve Bayesian）
- 中心向量法 （ Rocchio法）
- K近邻法（KNN）
- 支持向量机（SVM）
- 决策树（Decision Tree）
- 回归法

分类(classification)

- 给定一组数据样本
  - ☞每个样本用一组属性(attribute)来描述
  - ☞每个样本具有已知的类别(class)
  - ☞称为训练数据集(training set)
- 找到一个模型(model, 属性集到类别的映射函数)

  class = *f* (attributes)
- 目标：对于未知类别的样本(其属性new_attribute已知)

  ***f* (new_attribute) → 预测的类别**
- 分类是有指导的机器学习

分类流程