

课程信息

- 名称：机器学习
- 考核：算法笔记+实验报告
- 推荐教材
 - 《机器学习》，清华大学出版社，周志华，**2016**
 - 《机器学习从原理到应用》，人民邮电出版社，卿来云、黄庆明，2020
 - 《统计学习方法》，清华大学出版社，李航，2019
 - 《Machine Learning: The Art and Science of Algorithms that Make Sense of Data》，Cambridge University Press, Peter Flach, 2012

教材建议使用方式

1. 初学机器学习的第一本书：
通读、速读；细节不懂处略过
了解机器学习的疆域和基本思想，理解基本概念
“观其大略”
2. 阅读其他关于机器学习具体分支的读物（三月、半年）
3. 再读、对“关键点”的理解：
理解技术细冗后的本质，升华认识
“提纲挈领”
4. 对机器学习多个分支有所了解（1-3年）
5. 再读、细思：
不同内容的联系，不同的描述方式、出现位置蕴涵的意义、.....个别字句的
启发可能自行摸索数年不易得
“疏通经络”

<http://www.lamda.nju.edu.cn/zhoush/zhoush.files/publication/MLbook2016.htm>



课程定位

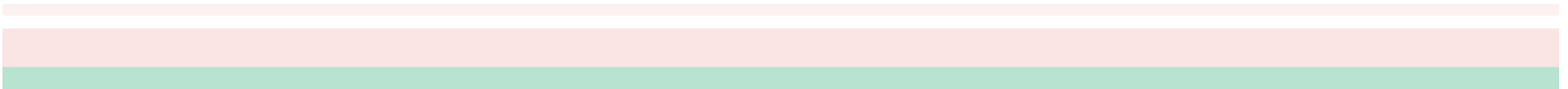
- 机器学习是什么学科？

- 科学

- 技术

- 工程

- 应用



课程定位

- 机器学习是什么学科？
 - 科学：是什么，为什么
 - 技术：怎么做
 - 工程：多、快、好、省
 - 应用：在相关行业和领域的实际应用

第一章：绪论

The bottom of the slide features three horizontal bars of equal width. The top bar is light pink, the middle bar is a slightly darker shade of pink, and the bottom bar is a light green color.

大纲

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

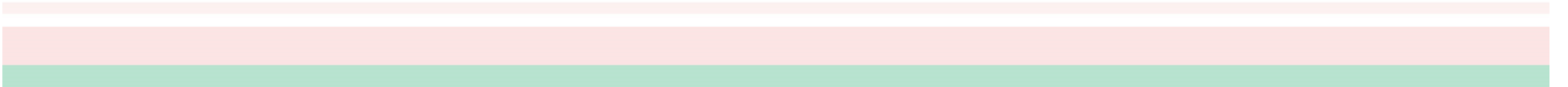
什么是机器学习

“Machine learning is the study of computer algorithms that can improve automatically through experience and by the use of data.” (Mitchell, 1997)

Learning = Improving with experience at some task

- improve over task T
- with respect to performance measure P
- based on experience E

E.g., Learn to play checkers

- T : play checkers
 - E : opportunity to play against self
 - P : % of games won in world tournament
- 

什么是机器学习

“假设用 P 来评估计算机程序在某类任务 T 上的性能，若一个程序通过利用经验 E 在任务 T 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” (Mitchell, 1997)

例1：垃圾邮件分类

- 任务 T ：标记每一封邮件是否为垃圾邮件
- 经验 E ：收集到很多邮件，包含垃圾邮件和正常邮件
- 性能指标 P ：机器标记垃圾邮件的准确率

例2：玩跳棋

- 任务 T ：下棋
- 经验 E ：玩很多跳棋游戏的经验
- 性能指标 P ：机器赢得比赛的概率

什么是机器学习

“假设用 P 来评估计算机程序在某类任务 T 上的性能，若一个程序通过利用经验 E 在任务 T 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” (Mitchell, 1997)

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

什么是机器学习

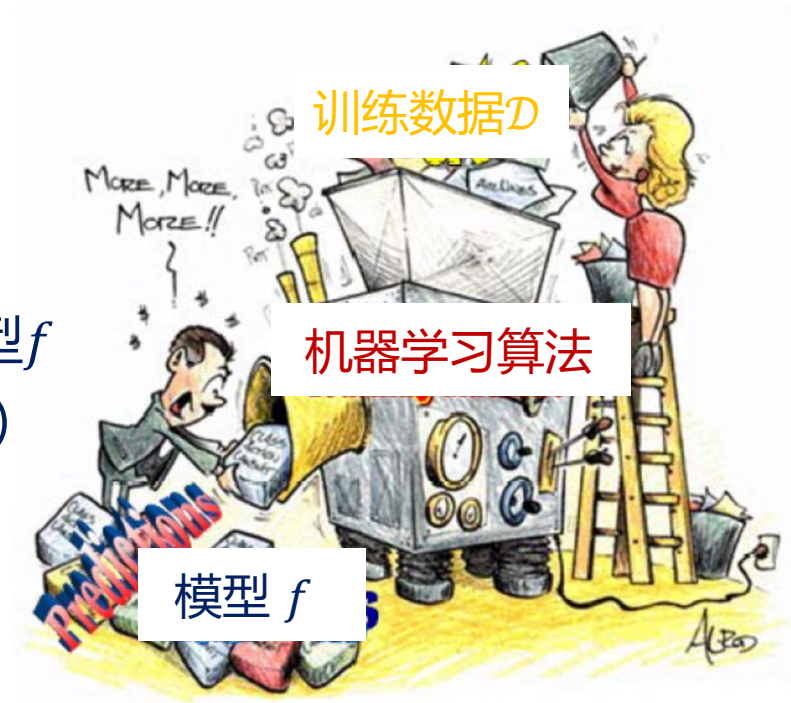
“假设用 P 来评估计算机程序在某类任务 T 上的性能，若一个程序通过利用经验 E 在任务 T 上获得了性能改善，则我们就说关于 T 和 P ，该程序对 E 进行了学习” (Mitchell, 1997)

经验 E : 训练数据 \mathcal{D}

模型: 预测函数 f

学习算法: 怎样从经验 E 中得到模型 f

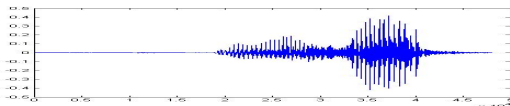
性能评价: 模型有多好 (测试数据)



机器学习 \approx 找到一个函数 f

根据经验


■ 语音识别

$$f(\text{  }) = \text{“How are you”}$$

■ 图像识别

$$f(\text{  }) = \text{“猫”}$$

■ 下棋

$$f(\text{  }) = \text{“5-5” (下一次移动)}$$

与其他领域的关系

- Pattern recognition is the automated recognition of patterns and regularities in data
- In machine learning, pattern recognition is the assignment of a label to a given input value

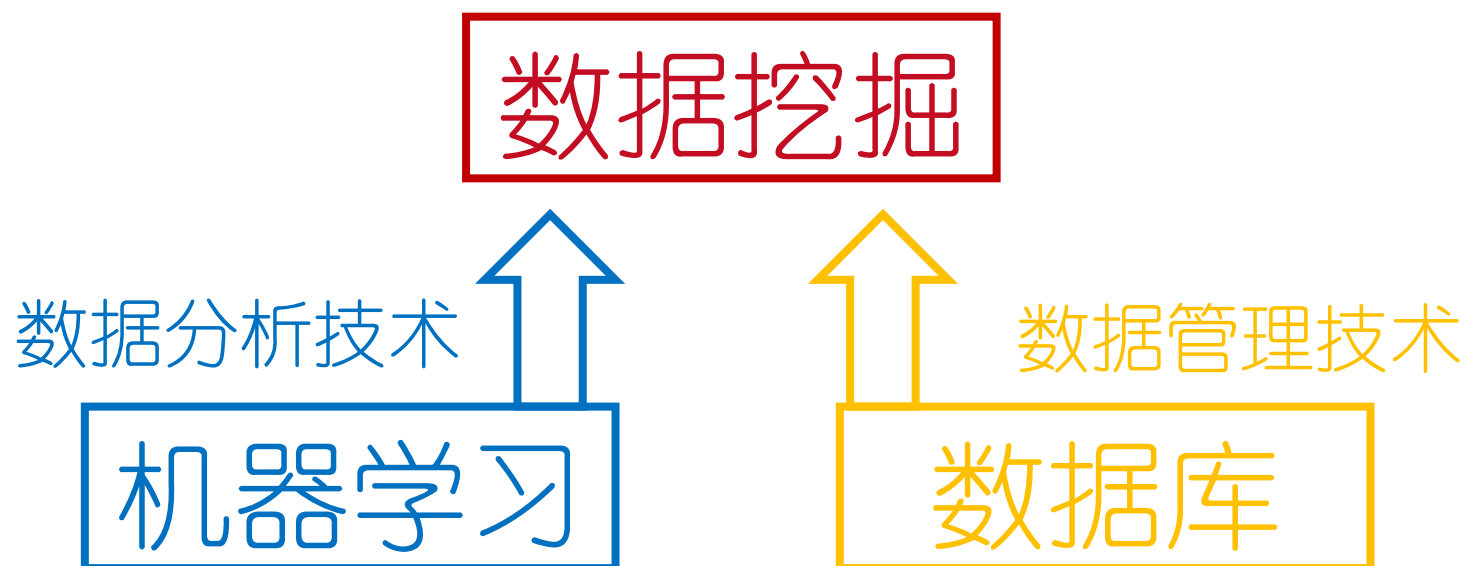
https://en.wikipedia.org/wiki/Pattern_recognition

机器学习

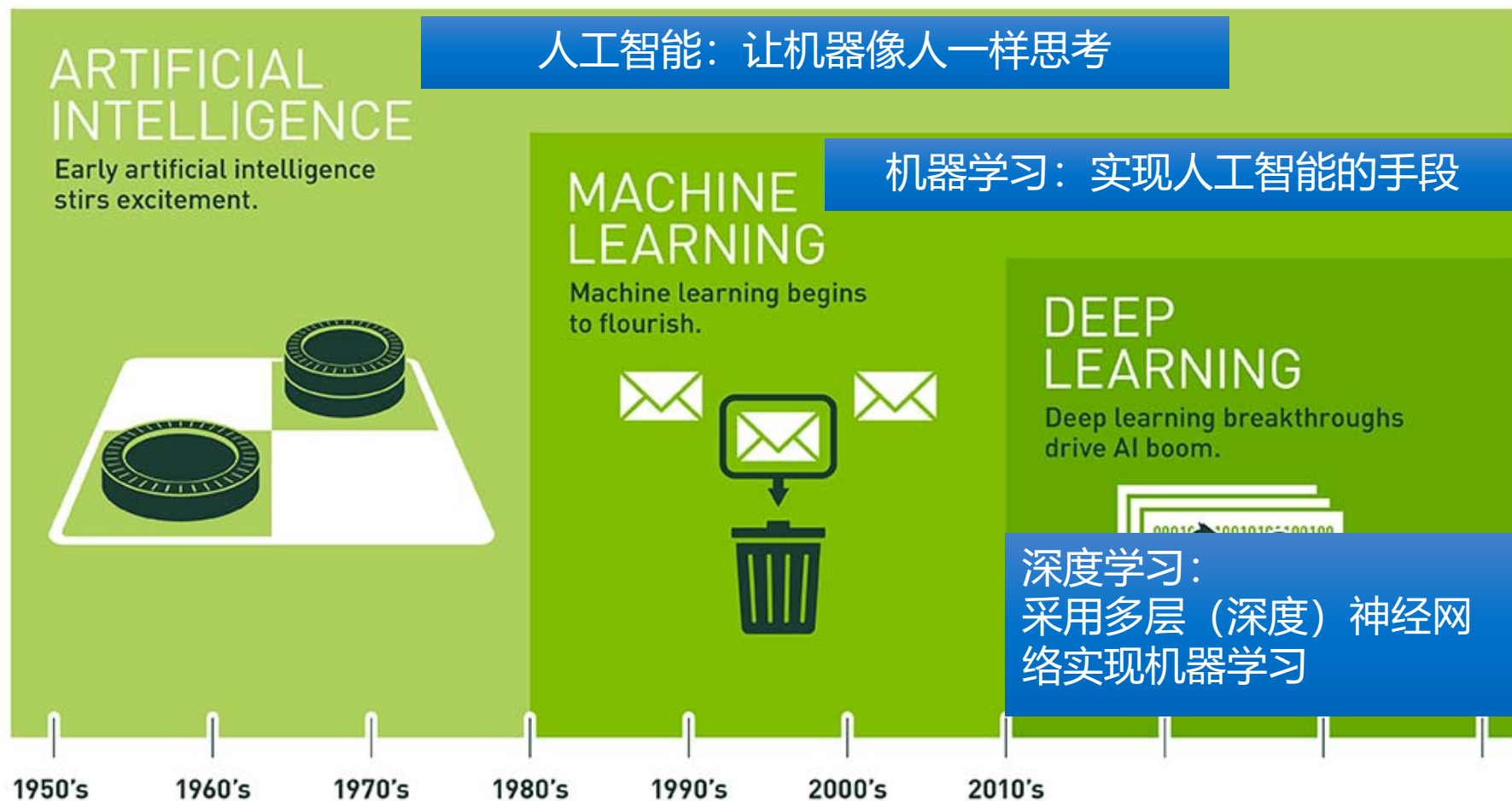
vs.

模式识别

与其他领域的关系



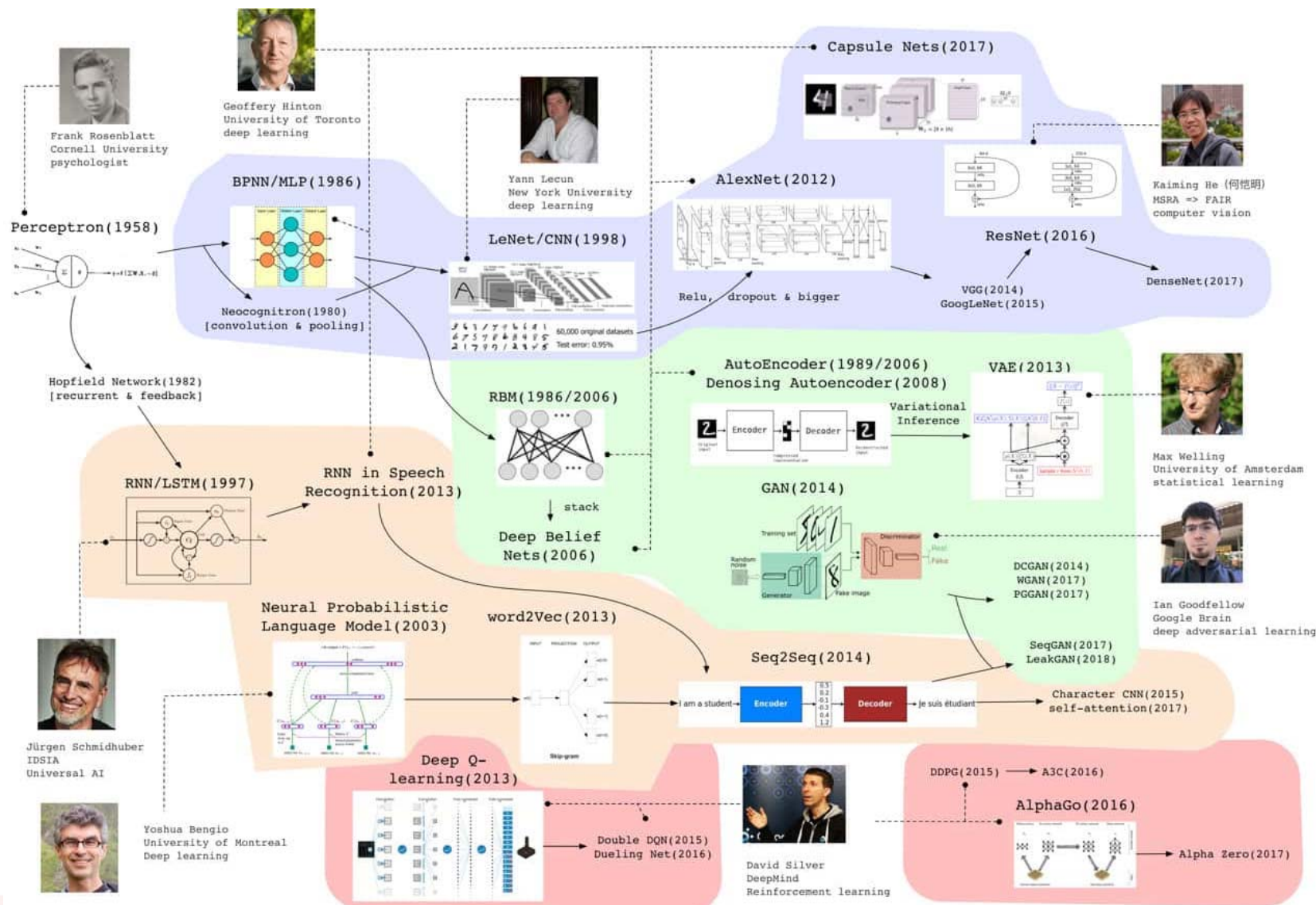
与其他领域的关系



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

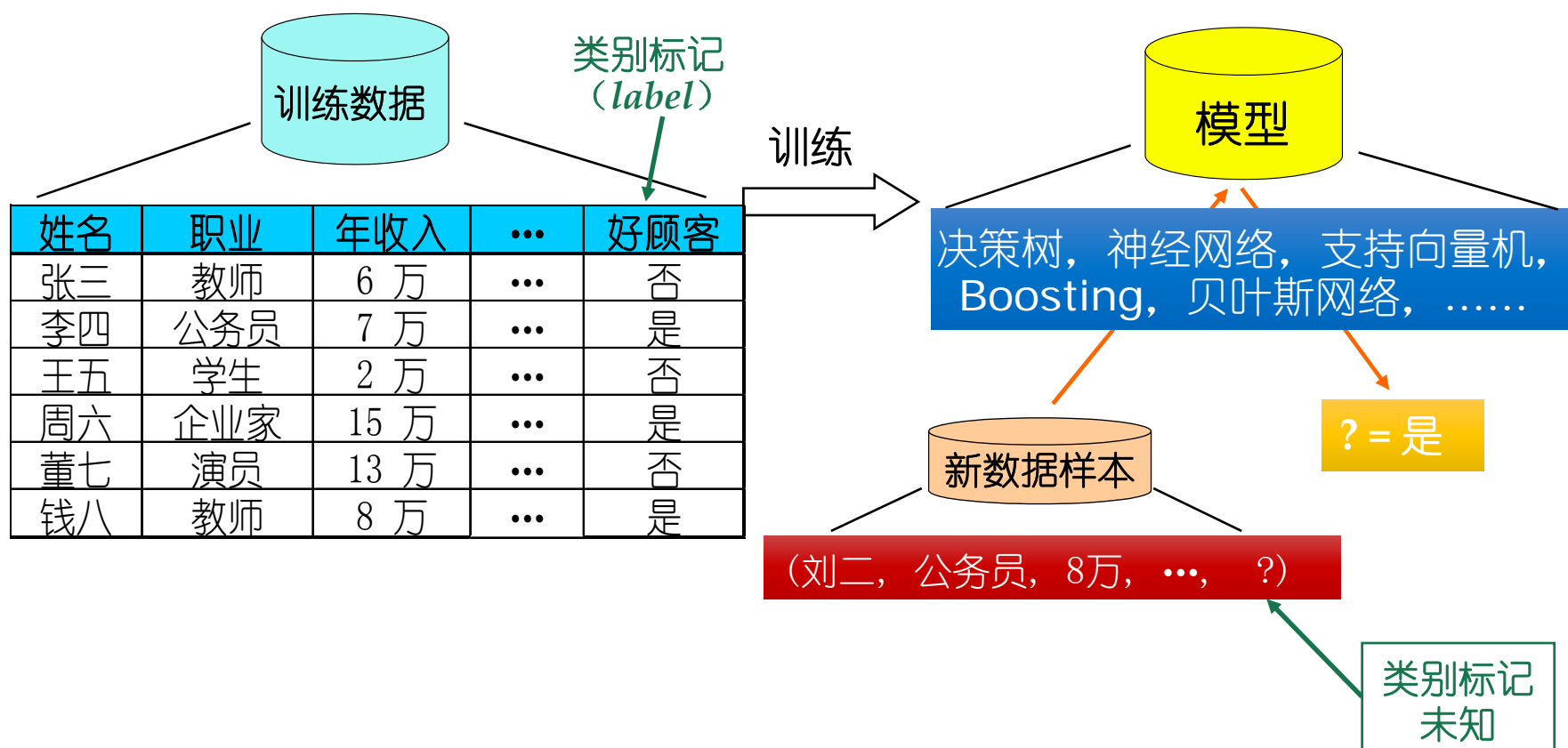
图片：<https://blogs.nvidia.com.tw/2016/07/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>

深度学习模型近年的重要进展



典型的机器学习过程

使用学习算法 (*learning algorithm*)



机器学习有坚实的理论基础

□ 计算学习理论(Computational learning)

➤ 概率近似正确(Probably Approximately Correct, PAC)

我们希望以比较大的把握学得比较好的模型, 即以较大概率学得误差满足预设上限的模型.

$$P(E(h) \leq \epsilon) \geq 1 - \delta,$$

这样的学习算法 \mathcal{L} 能以较大概率(至少 $1 - \delta$) 学得目标概念的近似(误差最多为 ϵ).

机器学习有坚实的理论基础

- 对于给定训练集 D , 我们希望基于学习算法 \mathcal{L} 学得模型所对应的假设尽可能接近目标概念.

为什么不是希望精确地学到目标概念呢？

机器学习过程受到很多因素的制约

- 获得的训练集 D 往往仅包含有限数量的样例, 因此通常会存在一些在 D 上 “等效” 的假设, 学习算法无法区别这些假设;
- 从分布 \mathcal{D} 采样得到 D 的过程有一定的偶然性, 即便对同样大小的不同训练集, 学得结果也可能有所不同.

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

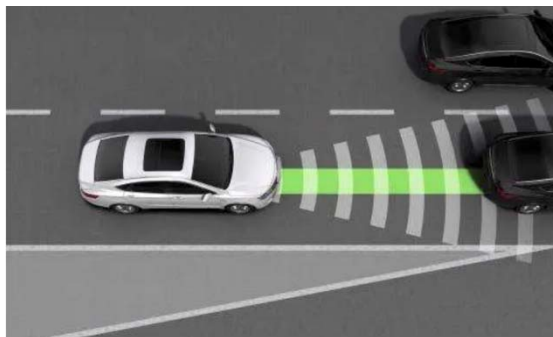
□ 具有自然科学探索色彩：

- 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”

应用现状：计算机视觉



监控



自动驾驶



人脸识别



农业

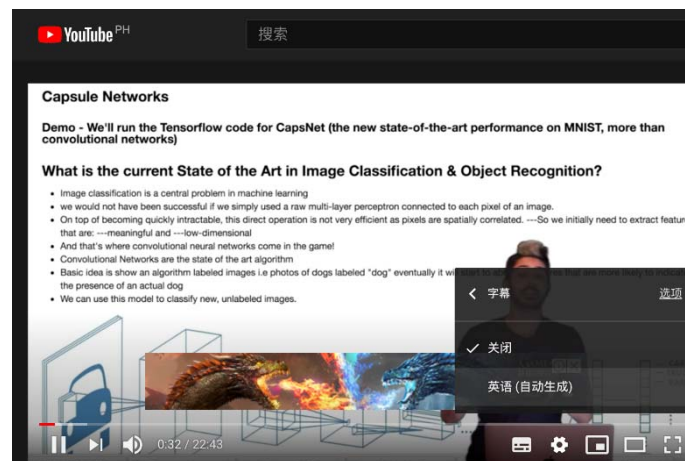


制造业



医疗

应用现状：语音识别



应用现状：搜索



应用现状：推荐

亚马逊
amazon.cn

You Tube



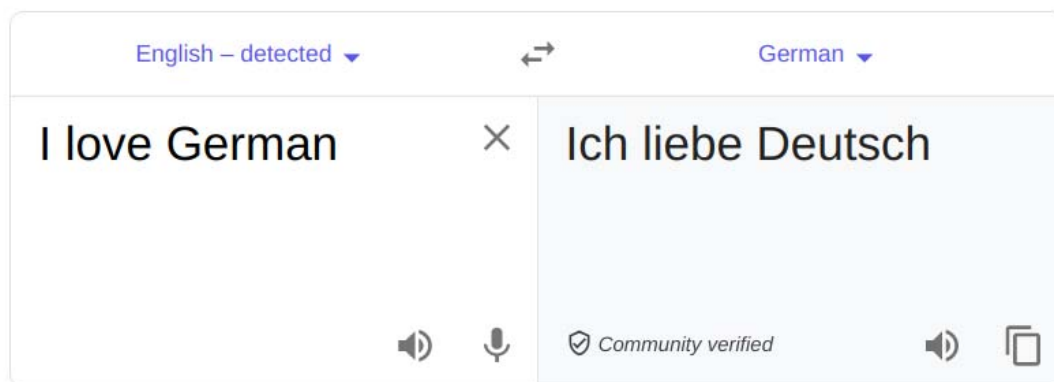
30%以上的购买来自推荐



75%以上的观看来自推荐



应用现状：自然语音处理



机器翻译



社交媒体监控

The most common type of marketing channel is the wholesale market. Varies kinds of **produce** are supplied from different areas are assembled at one place and sold through smaller regional markets, etc. Fruits and vegetables are handled and transport methods.

Replace the word

products

Dismiss

Suggested by Grammarly

语法检查



电子邮件过滤

应用现状：游戏



AlphaGo & 围棋



AlphaStar & 星海争霸II



Libratus & 德州扑克



绝悟AI & 王者荣耀

大纲

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

		特征 ↑			标记 ↑
	编号	色泽	根蒂	敲声	好瓜
训练集 ←	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
测试集 ←	1	青绿	蜷缩	沉闷	?

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

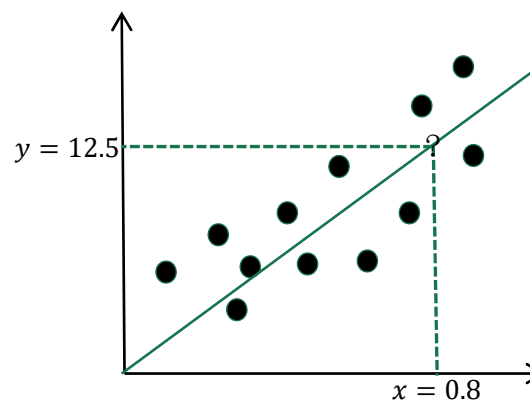
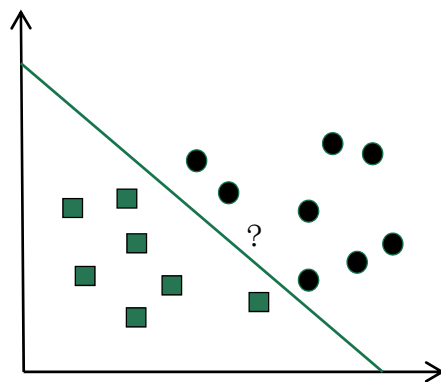
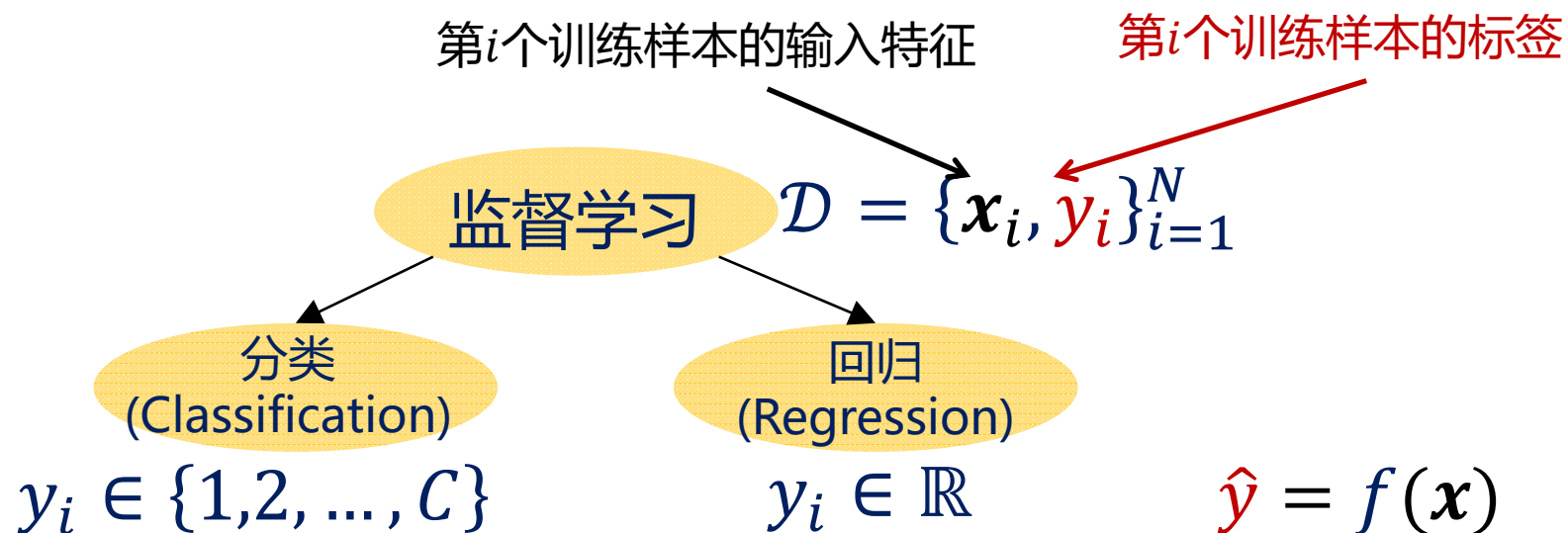
- 聚类: 无标记信息

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、降维、密度估计
- 半监督学习：两者结合
- 强化学习：具有“延迟标记信息”的监督学习

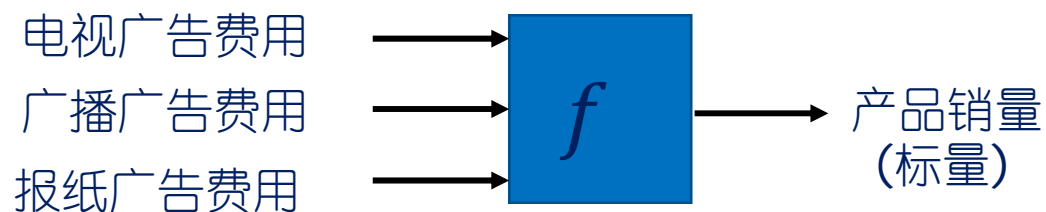
基本术语-任务



回归

例：产品销量预测

函数 f 的输出为标量 $\hat{y} \in \mathbb{R}$



训练数据:

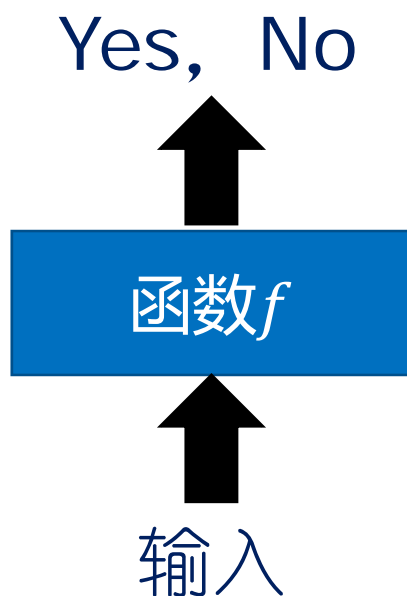
输入

电视广告费用	广播广告费用	报纸广告费用	产品销量
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

输出

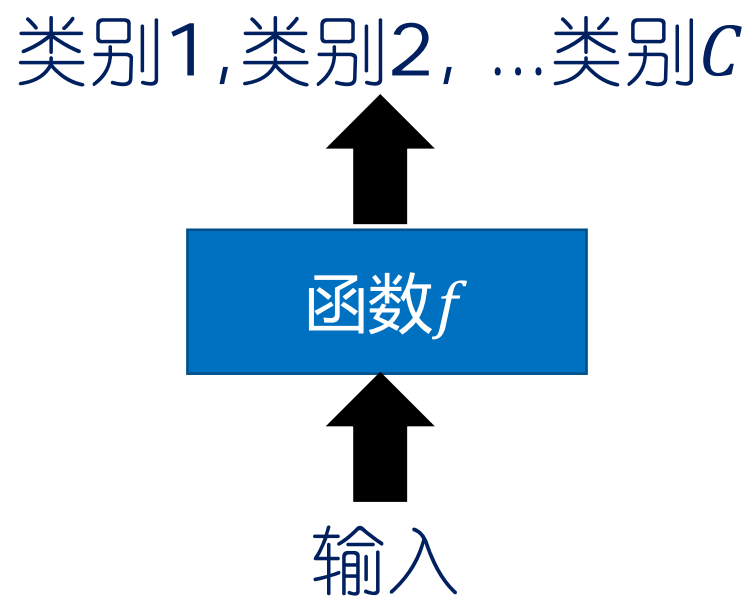
分类

- 二分类



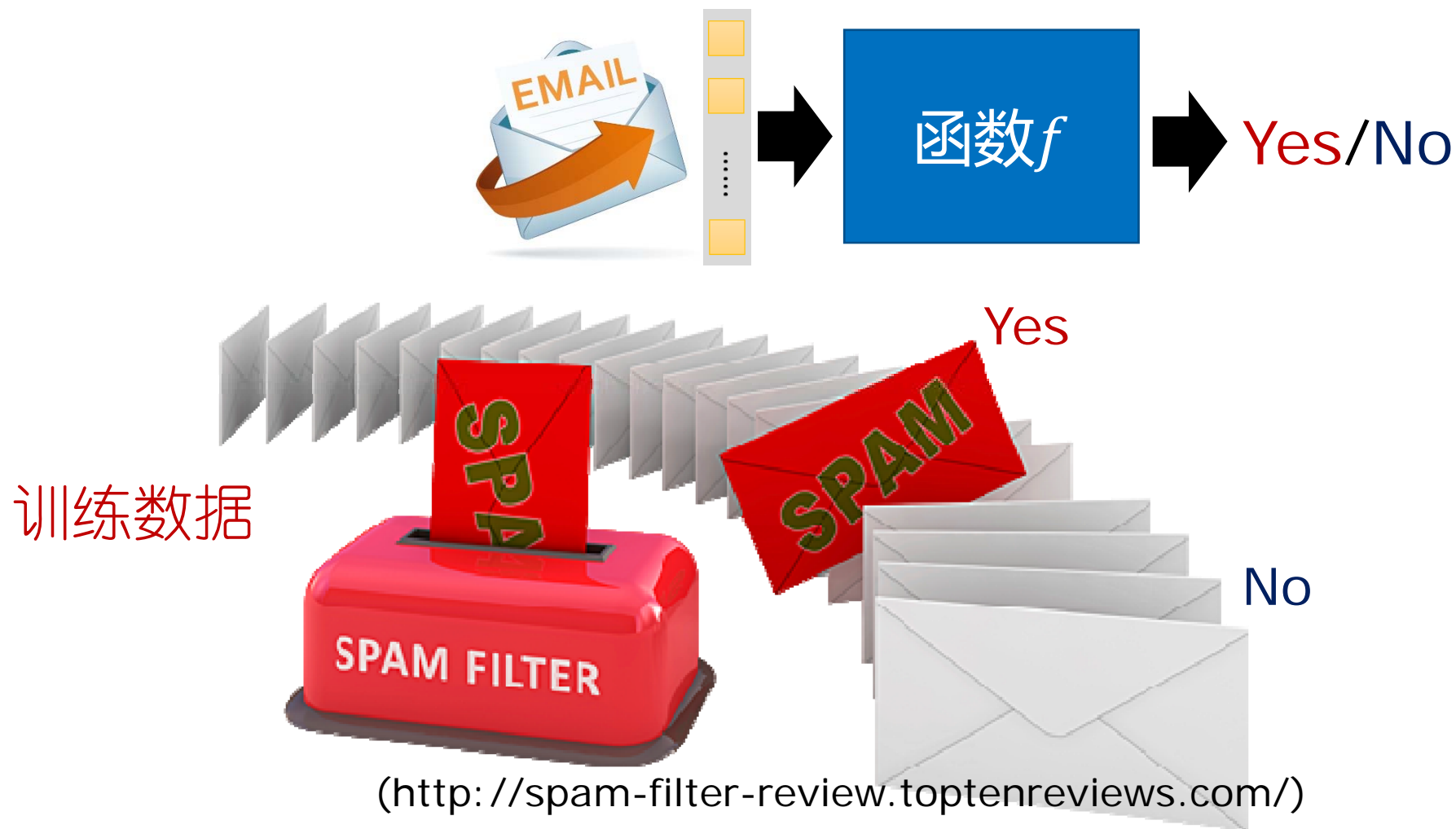
函数 f 的输出为离散值

- 多分类



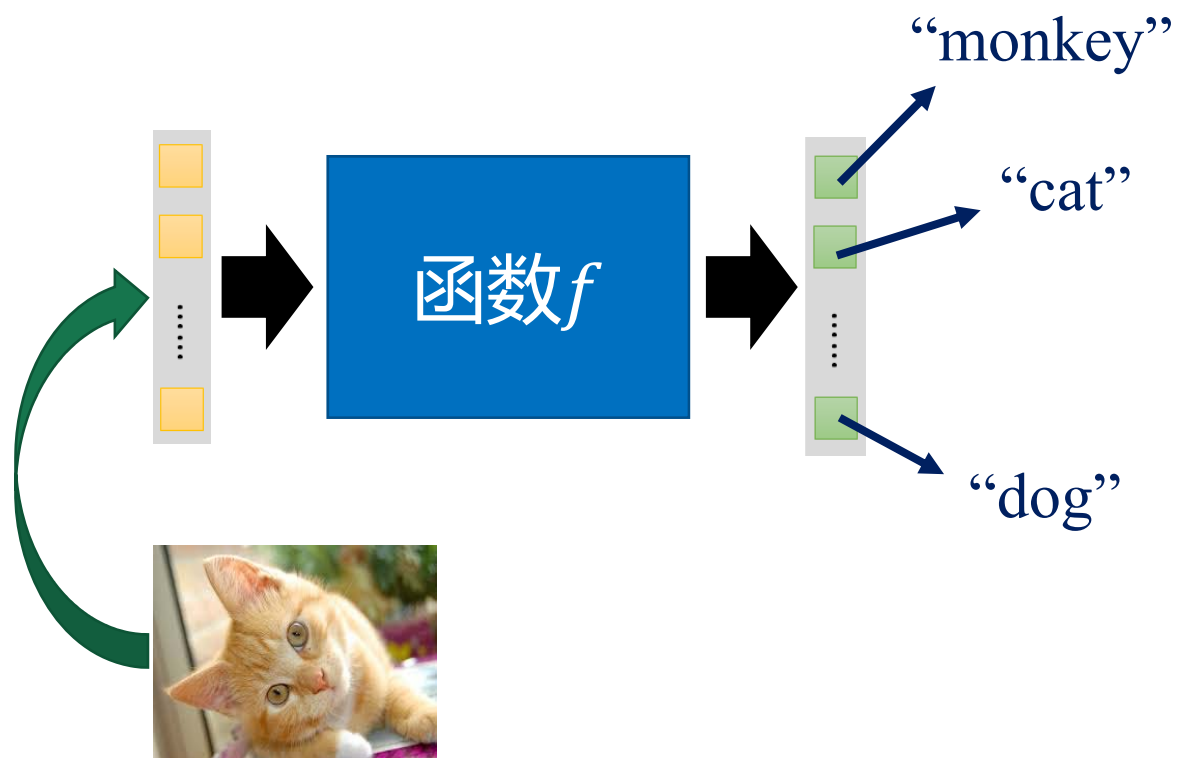
二分类

垃圾邮件过滤

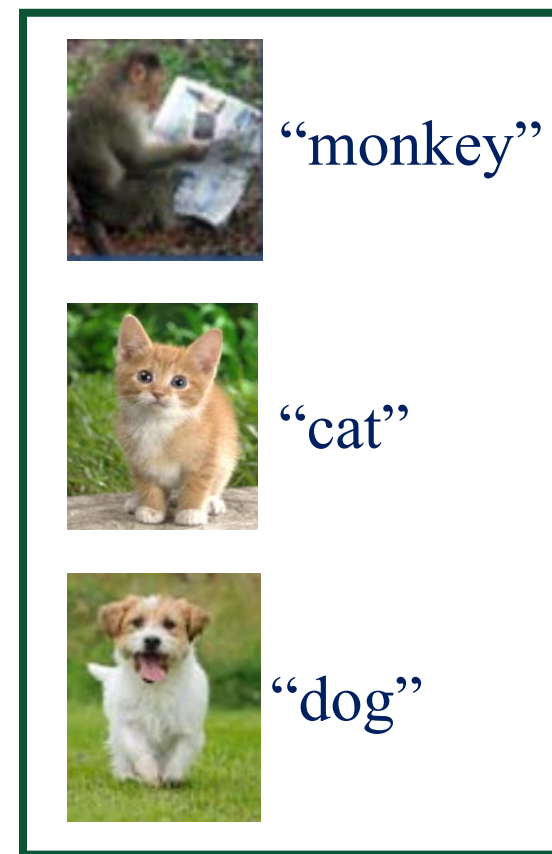


多分类

图像识别

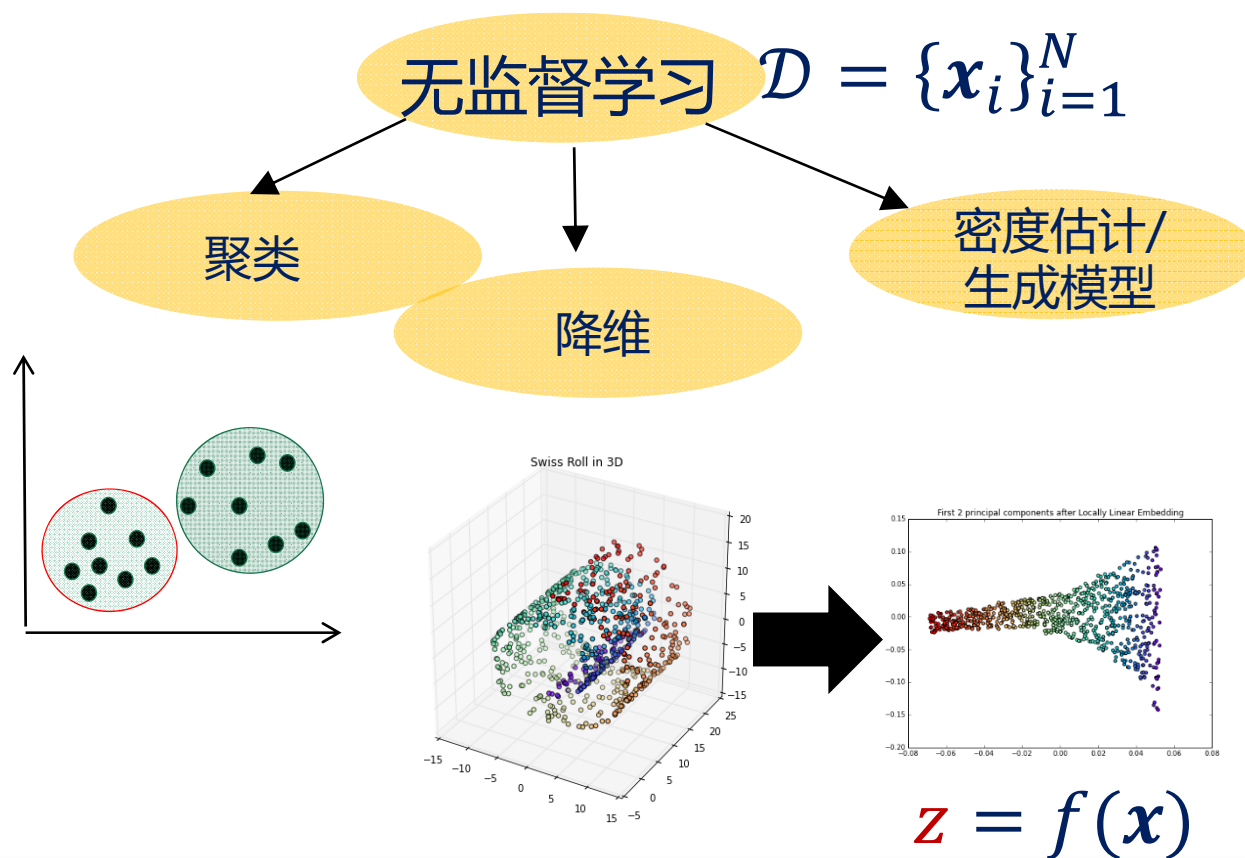


训练数据



无监督学习

发现数据中的“有意义的模式”



聚类



形状



尺寸



线种



颜色



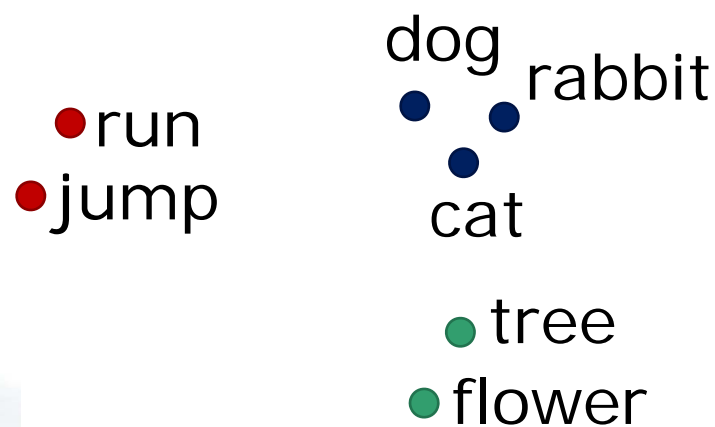
降维/嵌入表示

词嵌入：机器读入很多文档，学习到每个单词的表示



<http://top-breaking-news.com/>

Word Embedding



强化学习

■ 监督学习



下一次移动:
"5-5"



下一次移动:
"3-3"

■ 强化学习

第一次移动 ➡ 移动..... ➡ 赢!

- 从行为的反馈（奖励或惩罚）中学习
 - 设计一个回报函数
 - 强化学习的任务：找到一条回报值最大的路径

AlphaGo: 监督学习 + 强化学习

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

基本术语

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)

使用学习算法 (learning algorithm)

训练数据

类别标记 (label)

色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

训练

模型

决策树, 神经网络, 支持向量机,
Boosting, 贝叶斯网,

新数据样本

(浅白, 蜷缩, 浊响, ?)

? = 是

类别标
记未知

- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

- 分类, 回归
- 二分类, 多分类
- 正类, 反类

- 数据集; 训练, 测试
- 示例(instance), 样例(example)
- 样本(sample)
- 属性(attribute), 特征(feature); 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间

- 未见样本(unseen instance)
- 未知 “分布”
- 独立同分布(i.i.d.)
- 泛化(generalization)

大纲

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

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

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

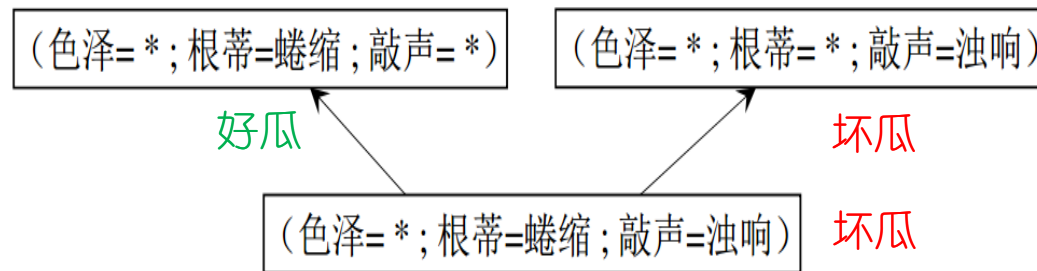
在模型空间中搜索不违背训练集的假设
假设空间大小： $4*4*4+1=65$

大纲

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

归纳偏好

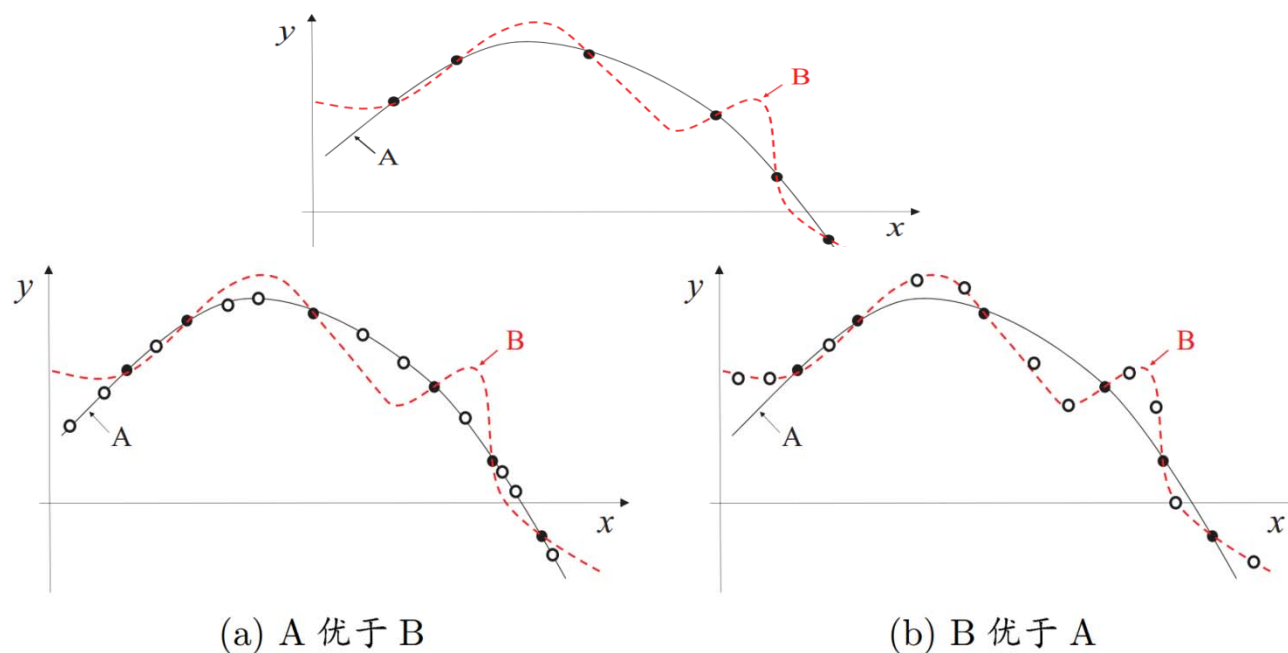
假设空间中有三个与训练集一致的假设，但他们对
(色泽=青绿；根蒂=蜷缩；敲声=沉闷)的瓜会预测
出不同的结果：



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



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 $P(h|X, \mathcal{L}_a)$ 代表算法 \mathcal{L}_a 基于训练数据 X 产生假设 h 的概率, 再令 f 代表要学的目标函数, \mathcal{L}_a 在训练集之外所有样本上的总误差为

$$E_{ote}(\mathcal{L}_a|X, f) = \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathcal{L}_a)$$

$\mathbb{I}(\cdot)$ 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：

$$\begin{aligned} \sum_f E_{ote}(\mathcal{L}_a | X, f) &= \sum_f \sum_h \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) P(h | X, \mathcal{L}_a) \\ &= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \sum_f \mathbb{I}(h(\mathbf{x}) \neq f(\mathbf{x})) \\ &= \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\ &= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \sum_h P(h | X, \mathcal{L}_a) \\ &= 2^{|\mathcal{X}|-1} \sum_{\mathbf{x} \in \mathcal{X} - X} P(\mathbf{x}) \cdot 1. \quad \text{总误差与学习算法无关!} \end{aligned}$$

NoFreeLunch

NFL定理的重要前提：

所有“问题”出现的机会相同、或所有问题同等重要

实际情形并非如此；我们通常只关注自己正在试图解决的问题

脱离具体问题，空泛地谈论“什么学习算法更好”毫无意义！

具体问题，具体分析！

现实应用

- 把机器学习的“十大算法”“二十大算法”都弄熟，逐个试一遍，是否就“止于至善”了？
- 机器学习并非“十大套路”“二十大招数”的简单堆积
- 现实任务千变万化，以有限的“套路”应对无限的“问题”，焉有不败？
- 最优方案往往来自：按需设计、度身定制

大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 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：

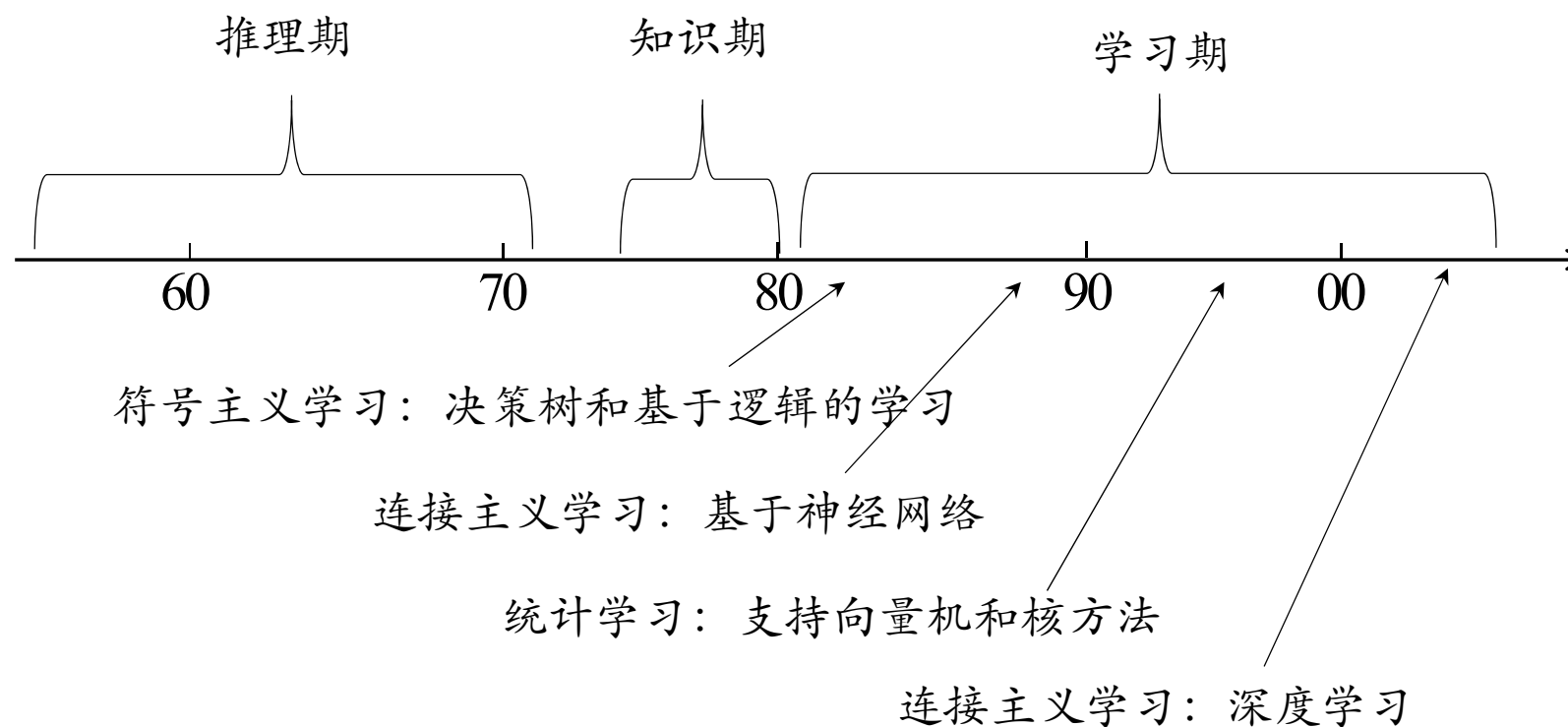
- 大量专家系统问世，在很多应用领域取得大量成果
- 但是由人来总结知识再交给计算机相当困难

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



大纲

- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 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(NeurIPS)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 Machine Learning Research和Machine Learning (CCF推荐会议/期刊)
- 国内不少书记包含机器学习方面的内容,例如[陆汝钐, 1996]. [李航, 2012]是以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)