周志华 著

MACHINE LEARNING

机器学习

清华大学出版社

本章课件致谢：
胡鹏

第0-1章 课程信息&绪论

0、课程信息

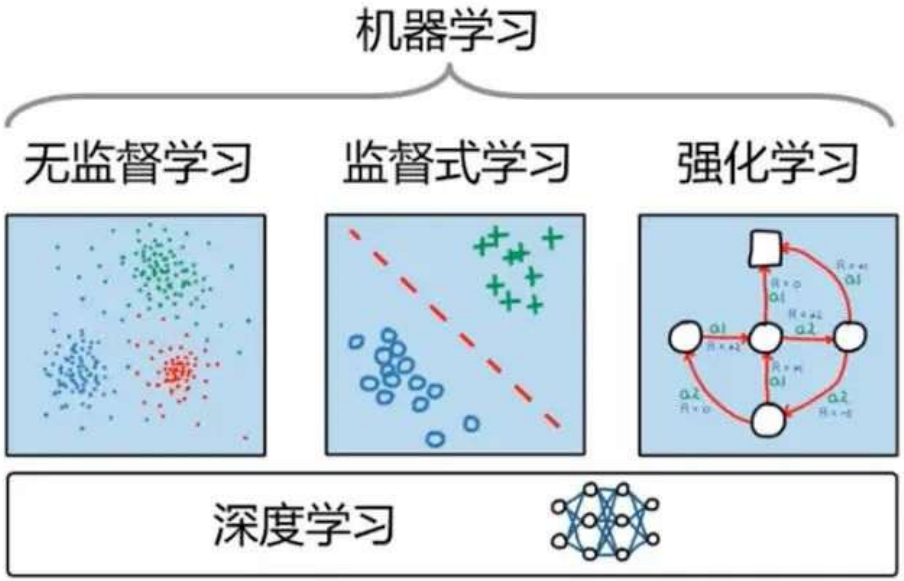
1） 机器学习是学习什么的？

2） 为什么学习机器学习？

3） 怎样学好机器学习?

0、课程信息

1） 机器学习是学习什么的？



涉及到算法理论、数学模型、数据结构、学习范式以及计算复杂性等多层面的深度交互

0、课程信息

1） 机器学习是学习什么的？

机器学习的发展历史显示，其根本目标是通过数据和经验实现自动化的性能改进。传统的机器学习算法强调特征工程和统计建模，其方法论深受统计学、优化理论和计算理论的支撑。随着大数据技术和高性能计算的发展，深度学习逐渐成为研究热点，通过多层神经网络结构实现了从原始数据到高级抽象特征的自动化学习。另一方面，强化学习强调智能体与环境的交互，通过延迟奖励信号指导策略改进，其数学基础依赖于马尔可夫决策过程、动态规划和最优控制理论。

0、课程信息

1） 机器学习是学习什么的？

机器学习的发展历史显示，其根本目标是通过数据和经验实现自动化的性能改进。传统的机器学习算法强调特征工程和统计建模，其方法论深受统计学、优化理论和计算理论的支撑。随着大数据技术和高性能计算的发展，深度学习逐渐成为研究热点，通过多层神经网络结构实现了从原始数据到高级抽象特征的自动化学习。另一方面，强化学习强调智能体与环境的交互，通过延迟奖励信号指导策略改进，其数学基础依赖于马尔可夫决策过程、动态规划和最优控制理论。

0、课程信息

1）为什么学习机器学习？

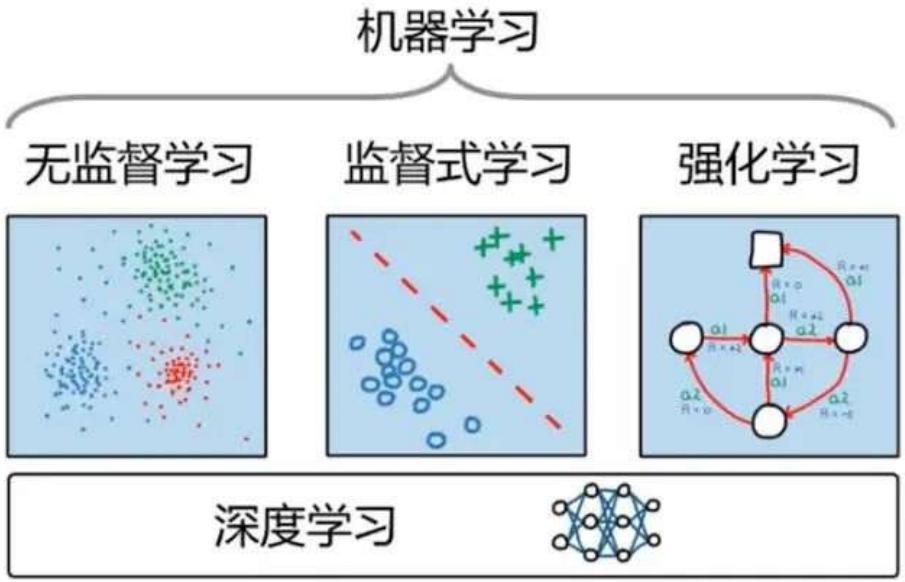
0、课程信息

3） 怎样学好机器学习?

考核方式：考试

机器学习构成

---数据标注情况和任务目标



监督学习（Supervised Learning）

- **数据集由输入-输出对组成**

算法通过最小化预测输出与真实输出之间的误差函数来训练模型。常见算法包括线性回归、支持向量机、决策树和神经网络。

- **核心问题**

模型泛化能力：即训练出的模型在未见样本上的表现是否可靠。

- **思考问题：**如果训练数据覆盖范围有限，监督学习模型是否能够应对未知输入的分布偏移？这种局限性在实际任务中如何被缓解？

无监督学习（Unsupervised Learning

- **数据没有明确标签**

其目标是从数据中发现潜在结构或模式。例如，聚类分析用于发现数据的自然分组，降维方法（如PCA）用于提取数据的主要特征方向。。

- **核心问题**

无监督学习强调数据内在结构的探索，而非直接预测。

- **思考问题：**在高维数据空间中，无监督方法是否总能捕获真实数据分布的本质？模型复杂度和数据量的限制会带来哪些理论约束？

强化学习（Reinforcement Learning, RL）

- **一种基于交互的学习方式**

智能体通过环境反馈的奖励信号调整策略以最大化长期回报。

- **核心问题**

在本质上与监督学习不同，因为它面对的是延迟反馈而非即时标签。其数学基础通常依赖马尔可夫决策过程（MDP），包括状态空间、动作空间、状态转移概率、奖励函数和折扣因子。

- **思考问题：**如何在有限试验中有效探索环境，并在探索与利用之间取得平衡？这涉及算法的收敛性、样本效率以及策略优化稳定性？

深度学习（Deep Learning, DL）

- **机器学习方法中的一个重要分支**

其核心思想是通过多层神经网络实现对数据的层次化表示学习，直接从原始数据中自动提取高层次特征。

- **核心**

核心贡献在于提供多层次、非线性的特征表示。传统机器学习算法（如线性回归、支持向量机）在处理高维、非线性问题时，依赖人工特征设计，表现有限。而深度网络通过端到端训练实现自动特征抽象，显著扩展了机器学习方法的适用范围。

- **思考问题：**深度网络的自动表示能力是否完全取代人工特征设计？在小样本或噪声数据情况下，深度学习是否仍然高效可靠？

大纲

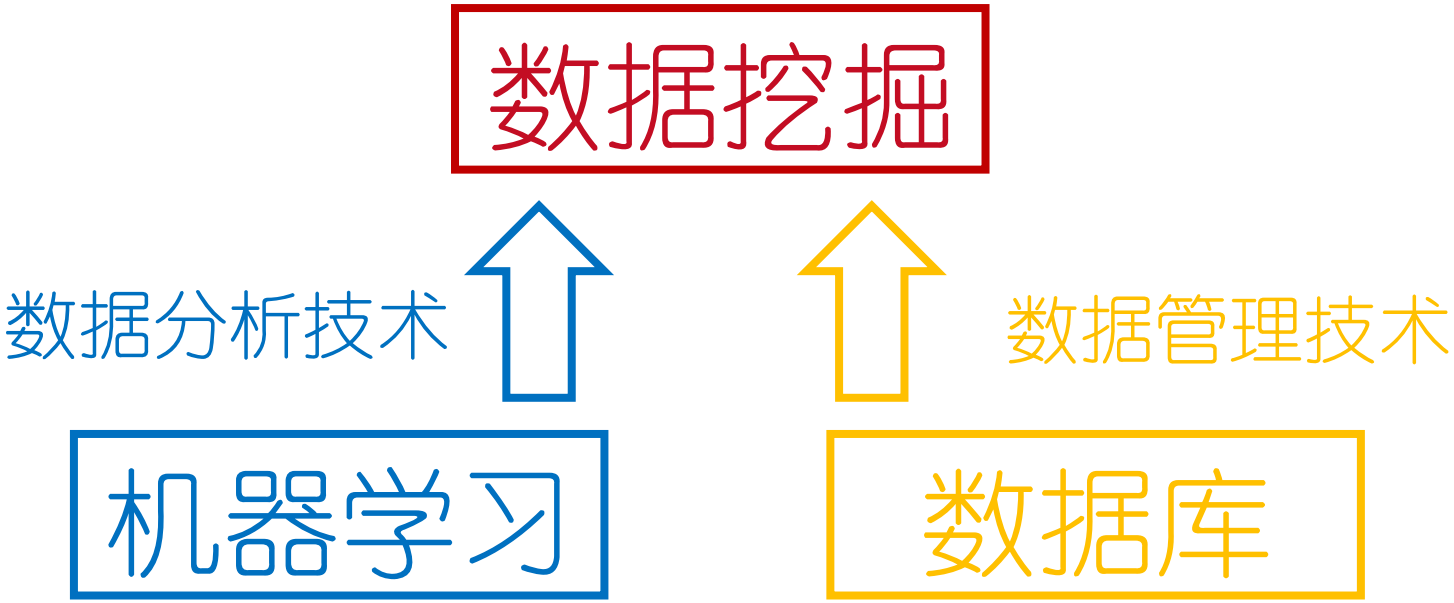
- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

机器学习

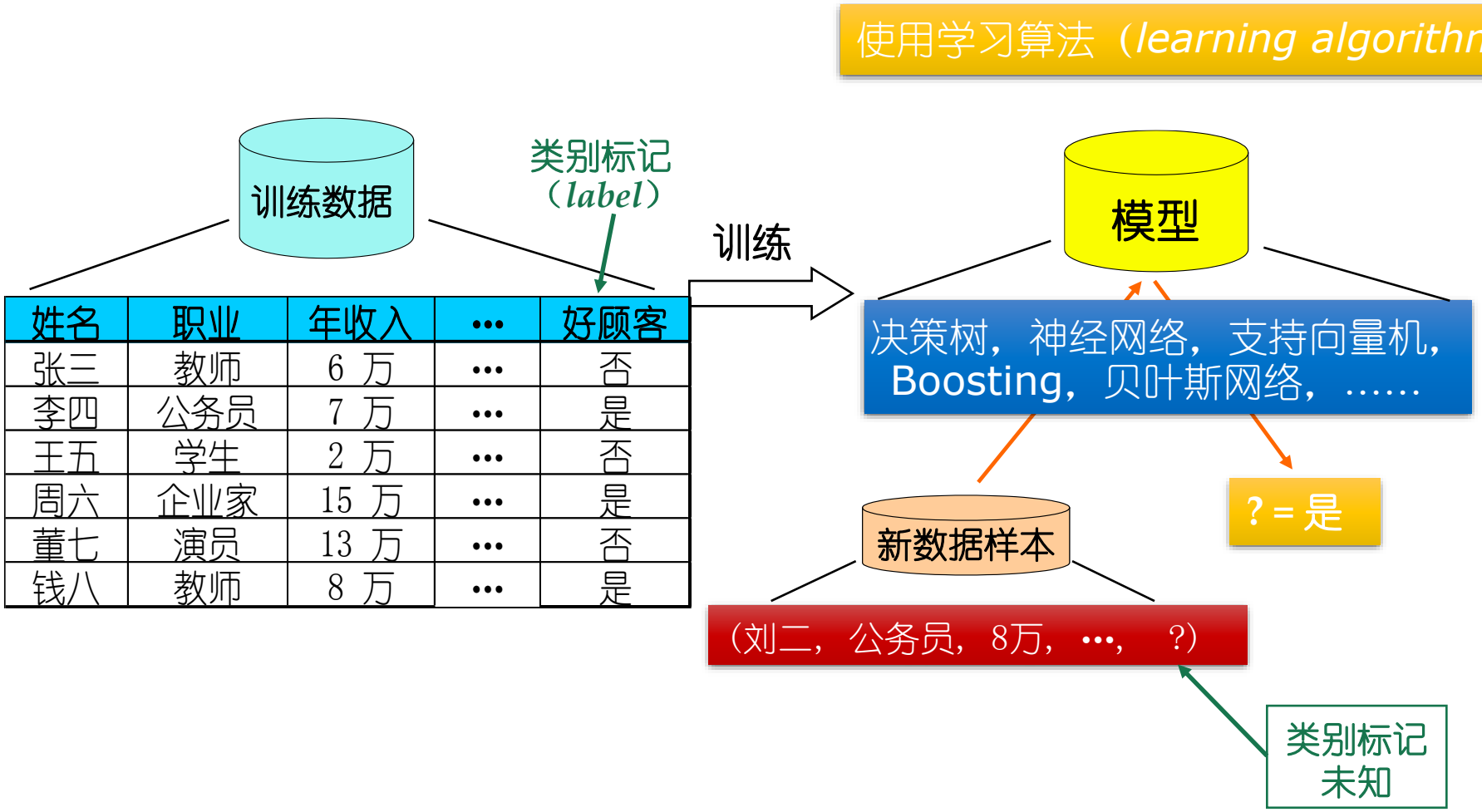
“假设用P来评估计算机程序在某任务类T上的性能，若一个程序通过利用经验E在T中任务上获得了性能改善，则我们就说关于T和P，该程序对E进行了学习”

机器学习致力于研究如何通过计算的手段，利用经验来改善系统自身的性能，从而在计算机上从数据中产生“模型”，用于对新的情况给出判断。

机器学习与数据挖掘



典型的机器学习过程


使用学习算法（learning algorithm）
训练数据
类别标记
（label）
训练
模型
姓名 职业 年收入 … 好顾客
张三 教师 6 万 … 否
李四 公务员 7 万 … 是
王五 学生 2 万 … 否
周六 企业家 15 万 … 是
董七 演员 13 万 … 否
钱八 教师 8 万 … 是
决策树，神经网络，支持向量机，
Boosting，贝叶斯网络，……
新数据样本
？= 是
（刘二，公务员，8万，…，？）
类别标记
未知

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

基本术语-数据

特征 ↑　　标记 ↑

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

训练集←

测试集←

基本术语-任务

- 预测目标：
  - 分类：离散值
    - 二分类：好瓜；坏瓜
    - 多分类：冬瓜；南瓜；西瓜
  - 回归：连续值

    瓜的成熟度
  - 聚类：无标记信息

基本术语-任务

- 有无标记信息
  - 监督学习：分类、回归
  - 无监督学习：聚类
  - 半监督学习：两者结合

基本术语-泛化能力

机器学习的目标是使得学到的模型能很好的适用于“新样本”，而不仅仅是训练集合，我们称模型适用于新样本的能力为泛化(generalization)能力。

通常假设样本空间中的样本服从一个未知分布 $\mathcal{D}$,样本从这个分布中独立获得，即“独立同分布”(i.i.d)。一般而言训练样本越多越有可能通过学习获得强泛化能力的模型

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

假设空间

计算假设空间大小：每个属性的可能取值相乘，再加上空集

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

$(色泽=?) \wedge (根蒂=?) \wedge (敲声=?) \leftrightarrow 好瓜$

在模型空间中搜索不违背训练集的假设

假设空间大小：3*4*4+1=49

假设空间

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

好瓜的假设需要考虑：

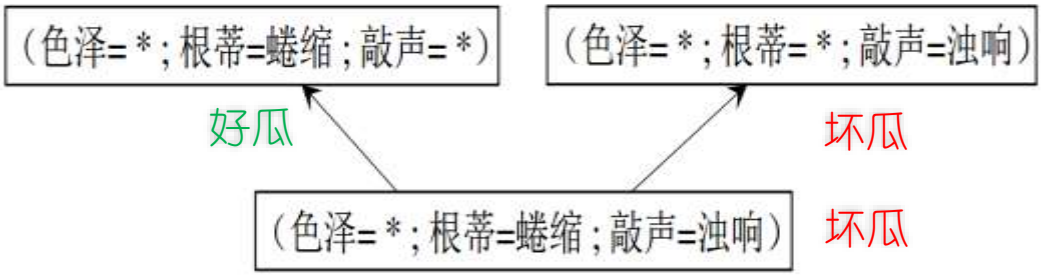
- 每个属性的可能取值；
- 通配符（取任意值都可）；
- 空集（表示没有符合条件的假设）

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

归纳偏好

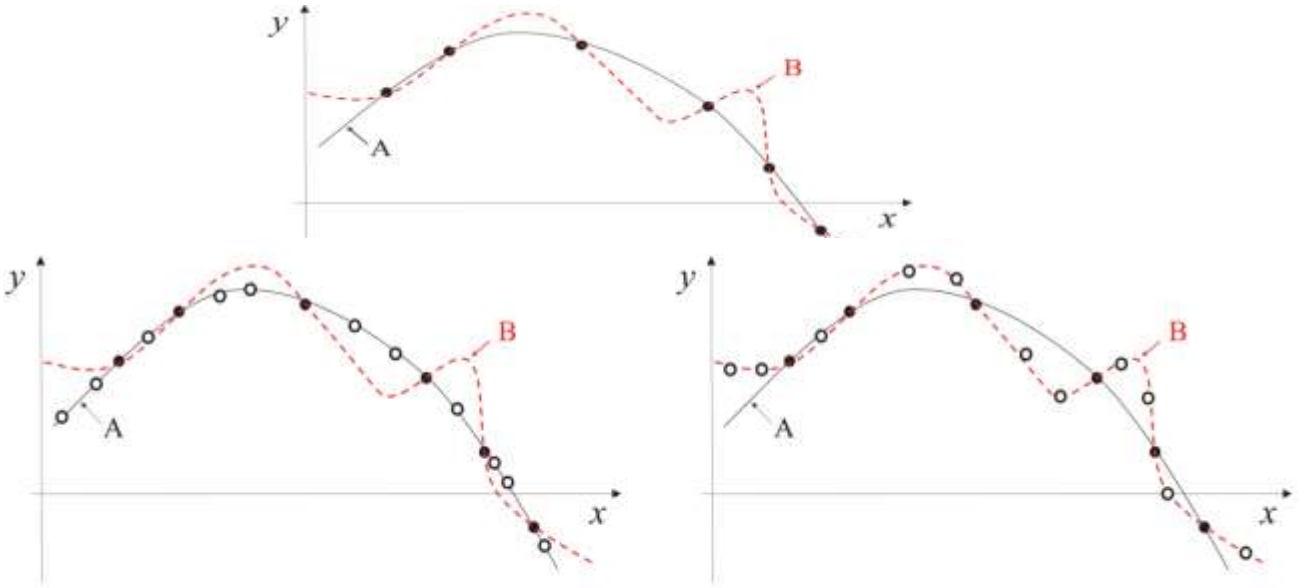
假设空间中有三个与训练集一致的假设，但他们对（色泽=青绿；根蒂=蜷缩；敲声=沉闷）的瓜会预测出不同的结果：



选取哪个假设作为学习模型？

归纳偏好

学习过程中对某种类型假设的偏好称作归纳偏好



(a) A 优于 B　　(b) B 优于 A

没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

归纳偏好

归纳偏好可看作学习算法自身在一个可能很庞大的假设空间中对假设进行选择的启发式或“价值观”.

“奥卡姆剃刀”是一种常用的、自然科学研究中最基本的原则，即“若有多个假设与观察一致，选最简单的那个”.

具体的现实问题中，学习算法本身所做的假设是否成立，也即算法的归纳偏好是否与问题本身匹配，大多数时候直接决定了算法能否取得好的性能.

NoFreeLunch

一个算法ξ_a如果在某些问题上比另一个算法ξ_b好，必然存在另一些问题，ξ_b比ξ_a好，也即没有免费的午餐定理。

简单起见，假设样本空间$\mathcal{X}$和假设空间$\mathcal{H}$离散，令$P(h|X,\mathfrak{L}_a)$代表算法$\mathfrak{L}_a$基于训练数据X产生假设h的概率，在令f代表要学的目标函数，$\mathfrak{L}_a$在训练集之外所有样本上的总误差为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\,\mathbb{I}\left(h\left(\boldsymbol{x}\right)\neq f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)$$

$\mathbb{I}(\cdot)$为指示函数，若$\cdot$为真取值1，否则取值0

NoFreeLunch

考虑二分类问题，目标函数可以为任何函数 $\mathcal{X} \mapsto \{0,1\}$ ，函数空间为 $\{0,1\}^{|\mathcal{X}|}$ ，对所有可能f按均匀分布对误差求和，有：

$$
\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\ \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))\ P(h \mid X, \mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \\
&= 2^{|\mathcal{X}|-1} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \cdot 1 \ .
\end{aligned}
$$

总误差与学习算法无关！

实际问题中，并非所有问题出现的可能性都相同

脱离具体问题，空谈“什么学习算法更好”毫无意义

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

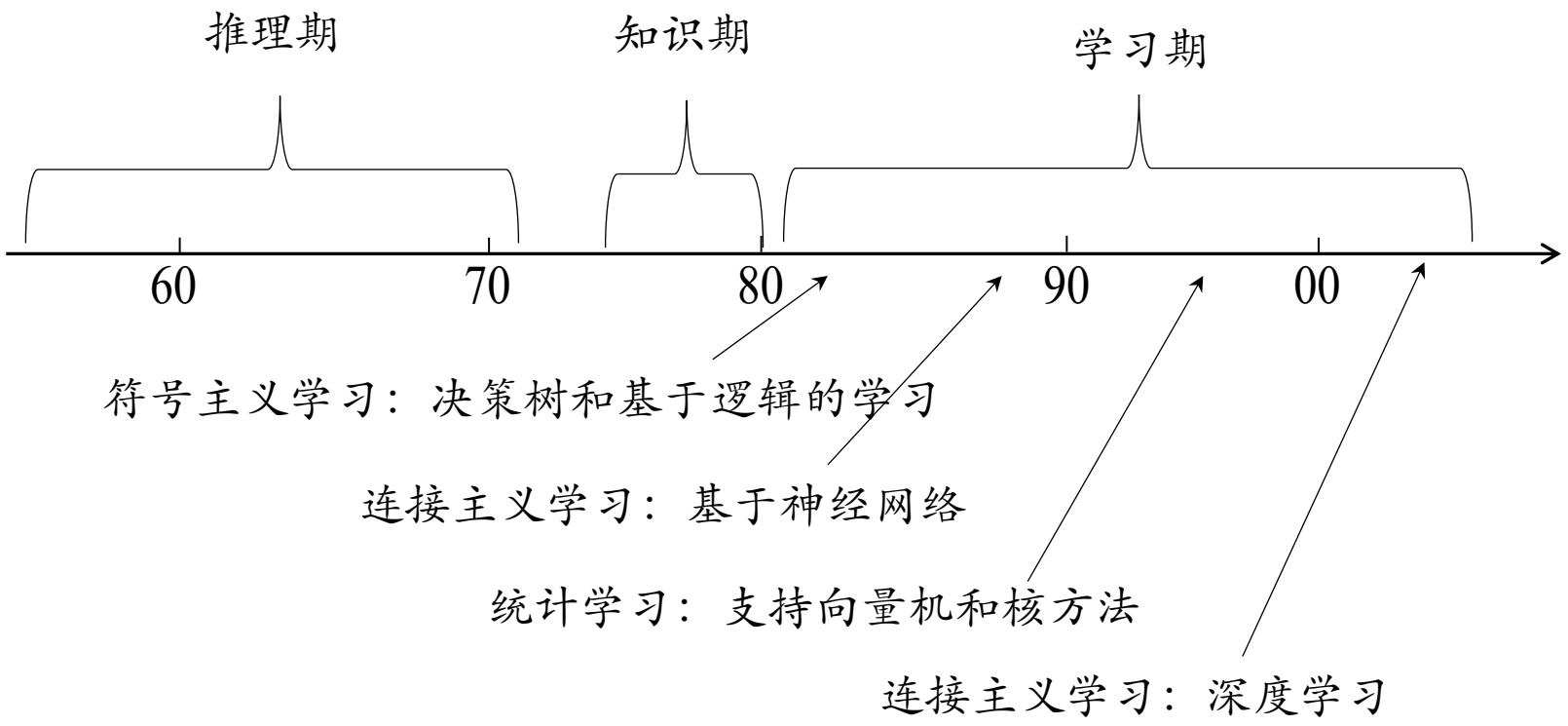
发展历程

- 推理期：
  - A. Newell和H. Simon的“逻辑理论家”(Logic Theorist)程序以及伺候的“通用问题求解”(General Problem Solving)程序等在当时取得了令人振奋的结果。
  - 2006年卡耐基梅隆大学宣告成立第一个“机器学习系”，机器学习奠基人之一T.Mitchell教授任系主任。
- 知识期：
  - 大量专家系统问世，在很多应用领域取得大量成果；
  - 但是由人来总结知识再交给计算机相当困难。

发展历程

- 学习期：
  - 符号主义学习
    - 决策树：以信息论为基础，最小化信息熵，模拟了人类对概念进行判定的树形流程
    - 基于逻辑的学习：使用一节逻辑进行知识表示，通过修改扩充逻辑表达式对数据进行归纳
  - 连接主义学习
    - 神经网络
  - 统计学习
    - 支持向量机及核方法

发展历程


推理期
知识期
学习期
60
70
80
90
00
符号主义学习：决策树和基于逻辑的学习
连接主义学习：基于神经网络
统计学习：支持向量机和核方法
连接主义学习：深度学习

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

应用现状

- 计算机领域最活跃的研究分支之一：
  - NASA_JPL科学家在Science撰文指出机器学习对科学研究起到越来越大的支撑作用
  - DARPA启动PAL计划，将机器学习的重要性提高到国家安全的高度来考虑
  - 2006年卡耐基梅隆大学宣告成立第一个“机器学习系”，机器学习奠基人之一T.Mitchell教授任系主任。

- 与普通人的生活密切相关：
  - 天气预报、能源勘探、环境监测、搜索引擎、自动驾驶汽车等

应用现状

- 影响到人类社会的政治生活：
  - 2012美国大选期间奥巴马麾下的机器学习团队，对社交网络等各类数据进行分析，为其提示下一步的竞选行动。

- 具有自然科学探索色彩：
  - P.Kanerva在二十世纪八十年代中期提出SDM(Sparse Distributed Memory)模型时并没有刻意模仿脑生理结构，但后来神经科学的研究发现，SDM的稀疏编码机制在视觉、听觉、嗅觉功能的脑皮层中广泛存在，促进理解“人类如何学习”

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

阅读材料

- [Mitchell, 1997]是第一本机器学习专门教材. [Duda et al., 2001; Alpaydin, 2004; Flach, 2012]为出色的入门读物. [Hastie et al., 2009]为进阶读物, [Bishop, 2006]适合于贝叶斯学习偏好者. [Shalev-Shwartz and Ben-David, 2014]适合于理论偏好者.

- 《机器学习:一种人工智能途径》 [Michalski et al., 1983]汇集了20位学者撰写16篇文章，是机器学习早期最重要的文献. [Dietterich, 1997] 对机器学习领域的发展进行了评述和展望。

阅读材料

- 机器学习领域最重要的国际学术会议是国际机器学习会议(ICML)、国际神经信息处理系统会议(NIPS)和国际学习理论会议(COLT),重要的区域性会议主要有欧洲机器学习会议(ECML)和亚洲机器学习会议(ACML);最重要的国际学术期刊是Journal of Maching Learning Research和Machine Learning.

- 国内不少书记包含机器学习方面的内容，例如[陆汝钤,1996].[李航,2012]是一统计学习为主题的读物. 国内机器学习领域最重要的活动是两年一次的中国机器学习大会(CCML)以及每年举行的“机器学习及其应用”研讨会(MLA).