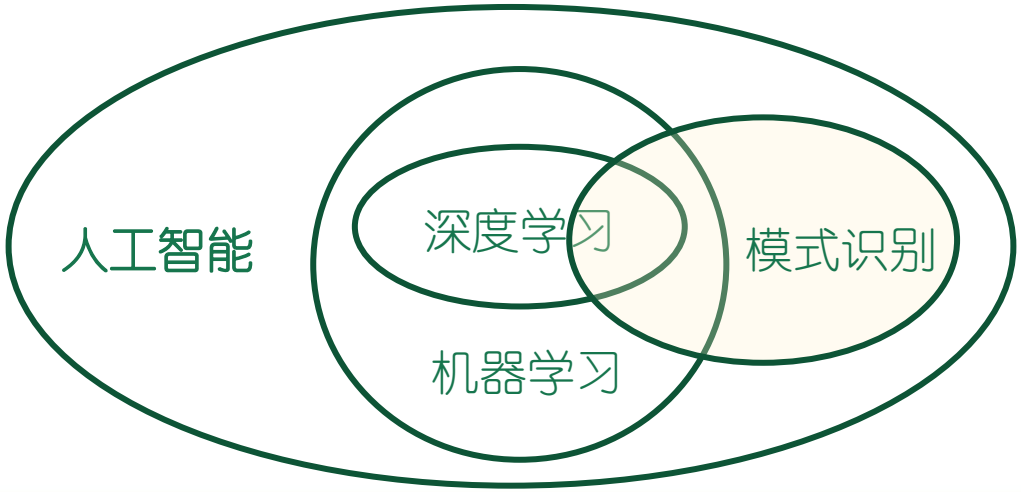
第一章：绪论

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

- 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。
- “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”



什么是人工智能

人工智能（Artificial Intelligence），英文缩写为**AI**。它是研究、开发用于**模拟、延伸和扩展人的智能**的理论、方法、技术及应用系统的一门新技术科学。

- ✓ 结构模拟：机器人学
- ✓ 功能模拟：以任务为核心

  模式识别、**机器学习、深度学习**

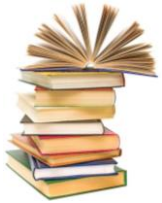
  自然语言处理、音频识别、定位跟踪、图像理解、知识推理、数据预测等等

人工智能

使一部机器的反应方式像人一样进行**感知、认知、决策、执行**的人工程序或系统

什么是机器学习

机器学习模型



经验=数据

Task 1-model

Task 2-model

Task 3-model

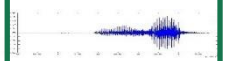
...

智能计算机

- ✓ 什么是机器学习任务?
- ✓ 什么是机器学习学习模型?
- ✓ 什么是机器学习经验/数据?
- ✓ 什么是机器学习性能指标?

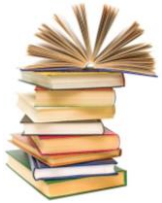
机器学习

对于某类任务T（Task）和性能度量P（Performance Measure），一个计算机程序被认为可以从经验E（Experience）中学习是指，通过经验E改进后，计算机程序在任务T上由性能度量P衡量的性能有所提升

任务	模型	输入	性能评价	输出
图像识别	$f($		$)=$	“Cat”
语音识别	$f($		$)=$	“你好”
文本分类	$f($		$)=$	“体育新闻”
机器翻译	$f($	“你几岁了”	$)=$	“How old are you”

什么是深度学习

机器学习模型



经验=数据

Task 1-model

Task 2-model

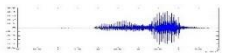
Task 3-model

...

智能计算机

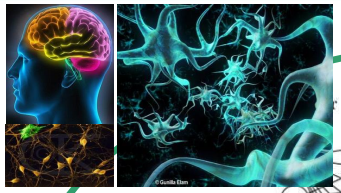
人是地球上具有最高智慧的动物，人类靠大脑进行思考、联想、记忆和推理判断。

建立**模仿人类大脑**的模型？

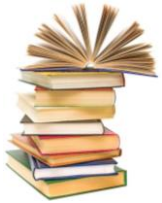
图像识别	$f($		$)=$	“Cat”
语音识别	$f($		$)=$	“你好”
文本分类	$f($		$)=$	“体育新闻”
机器翻译	$f($	“你几岁了”	$)=$	“How old are you”

模型

什么是深度学习



机器学习模型



经验=数据

Task 1-model

Task 2-model

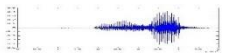
Task 3-model

...

智能计算机

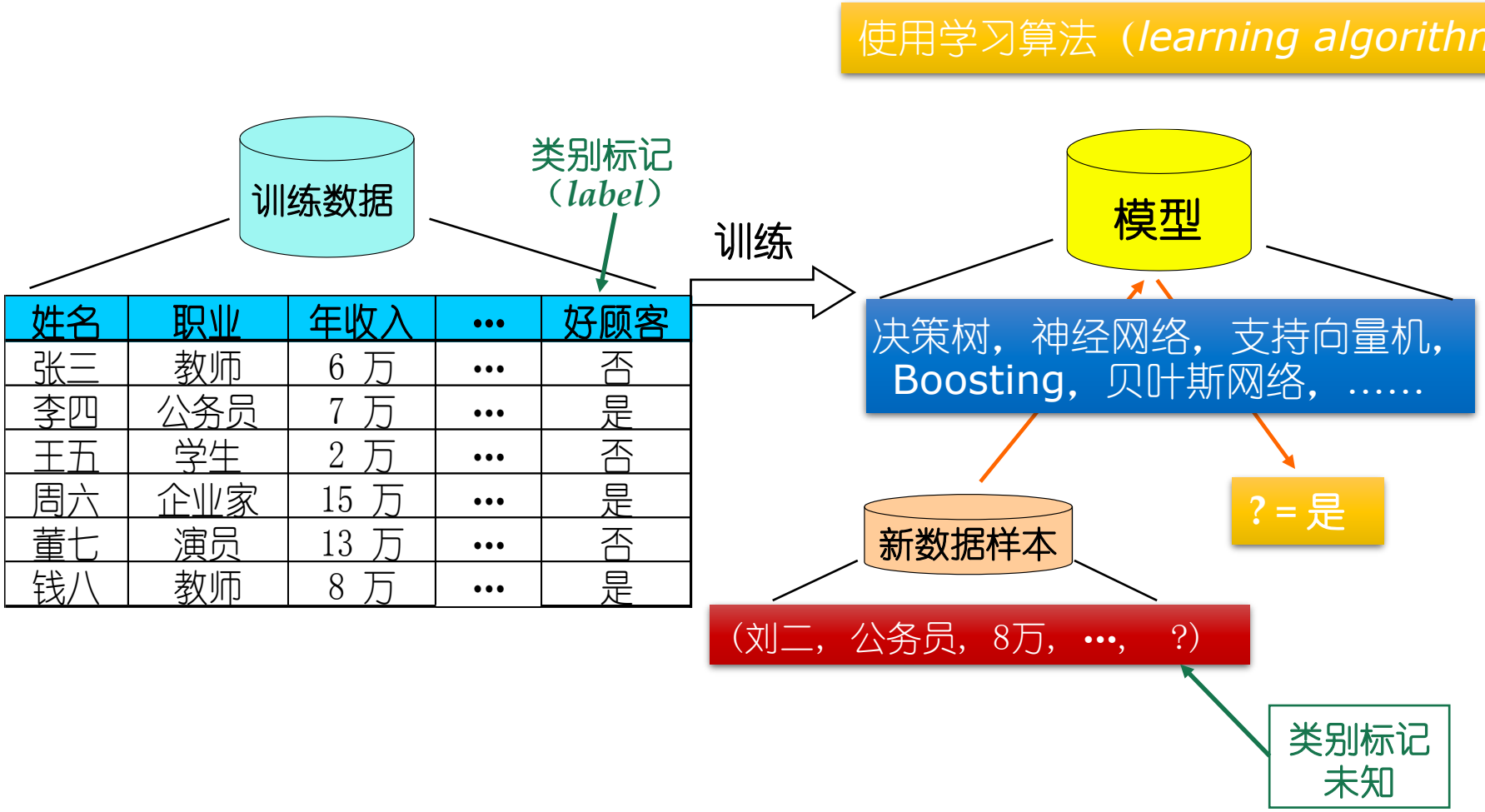
人是地球上具有最高智慧的动物，人类靠大脑进行思考、联想、记忆和推理判断。

建立**模仿人类大脑**的模型？

图像识别	$f($		$)=$	“Cat”
语音识别	$f($		$)=$	“你好”
文本分类	$f($		$)=$	“体育新闻”
机器翻译	$f($	“你几岁了”	$)=$	“How old are you”

模型

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

$$D = \{x_1, x_2, \cdots, x_m\}$$

$$x_i = (x_{i1}, x_{i2}; \cdots; x_{id})$$

(x_i, y_i)

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

机器学习方法

有监督学习（supervised learning）：从给定的**有标注的训练数据集**中学习出一个函数（模型参数），当新的数据到来时可以根据这个函数预测结果。常见任务包括**分类**与**回归**。

Classification: Y is discrete

Y：年轻人(1)，老年人(-1)

X： x_1 黑头发的比例， 值域（0，1）；

x_2 行走速度， 值域（0，100）米/每分钟.

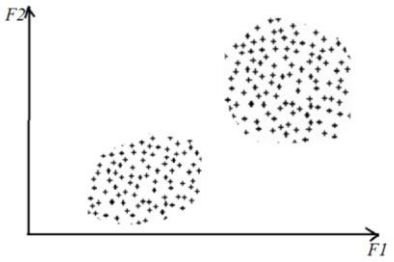
Training Data:

Y=1：(1，99)、(0.9，80)、(0.80，100) …

Y=-1：(0.2，30)、(0.5，50)、(0.4，30) …

Test:

X=(0.85, 98), Y=?



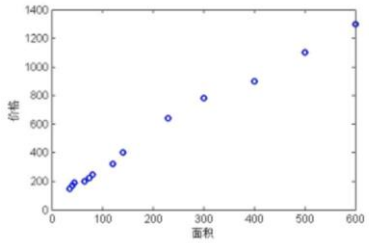
Regression: Y is continue

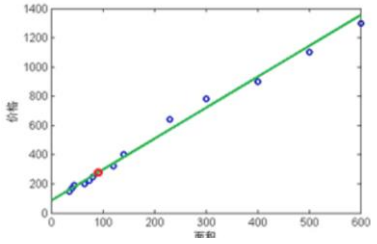
Y：房屋价钱（万元），值域 Y>=0.

X: x1=房屋面积 m^2

Training Data:

35	150
40	170
45	190
65	200
74	224
80	245
120	320
140	400
230	640
300	780
400	900
500	1100
600	1300



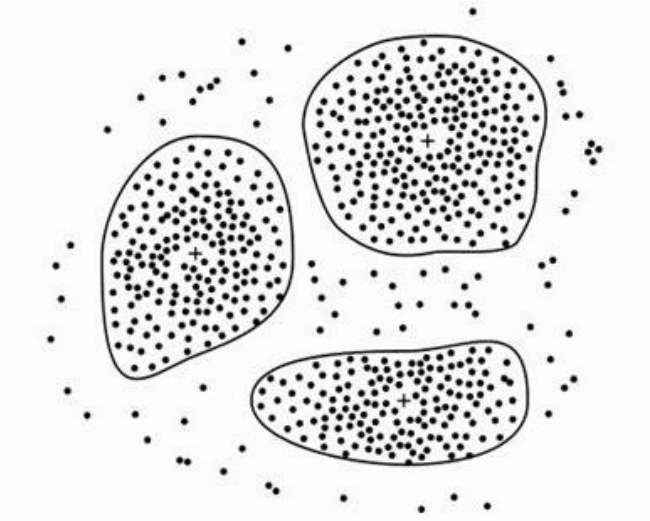


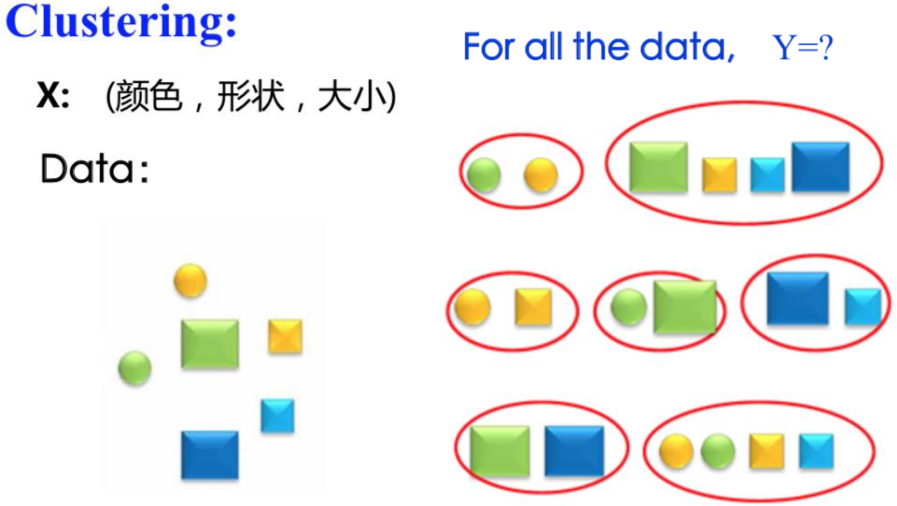
$y=ax+b$

Test: X=90 Y= ?

机器学习方法

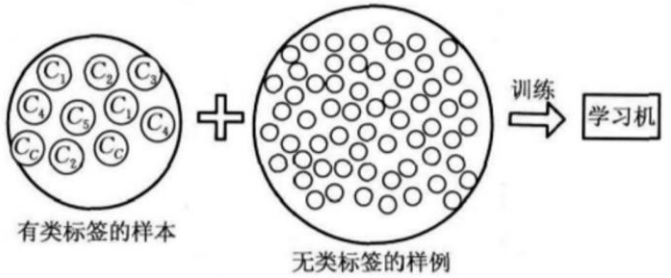
无监督学习（unsupervised learning）：**没有标注的训练数据集**，需要根据样本间的统计规律对样本集进行分析，常见任务如**聚类**等。





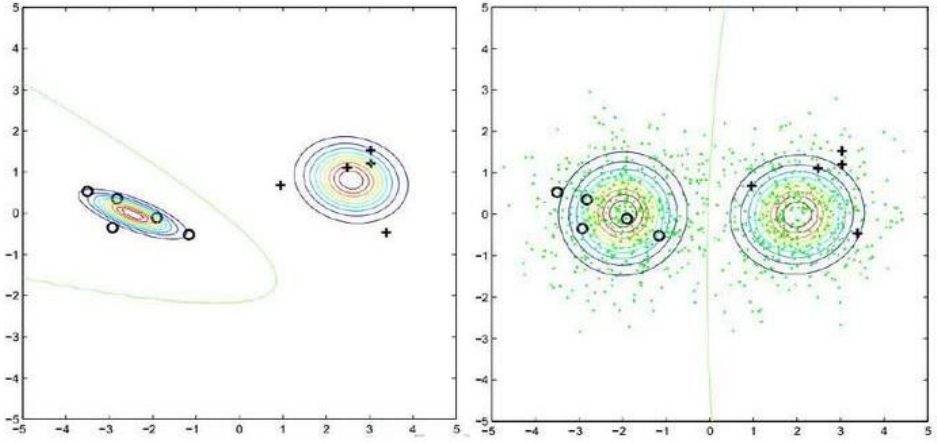
机器学习方法

半监督学习（Semi-supervised learning）：结合**（少量的）标注训练数据**和**（大量的）未标注数据**来进行数据的分类学习。



两个基本假设：

- **聚类假设**：处在相同聚类中的样本示例有较大的可能拥有相同的标记。根据该假设，决策边界就应该尽量通过数据较为系数的地方。
- **流形假设**：处于一个很小的局部区域内的示例具有相似的性质，因此，其标记也应该相似。在该假设下，大量未标记示例的作用就是让数据空间变得更加稠密，从而有助于更加准确地刻画局部特性，使得决策函数能够更好地进行数据拟合。

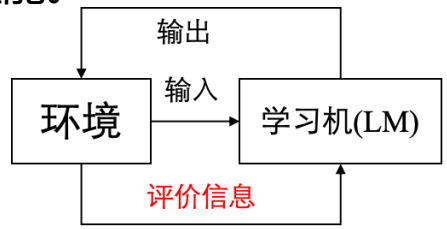


机器学习方法

增强学习（Reinforcement Learning）：外部环境对输出只给出评价信息而非正确答案，学习机通过强化受奖励的动作来改善自身的性能。

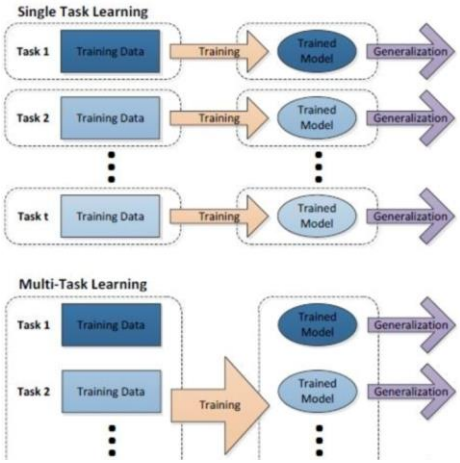
如：**让计算机学着去玩Flappy Bird**

我们不需要设置具体的策略，比如先飞到上面，再飞到下面，我们只是需要给算法定一个“小目标”！比如当计算机玩的好的时候，我们就给它一定的奖励，它玩的不好的时候，就给它一定的惩罚，在这个算法框架下，它就可以越来越好，超过人类玩家的水平。



多任务学习（Multi-task Learning）：把多个相关（related）的任务放在一起同时学习。

单任务学习时，各个任务之间的模型空间（Trained Model）是相互独立的，但现实世界中很多问题不能分解为一个一个独立的子问题，且这样忽略了问题之间所包含的丰富的关联信息。多任务学习就是为了解决这个问题而诞生的。多个任务之间共享一些因素，它们可以在学习过程中，共享它们所学到的信息，相关联的**多任务学习比单任务学习具备更好的泛化（generalization）效果。**



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

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

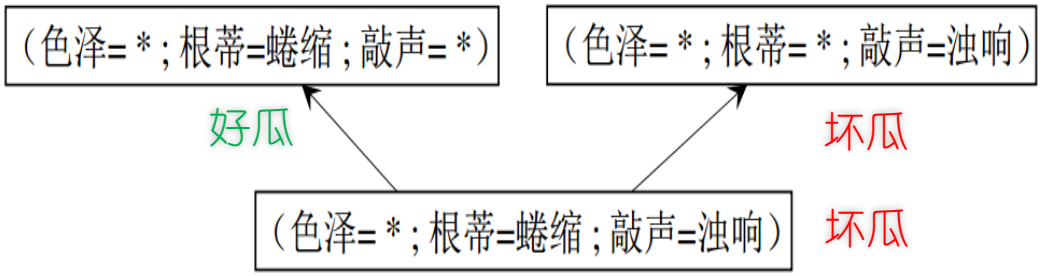
假设空间大小：4*4*4+1=65

大纲

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

归纳偏好

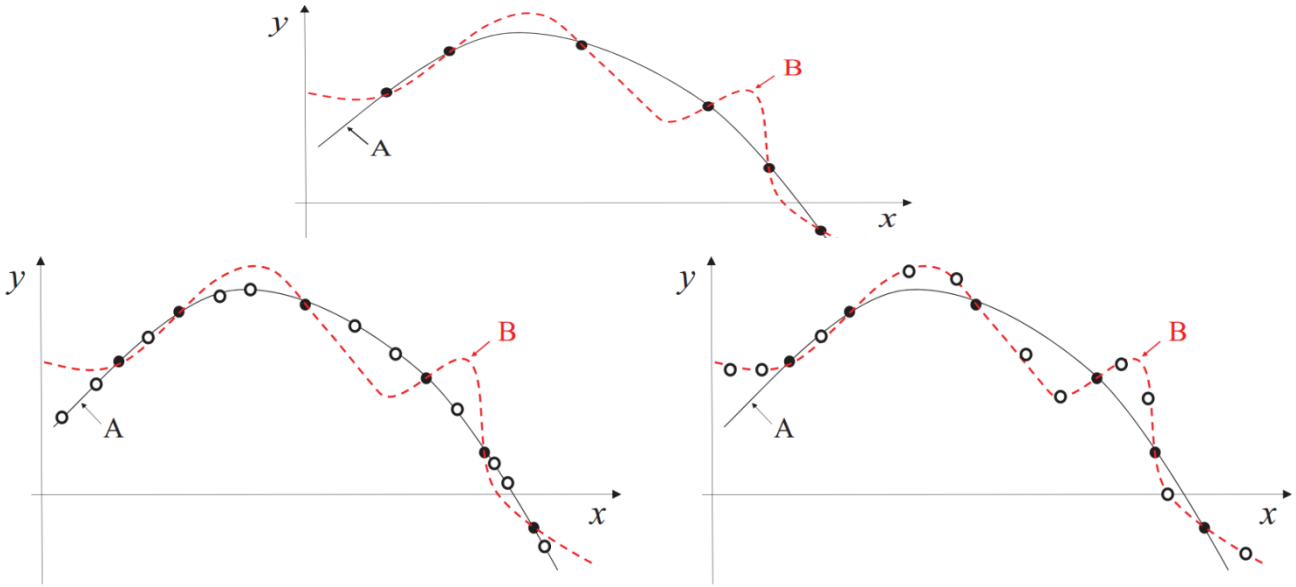
假设空间中有三个与训练集一致的假设，但他们对（色泽=青绿；根蒂=蜷缩；敲声=沉闷）的瓜会预测出不同的结果：



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 $\mathcal{H}$ 是离散的，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。

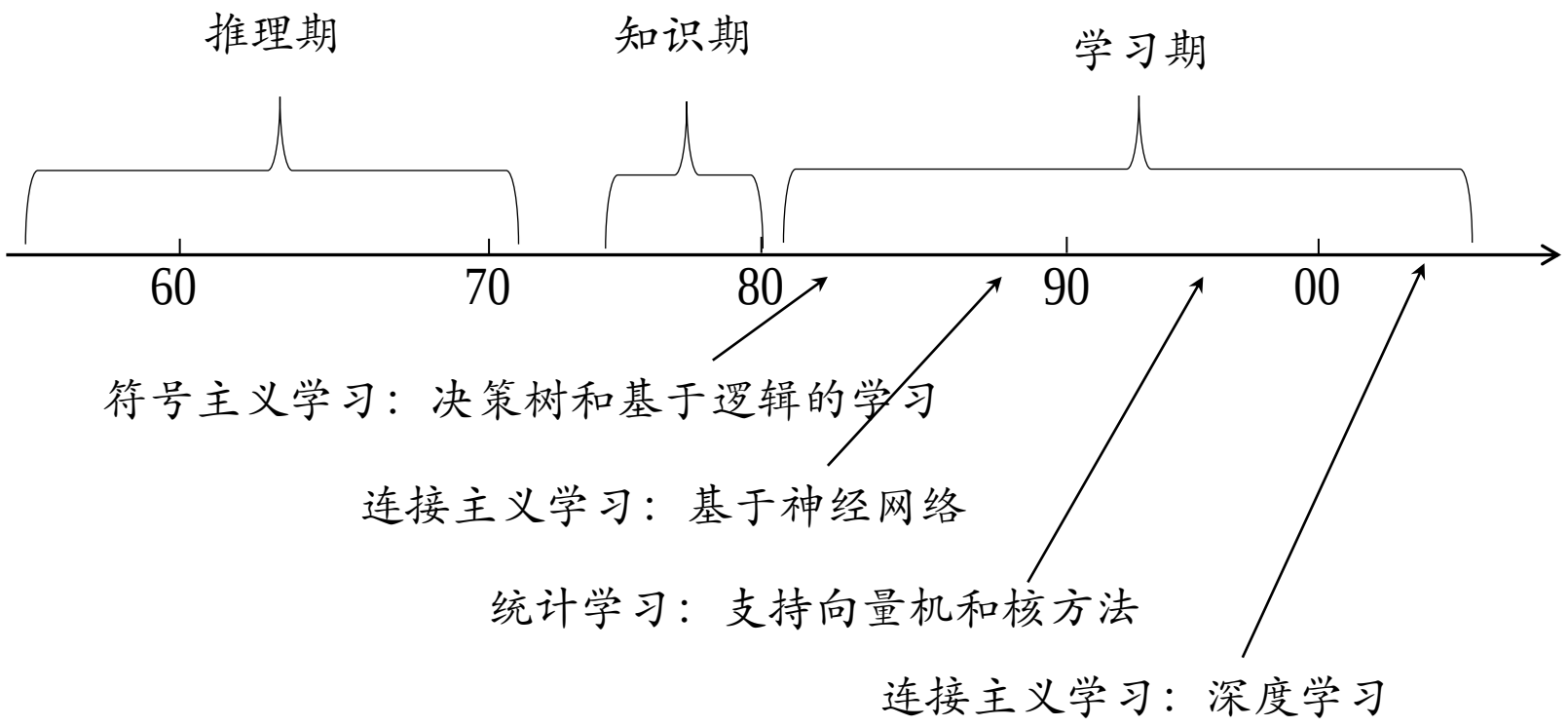
DENDRAL系统（1968年，斯坦福大学E.A.Feigenbaum等人）——推断化学分子结构的专家系统

发展历程

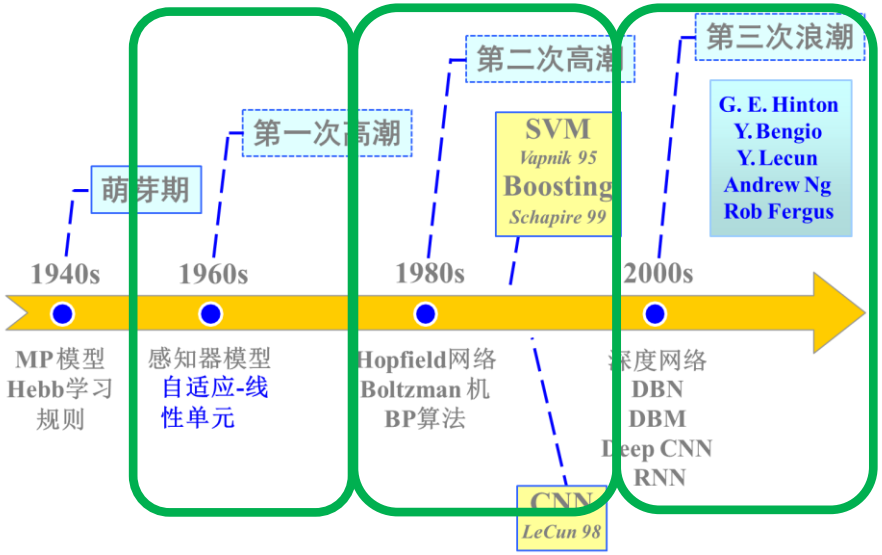
- 学习期：
  - 符号主义学习
    - 决策树：以信息论为基础，最小化信息熵，模拟了人类对概念进行判定的树形流程
    - 基于逻辑的学习：使用一阶逻辑进行知识表示，通过修改扩充逻辑表达式对数据进行归纳
  - 连接主义学习
    - 神经网络
  - 统计学习
    - 支持向量机及核方法

1952, Arthur Samuel at IBM

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

神经网络发展历程



第一次高潮

- 1962年， Frank Rosenblatt 提出感知器模型
- 1969年，M. Minsky 等人指出感知器不能解决高阶谓词问题和异或问题，将感知器拉下神坛

第二次高潮

- 人工智能对自动制导车的失败，而利用神经网络有可能解决这个问题，导致了人工神经网络的第二次高潮。
- 1986年，Rumelhart等提出多层网络的学习算法—反向传播算法

第三次高潮

- 2006年，深度学习被提出，开始使用更深的网络模型。
- 2012年，深度学习算法在语言和视觉识别上实现突破。
- 2016年，采用深度学习和强化学习的AlphaGo在围棋上战胜人类。
- 2022年，GPT等大模型在AIGC(AI Generated Content)领域实现突破。

大纲

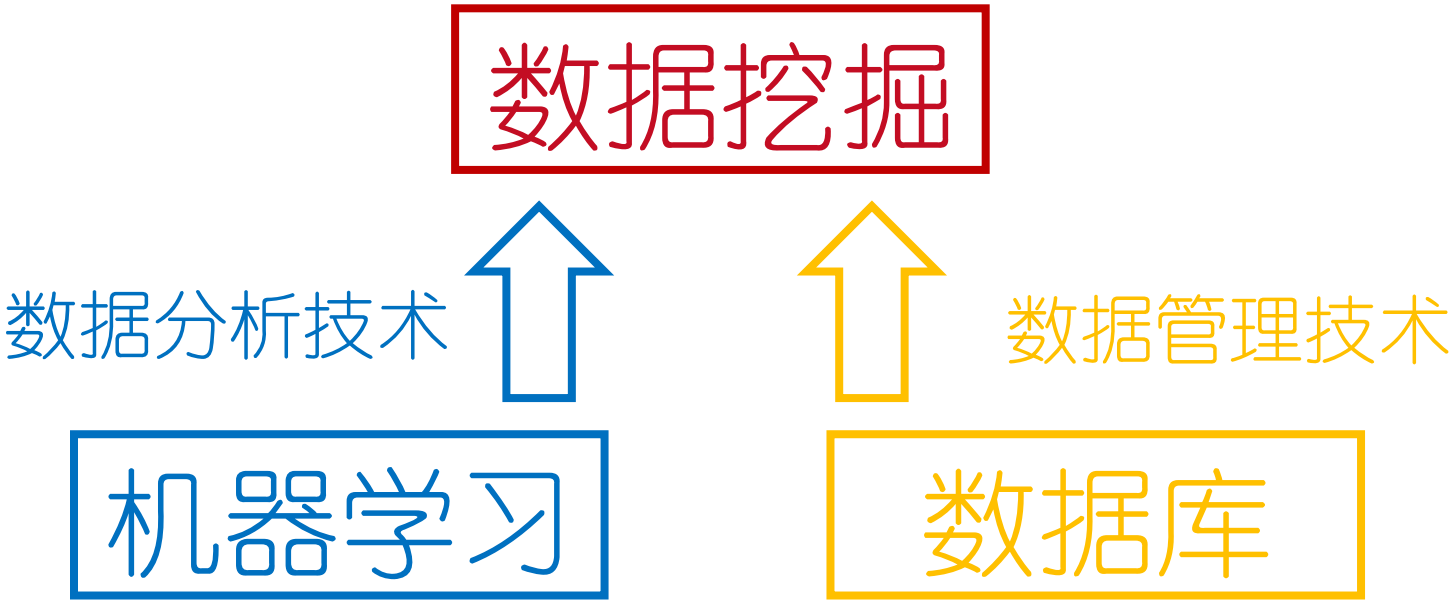
- 引言
- 基本术语
- 假设空间
- 归纳偏好
- 发展历程
- 应用现状
- 阅读材料

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。

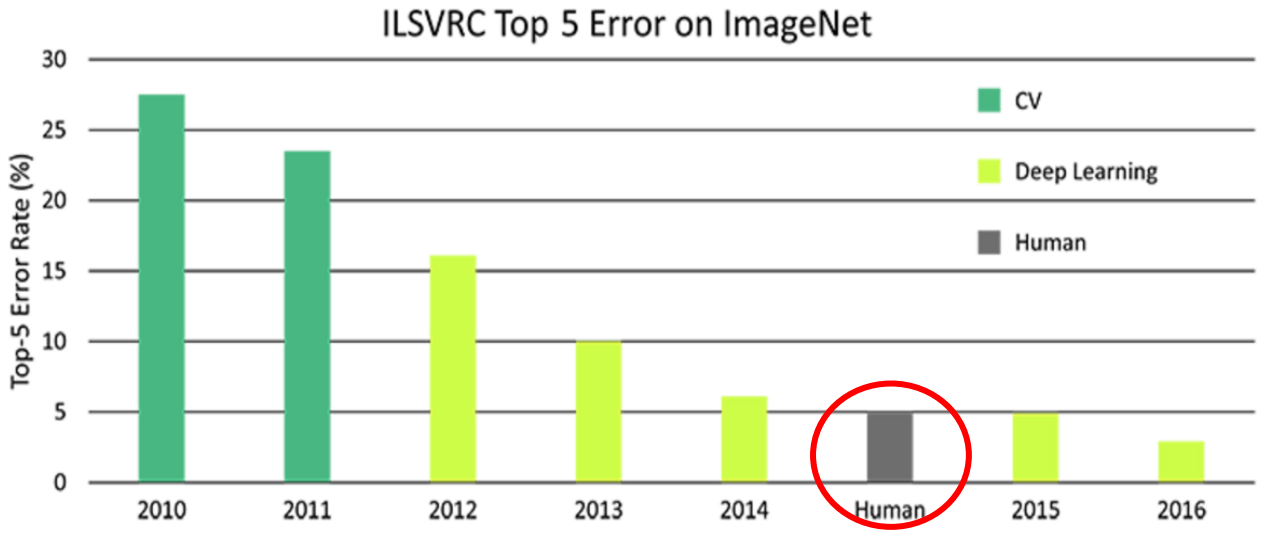
- 与普通人的生活密切相关：
  - 天气预报、能源勘探、环境监测、搜索引擎、自动驾驶汽车等

机器学习与数据挖掘

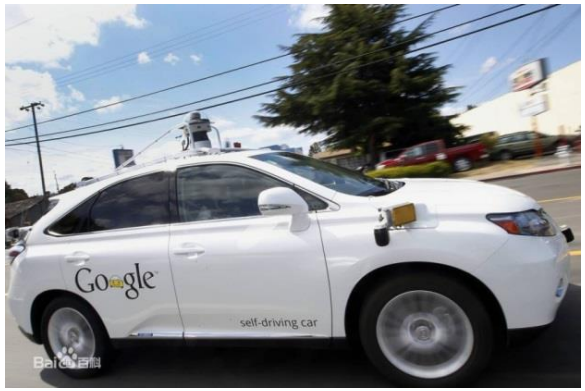


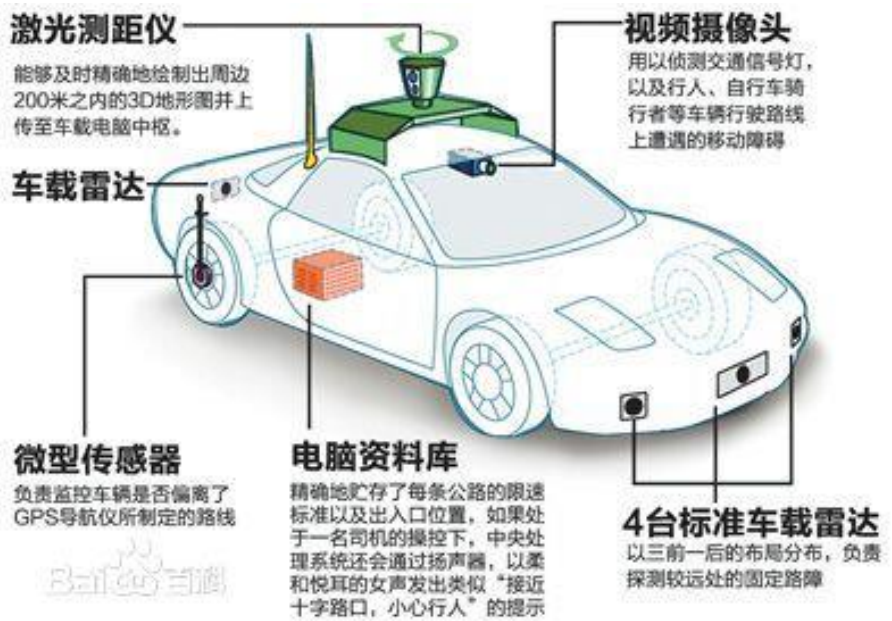
AlphaGo

- 2015年，微软亚洲研究院的ResNet，在ImageNet的分类比赛中取得了3.57%的错误率，超过了人在ImageNet数据集上对图片进行分类的成绩（5.1%）。
- 2016年初万众瞩目的围棋人机大战中，AlphaGo击败人类职业围棋选手。2017年10月又推出新一代AlphaGo Zero可以通过自我对弈来自我训练。



无人驾驶





百度发布“**Apollo（阿波罗）**”软件平台，向汽车行业及自动驾驶领域提供一套完整的平台。

无人驾驶主要包括三个环节：感知、**决策**、和控制

核心技术：异步多传感器同步+深度**数据融合**

- 我们身边的机器学习？

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.

- 国内不少书籍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).