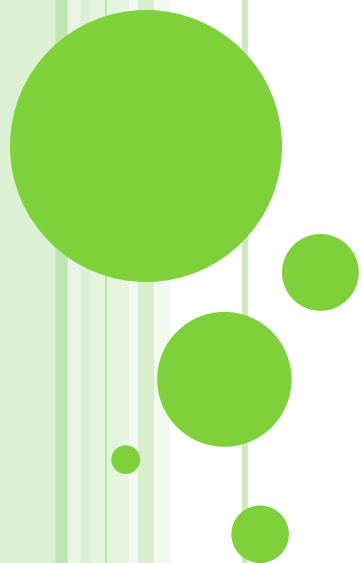


# 第一章 绪 论



# 目录

---

- 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



# 引言

## 1 什么是机器学习？

傍晚小街路面上沁出微雨后的湿润，和煦的细风吹来，抬头看看天边的晚霞，嗯，明天又是一个好天气。走到水果摊旁，挑了个根蒂蜷缩、敲起来声音浊响的青绿西瓜，一边满心期待着皮薄肉厚瓢甜的爽落感，一边愉快地想着，这学期狠下了工夫，基础概念弄得清清楚楚，算法作业也是信手拈来，这门课成绩一定差不了！

微湿路面、感到和风、看到晚霞



明天是好天

色泽青绿、根蒂蜷缩、敲声浊响



好瓜

类似的,我们从以往的学习经验知道

理清了概念、做好了作业



会取得好成绩



# 引言

## 1 什么是机器学习？

### ■ 为什么我们能做出准确的判断？

因为我们已经积累了很多经验，而利用经验，  
就能对新的情况作出有效的决策！

上面对经验的利用是靠我们人类自身完成的，  
计算机能帮忙吗？

机器学习正是这样一门学科



# 引言

## 1 什么是机器学习？

➤ 致力于通过计算手段，利用经验来改善系统自身的性能。

计算机中：

经验-----“数据”形式存储

所以，机器学习主要内容：

在计算机上从数据产生模型的算法，即学习算法

有了学习算法，把经验数据提供给他，就能基于数据产生模型，在新的情况下，模型会给我们提供相应的判断！



# 引言

## 1 什么是机器学习？

“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。

# 引言

## 1 什么是机器学习？

“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”

- 机器学习可以说是研究关于“学习算法”的学问
- 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。

# 引言

2

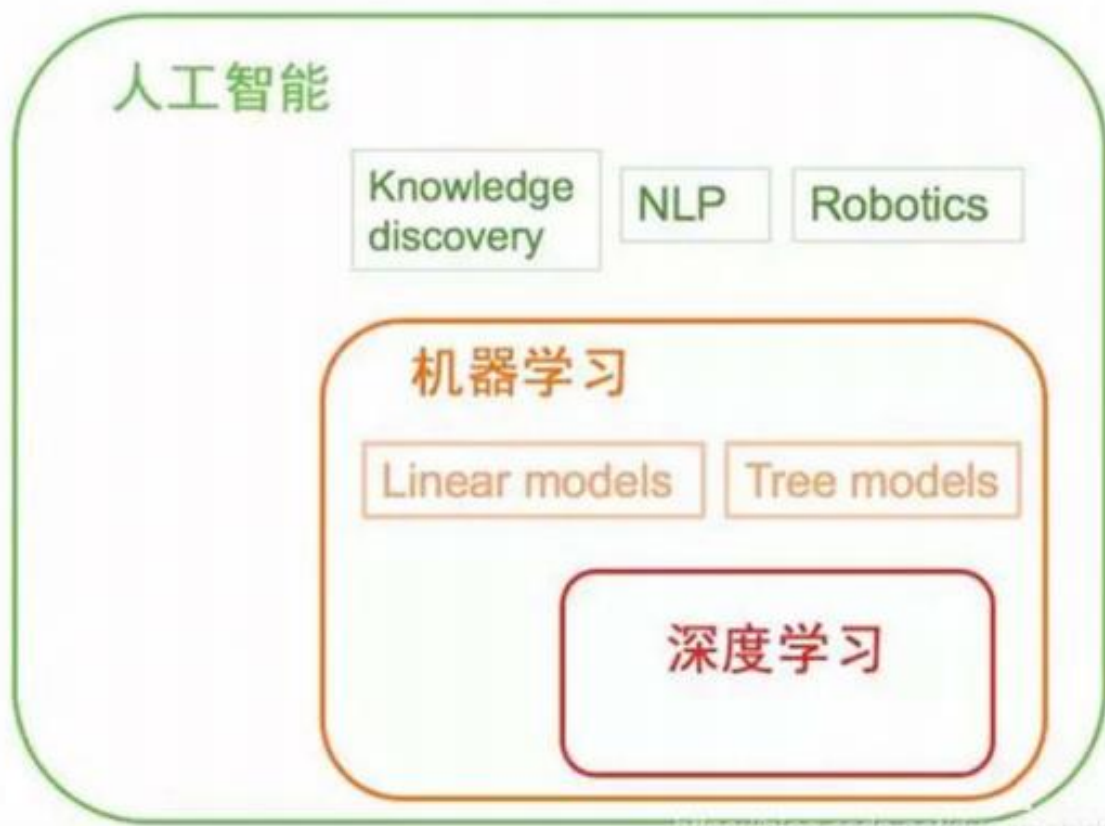
分清楚几个概念？

人工智能

机器学习

深度学习

数据挖掘





# 引言

## 2 分清楚几个概念？

### 人工智能的定义

人工智能，英文名为Artificial Intelligence，缩写为AI。人工智能和其它许多新兴学科一样，至今尚无一个统一的定义。

综合各种不同的观点，可以从“能力”和“学科”两个方面对人工智能进行定义。

➤ **从能力的角度来看**，人工智能是相对于人的自然智能而言的，

所谓人工智能是指用人工的方法在机器（计算机）上实现的智能。

➤ **从学科的角度来看**，人工智能是作为一个学科名称来使用的，

所谓人工智能是一门研究如何构造智能机器和智能系统，使它能模拟、延伸和扩展人类智能的学科。



特斯拉人形机器人“擎天柱2代”

# 引言

2

## 分清楚几个概念？

### 人工智能

- Artificial Intelligence，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的技术科学。
- 人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括语音识别、图像识别、机器人、自然语言处理、智能搜索和专家系统等。
- 人工智能可以对人的意识、思维的信息过程的模拟。人工智能不是人的智能，但能像人那样思考、也有可能超过人的智能。

# 引言

2

## 分清楚几个概念？

### 数据挖掘

- Data Mining，从海量数据中“挖掘”隐藏信息，这里的数据是“大量的、不完全的、有噪声的、模糊的、随机的实际应用数据”，信息指的是“隐含的、规律性的、人们事先未知的、但又是潜在有用的并且最终可理解的信息和知识”。在商业环境中，企业希望让存放在数据库中的数据能“说话”，支持决策。所以，数据挖掘更偏向应用。
- 数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统(依靠过去的经验法则)和模式识别等诸多方法来实现上述目标。

# 引言

2

## 分清楚几个概念？

### 机器学习

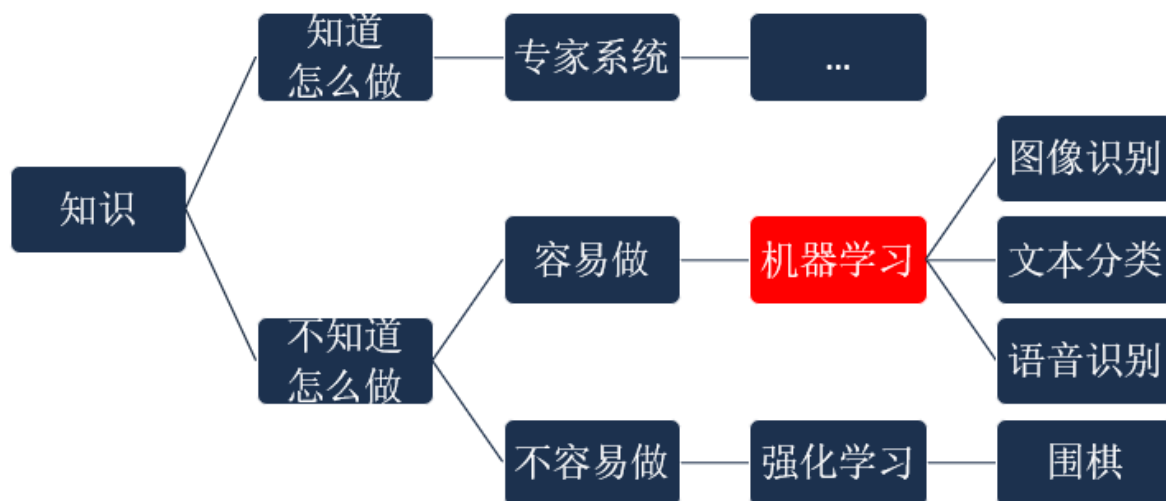
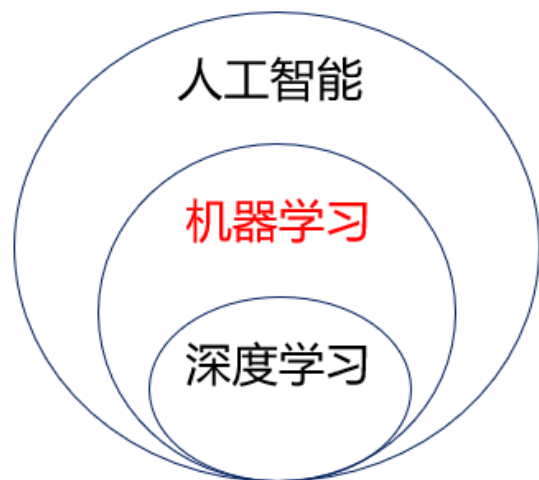
- Machine Learning是指用某些算法指导计算机利用已知数据得出适当的模型，并利用此模型对新的情境给出判断的过程。思想并不复杂，仅仅是对人类生活中学习过程的一个模拟。而在这整个过程中，最关键的是数据。
- 任何通过数据训练的学习算法的相关研究都属于机器学习，包括很多已经发展多年的技术，比如线性回归(Linear Regression)、K均值(K-means，基于原型的目标函数聚类方法)、决策树(Decision Trees，运用概率分析的一种图解法)、随机森林(Random Forest，运用概率分析的一种图解法)、PCA(Principal Component Analysis，主成分分析)、SVM(Support Vector Machine，支持向量机)以及ANN(Artificial Neural Networks，人工神经网络)。

# 引言

## 2 分清楚几个概念？

### ● 机器学习与知识获取

机器学习是一种人工智能技术。在概念学习的目标指引下，它使计算机能够在基于合理假设前提下，根据数据分布优化某种原型函数，形成与自变量相关的参数，并以参数的结构与集合作为知识。其中函数参数学习无需人工干预。



容易做是指：存在一种已知的有效算法或模型可以直接解决问题，而不需要大量的试验和错误。在这种情况下，如果能够找到一个合适的原型模型（比如一个简单的线性模型、决策树或其他监督学习模型），那么问题就可以相对容易地解决。

# 引言

2

## 分清楚几个概念？

### 深度学习

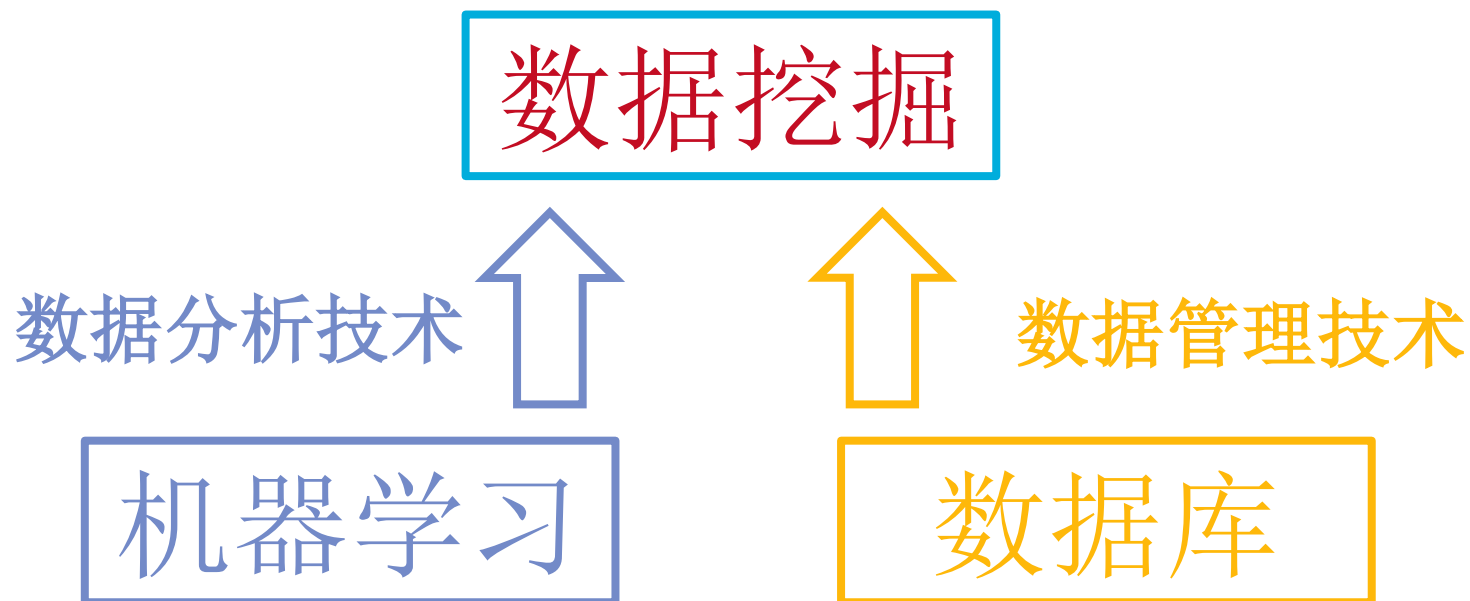
- 深度学习(Deep Learning)的概念源于人工神经网络的研究。含多隐层的多层感知器就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。
- 深度学习是机器学习研究中的一个新的领域，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据，例如图像，声音和文本。



# 引言

## 2 分清楚几个概念？

### 机器学习与数据挖掘

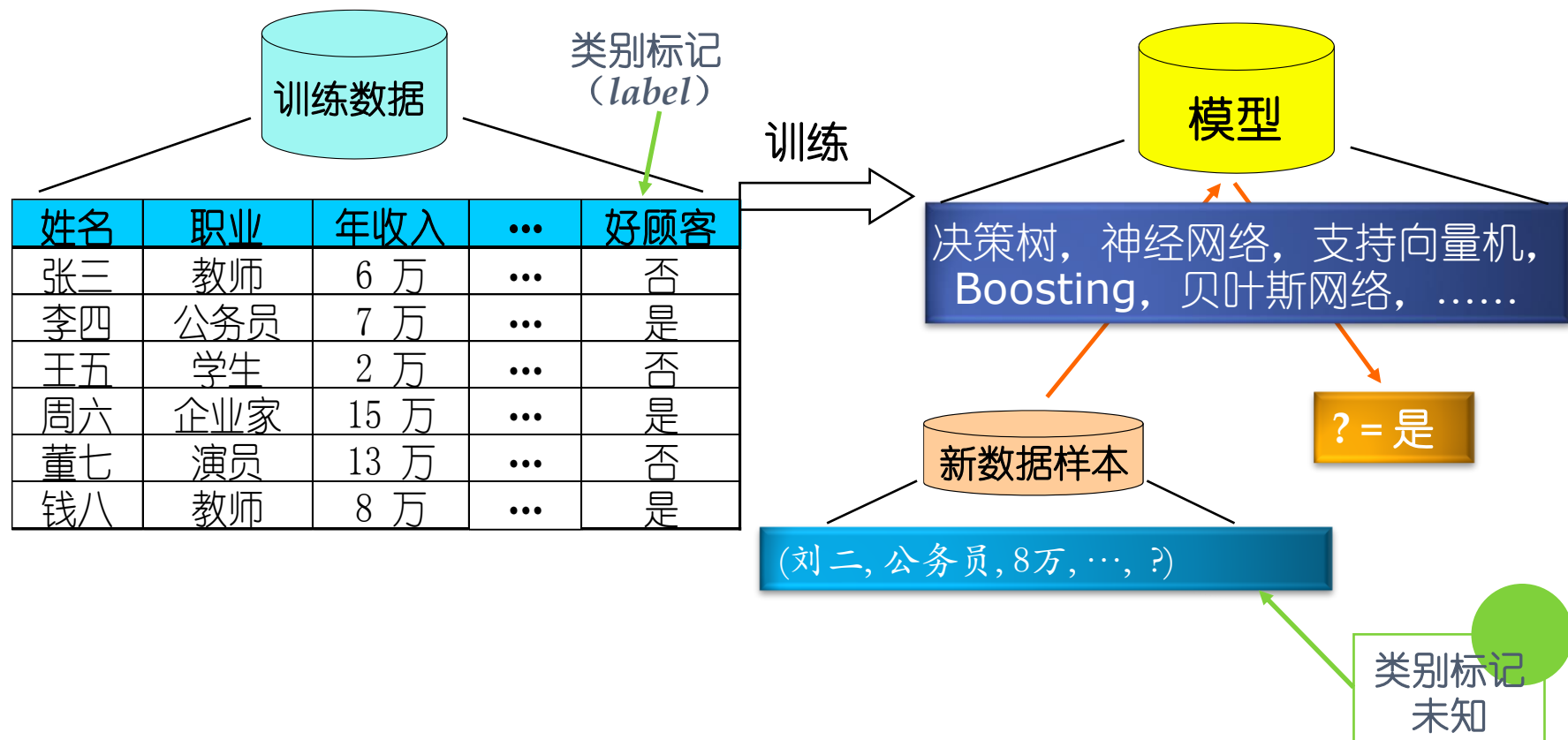


# 引言

## 3

## 典型的机器学习过程

使用学习算法 (*learning algorithm*)





# 目录

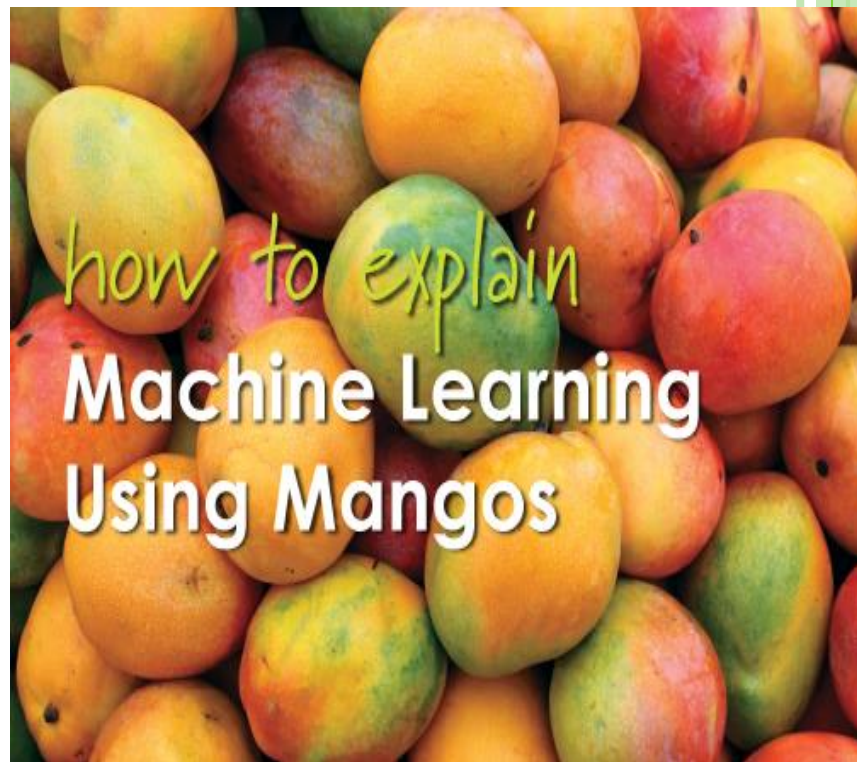
---

- 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

- 一个例子：甜芒果概念学习
  - ▶ 从市场上随机选取的芒果样本（训练数据），列出每个芒果的所有特征：
    - ▶ 如颜色，大小，形状，产地，品牌
  - ▶ 以及芒果质量（输出变量）：
    - ▶ 甜蜜，多汁，成熟度。
  - ▶ 设计一个学习算法来学习芒果的特征与输出变量之间的相关性模型。
  - ▶ 下次从市场上买芒果时，可以根据芒果（测试数据）的特征，使用前面计算的模型来预测芒果的质量。



# 基本术语—数据

在采集足够多数量的训练数据后，甜芒果概念学习可整理为一个二维表，列出所有相关的条件。

特征					标签
芒果编号	颜色	大小	形状	甜度	
训练数据 { 1	橙黄(1)	较大(0.85)	球状(1)	甜(1)	{ 甜(1)
2	淡黄(2)	一般(0.60)	椭球状(2)	不甜(0)	
3	橙黄(1)	较小(0.45)	球状(1)	甜(1)	
.....	.....	.....	.....	.....	
测试数据 { i	淡黄(2)	较大(0.75)	球状(1)	?	

**机器学习的直观解释：**甜芒果识别问题对应于一个基于训练数据的函数  $f(\cdot)$  构造及其参数学习过程，使得基于所学习的参数组合，能在测试数据完成类似任务。

# 基本术语—数据

## 西瓜数据集

		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	?

一批关于西瓜的数据

- **数据集**：记录的集合就是一个数据集，每条记录是一个示例或者样本
- **属性或者特征**：反映事件或对象在某方面的表现或性质的事项，  
一个样本的**属性**，如西瓜的色泽、根蒂、敲声
- 属性上的取值，称为**属性值**。如色泽是青绿的还是乌黑的。

# 基本术语—数据

	特征			标记
	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	?

➤ 属性张成的空间称为属性空间，或样本空间，或输入空间。

例如我们以色泽、根蒂、敲声作为西瓜的三个属性，

构造出一个三维的空间，称为属性空间、样本空间或者输入空间。

例如：坐标轴 色泽 根蒂 敲声，张成一个描述西瓜的三维空间，每个西瓜都可以在这个空间中找到自己的坐标位置。

特征向量：空间中的每个点对应一个坐标向量，

因此我们把每一个示例称为一个特征向量

# 基本术语—数据

		特征				标记
		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	?	

➤ 一般令  $D=\{x_1, x_2, \dots, x_m\}$  表示包含  $m$  个西瓜的数据集。

而每一个西瓜则有  $x_i = \{x_{i1}, x_{i2}, x_{i3}\}$  三个属性。

每个示例:  $x_i = \{x_{i1}, x_{i2}, \dots, x_{id}\}$   $\mathbf{x}_i = (x_{i1}; x_{i2}; \dots; x_{id})$

是  $d$  维样本空间中的一个向量

$d$  称为样本  $\mathbf{x}_i$  的“维数” (dimensionality).

# 基本术语—学习

从数据中学得模型的过程称为:

“学习” (learning)或 “训练” (training)

这个过程通过执行**某个学习算法**来完成

- 训练中使用的数据称为 **训练数据(training data)**
- 其中每个样本称为一个 **“训练样本” (training sample)**
- 训练样本组成的集合称为 **“训练集” (training set)**
- 学得了模型关于数据的某种潜在的规律  
**因此亦称 “假设” (hypothesis)**
- 这种潜在规律自身则称为: **“真相” 或 “真实” (ground-truth)**

**学习过程就是为了找出或逼近真相**



# 基本术语—任务

## ■ 预测目标:

### ● 分类:离散值

若我们预测的是离散值，例如“好瓜”，“坏瓜”，  
此类学习任务称为：“分类”(classification)

#### ➤ 二分类: 好瓜;坏瓜

只涉及两个类别的“二分类”(binary classification)任务，  
通常称其中一个类为“**正类**”(positive class)

另一个类为“**反类**”(negative class)

#### ➤ 多分类: 冬瓜; 南瓜; 西瓜

涉及多个类别时候，

则称为“多分类”(multi-class classification)任务



# 基本术语—任务

## ■ 预测目标:

一般的预测任务是希望通过对训练集

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$$

进行学习, 建立一个从输入空间 $\mathcal{X}$ 到 $\mathcal{Y}$ 的映射  $f: \mathcal{X} \mapsto \mathcal{Y}$ .

- 对二分类任务,通常: 令  $\mathcal{Y} = \{-1, +1\}$  或  $\{0, 1\}$ ;
- 多分类任务,  $|\mathcal{Y}| > 2$ ;
- 对回归任务,  $\mathcal{Y} = \mathbb{R}$ ,  $\mathbb{R}$  为实数集
- 学得模型后, 使用其进行预测的过程称为“测试”(testing)
- 被预测的样本称为“测试样本”(testing sample)

# 基本术语—任务

## ■ 预测目标:

### ● 回归:连续值

若欲预测的是连续值，例如西瓜成熟度0.95、0.37，  
此类学习任务称为“回归” (regression)

### ● 分类: 离散值

### ● 聚类:无标记信息

聚类 ( clustering ) ，即将训练集中的西瓜分成若干组，每组称为一个“簇” (cluster) ，这些自动形成的簇，可能对应一些潜在的概念，例如浅色瓜，深色瓜，甚至本地瓜，外地瓜

## 注意:

在聚类学习中，浅色瓜，本地瓜这样的概念我们事先是不知道的，  
而且学习过程中使用的训练样本通常不拥有标记信息

# 基本术语—任务

## □ 有无标记信息

根据训练数据**是否拥有标记信息**，学习任务可以大致划分为两大类：

“**监督学习**” (supervised learning) 和 “**无监督学习**” (unsupervised learning)

○ **监督学习**：分类、回归

监督学习的代表

○ **无监督学习**：聚类

无监督学习的代表

○ **半监督学习**：两者结合

**注意：**

机器学习的目标是**使学得的模型能很好的适用于新样本**，而不是仅仅在训练样本上工作的很好，即便对聚类这样的无监督学习任务，也希望学得的簇划分，能够适用于没在训练集中出现的样本

学得模型适应于新样本的能力称为 “**泛化**” (generalization) 能力

# 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

“独立同分布”(independent and identically distributed, 简称 *i.i.d.*)



# 目录

---

- 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

归纳（induction）与演绎（deduction）是科学推理的两大基本手段。

**归纳**：特殊到一般的泛化（generalization）过程，既从具体的事实归纳出一般性的规律；

**演绎**：从一般到“特殊化”（specialization）的过程；

“从样例中学习”显然是一个归纳过程，因此亦称为“归纳学习”（inductive learning）

归纳学习有狭义与广义之分，

广义的归纳学习大体相当于从样例中学习，

而狭义的归纳学习，则要求从训练数据中学得概念（concept），

因此也称“概念学习”或“概念形成”

概念学习中最基本的是布尔概念学习，

即对“是”“不是”这样的可表示为0/1布尔值的目标概念的学习



# 假设空间

我们学得的将是“好瓜是某种色泽，某种根蒂，某种敲声的瓜，

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

这样的概念，用布尔表达式写出来则是

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

这里的“问号？”表示尚未确定的取值，我们的任务是通过训练数据的训练把“问号？”确定下来



# 假设空间

**学习过程：** 看作一个在所有假设(hypothesis)组成的空间中进行搜索的过程；

**搜索目标：** 是找到与训练集“匹配”(fit)的假设，

即：能够将训练集中的瓜判断正确的假设。

假设的表示一旦确定，假设空间及其规模大小就确定了。

这里我们的假设空间由形如：

“(色泽=?) ∧ (根蒂=?) ∧ (敲声=?)”的可能取值所形成的假设组成。

例如：色泽有“青绿”“乌黑”“浅白”这三种可能取值；

**假设空间：** 是在已知属性和属性可能取值的情况下，

对所有可能满足目标（好瓜）的情况的一种毫无遗漏的假设集合。





# 假设空间

**例子：**假设一个瓜的好或不好，由三个属性确定。分别是色泽、根蒂、敲声。

其中，色泽有青绿、乌黑、浅白3种取值，

根蒂有蜷缩、稍蜷、硬挺3种取值，

敲声有浊响、清脆、沉闷3种取值。

那么假设空间由形如 “(色泽=? )  $\wedge$  (根蒂=? )  $\wedge$  (敲声=? )” 的所有假设组成。

除了考虑属性色泽、根蒂、敲声分别有3、3、3种可能取值，

还要考虑到一种属性可能比如色泽无论取什么值都合适（用通配符\*表示），

另外有一种情况就是好瓜这个概念根本不成立（用 $\emptyset$ 表示），则假设空间大小为

在模型空间中搜索不违背训练集的假设

假设空间大小： $3*3*4+1=37$

**版本空间：**可能有多个假设与训练集的一致，即存在着一个与训练集一致的“假设集合”，我们称之为版本空间。与训练集匹配的假设空间子集称为版本空间。

# 假设空间

**版本空间：**可能有多个假设与训练集的一致，即存在着一个与训练集一致的“假设集合”，我们称之为版本空间。**与训练集匹配的假设空间子集称为版本空间。**

提到的**匹配**：我们的假设空间写出来为，

**色泽青绿，根蒂蜷缩，敲声浊响**的瓜**是好瓜**，

把上文列举的37种可能性组合（不包含空集）去替换红色字迹部分，蓝色字迹部分不变，总共写成37+1行，这是假设空间的真面目！！！！

然而，**这些都是假设**，接下来要做的是**根据训练集去筛选这些假设**。

筛选规则书中有说：

**删除与正例不一致的假设**

**和 与反例一致的假设（即删除那些错误的假设）。**

**这个过程就是假设空间与训练集匹配。**

**最终得到的假设组成版本空间。**



# 目录

---

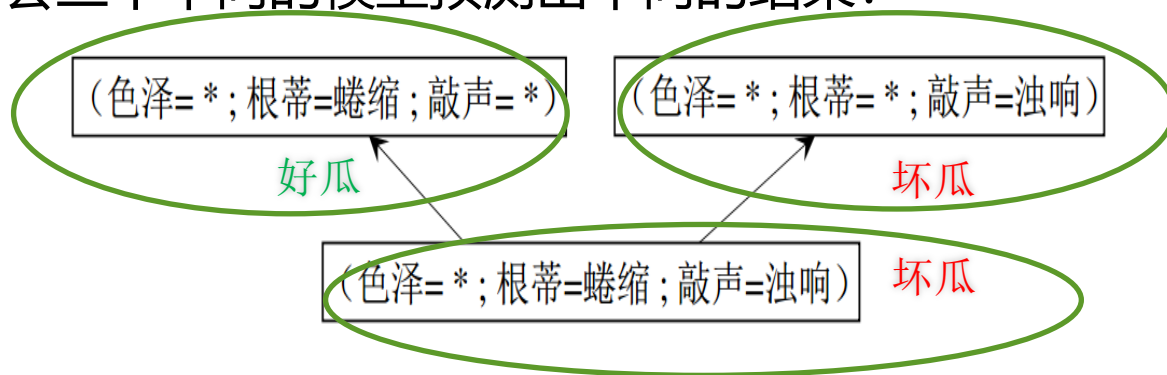
- 引言
- 基本术语
- 假设空间
- **归纳偏好**
- 发展历程
- 应用现状
- 阅读材料



# 归纳偏好

现在假设空间中：有三个与训练集一致的假设，但是与他们对应的模型产生新样本的时候有不同的输出，例如：

他们对新收来的(色泽=青绿；根蒂=蜷缩；敲声=沉闷)的瓜  
会三个不同的模型预测出不同的结果：



选取哪个假设作为学习模型？

# 归纳偏好

“归纳偏好” (inductive bias):

机器学习算法在学习过程中对某种类型假设的偏好,  
称为 “归纳偏好” (inductive bias), 或简称为“偏好”

e.g., 若算法喜欢 “特殊” 的模型,

则它会选择 “好瓜”  $\leftrightarrow (\text{色泽} = *) \wedge (\text{根蒂} = \text{蜷缩}) \wedge (\text{敲声} = \text{浊响})$ ;

而若算法喜欢 “一般” 的模型. 并且由于某种原因更 “相信” 根蒂,

则它会选择 “好瓜”  $\leftrightarrow (\text{色泽} = *) \wedge (\text{根蒂} = \text{蜷缩}) \wedge (\text{敲声} = *)$

归纳偏好

所有的机器学习算法都有其归纳偏好

一般性原则

奥卡姆剃刀 · 有多个选择的话, 选最简单的

NFL没有免费的午餐 · 总误差和学习算法无关

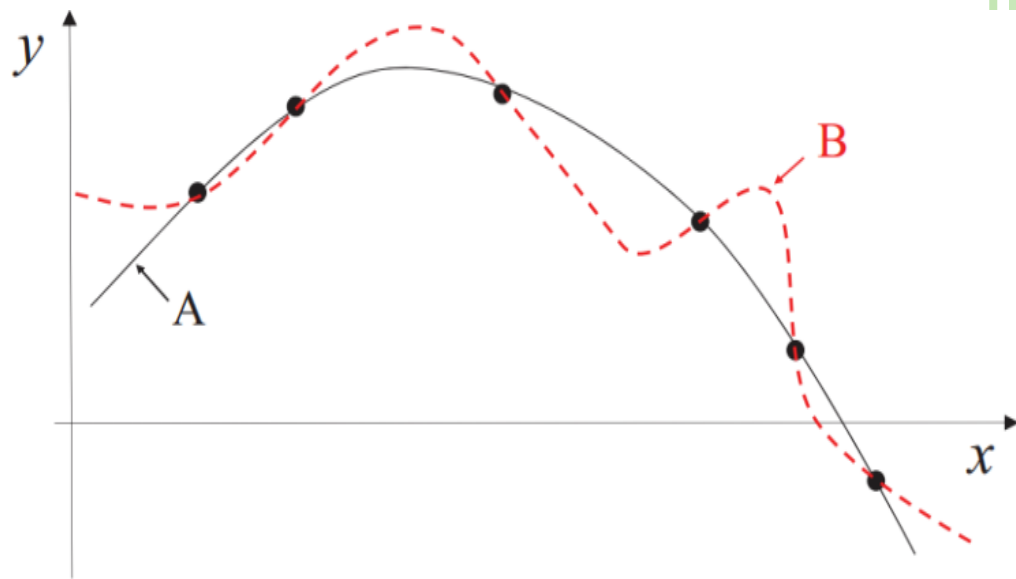
任何一个有效的机器学习算法必有其归纳偏好, 否则它将被假设空间中  
看似在训练集上 “等效” 的假设所迷惑而无法产生确定的学习结果.

# 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

如图解释：

这里的每个训练样本是图中的一个点 $(x, y)$ ，要学得一个与训练集一致的模型，相当于找到一条穿过所有训练样本点的曲线。



对有限个样本点组成的训练集，  
存在很多条曲线与其一致。

我们的学习算法必须有某种偏好，  
才能产出它认为“正确”的模型。

存在多条曲线与有限样本训练集一致

对同样一组数据集进行机器学习，然后采用样本A去进行测试，可能算法A比算法B结果好；  
那么，一定存在样本B，算法B一定比算法A好！

那你做的这个学习有啥用？空谈数据样本的话，本来就没用。

很可能现实样本A占了80%，样本B只占20%，让你在实际中选择算法A还是算法B？

# 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” (Occam's razor)**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,则选择最简单的那个”。
- 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,

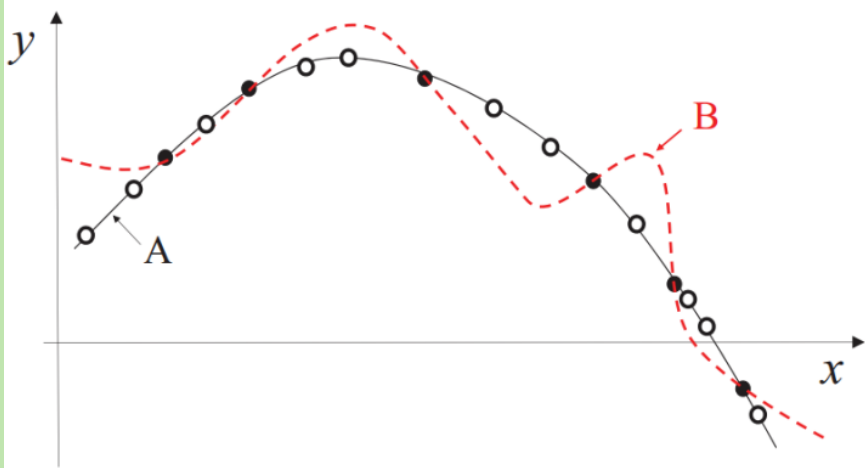
大多数时候直接决定了算法能否取得好的性能。

实际中,归纳偏好由超参数、测试-验证集的划分等影响,由Loss Function确定,且优化这个Loss Function的方法确定,归纳偏好就确定下来了。

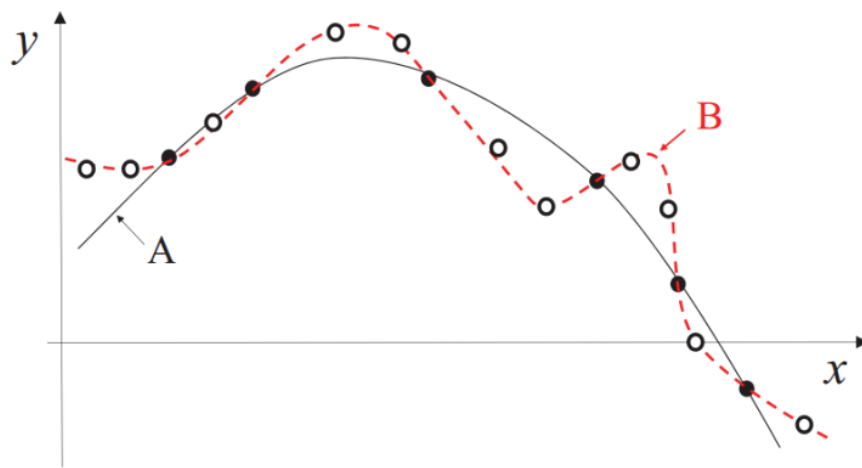
改变Loss Function,比如从Ridge Regression换成LASSO,就是在改变归纳偏好。

# 归纳偏好

一个算法 $\xi_a$ 如果在某些问题上比另一个算法 $\xi_b$ 好，必然存在另一些问题， $\xi_b$ 比 $\xi_a$ 好，也即**没有免费的午餐定理**。



(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

“没有免费的午餐” 定理(No Free Lunch Theorem, 简称NFL定理).

证明:

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, 在令  $f$  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



# 归纳偏好

“没有免费的午餐” 定理(No Free Lunch Theorem, 简称NFL定理).

证明:

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:

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

总误差与学习算法无关!

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

# 归纳偏好

总误差与学习算法无关！

对于任意两个学习算法  $\mathcal{L}_a$  和  $\mathcal{L}_b$ , 我们都有

$$\sum_f E_{ote}(\mathcal{L}_a \mid X, f) = \sum_f E_{ote}(\mathcal{L}_b \mid X, f)$$

也就是说, 无论学习算法  $\mathcal{L}_a$  多聪明、学习算法  $\mathcal{L}_b$  多笨拙

所有学习算法的期望性能都相同.

这就是 “没有免费的午餐” 定理

(No Free Lunch Theorem, 简称NFL定理).



# 归纳偏好

既然所有的学习算法期望性都和随机胡猜差不多，那有什么学的？？？

然而, NFL定理有一个重要前提:

所有“问题”出现的机会相同、或所有问题同等重要。

但实际情形并不是这样。并非所有问题出现的可能性都相同。

很多时候, 我们只关注自己正在试图解决的问题(例如某个具体应用任务), 希望为它找到一个解决方案, 至于这个解决方案在别的问题、甚至在相似的问题上是否为好方案, 我们并不关心。



# 归纳偏好

很多时候, 我们只关注自己正在试图解决的问题(例如某个具体应用任务), 希望为它找到一个解决方案, 至于这个解决方案在别的问题、甚至在相似的问题上是否为好方案, 我们并不关心. 例如:

例如, 为了快速从A地到达B地,

如果考虑的是A是南京鼓楼, B是南京新街口, 那么“骑自行车”是很好的解决方案;

如果A是南京鼓楼, B是北京新街口的情形, 这个方案显然很糟糕,

但我们对此并不关心

所以, NFL定理是让我们清楚地认识到,

脱离具体问题, 空泛地谈论 “什么学习算法更好” 毫无意义.

要谈论算法的相对优劣, 必须要针对具体的学习问题.



# 目录

---

- 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

## □ 学习期：符号主义学习

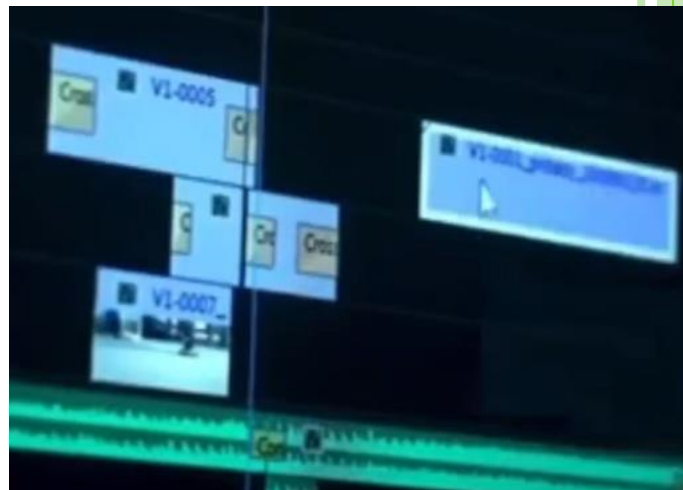
### 符号主义学派

符号主义（Symbolicism），又称逻辑主义（Logicism）、心理学派（Psychologism）或计算机学派（Computerism），是基于**物理符号系统假设和有限合理性原理**的人工智能学派。

**符号主义认为人工智能起源于物理逻辑**，人类认知（智能）的基本元素是符号（Symbol），认知过程是符号表示上的一种运算。

代表人物有：约翰·麦卡锡（John McCarthy）、艾伦·纽厄尔（Allen Newell）等人。

代表性成果：有打败国际象棋世界冠军卡斯帕罗夫的IBM超级计算机——“深蓝”。



“深蓝”打败国际象棋世界冠军卡斯帕罗夫

# 发展历程

## □ 学习期：连接主义学习 神经网络

### 联结主义学派

联结主义（Connectionism），又称仿生学派（Bionicsism）或生理学派（Physiologism），是基于神经网络及网络间的联结机制与学习算法的人工智能学派。

联结主义认为人工智能起源于仿生学，把人的智能归结为人脑的高层活动，强调智能的产生是由大量简单的单元通过复杂的相互联结和并行运行的结果。

代表人物有：麦克洛奇（W. McCulloch）、皮兹（W. Pitts）等人。

代表性成果有：战胜世界围棋高手柯洁的AlphaGo。



AlphaGo战胜世界  
围棋高手柯洁

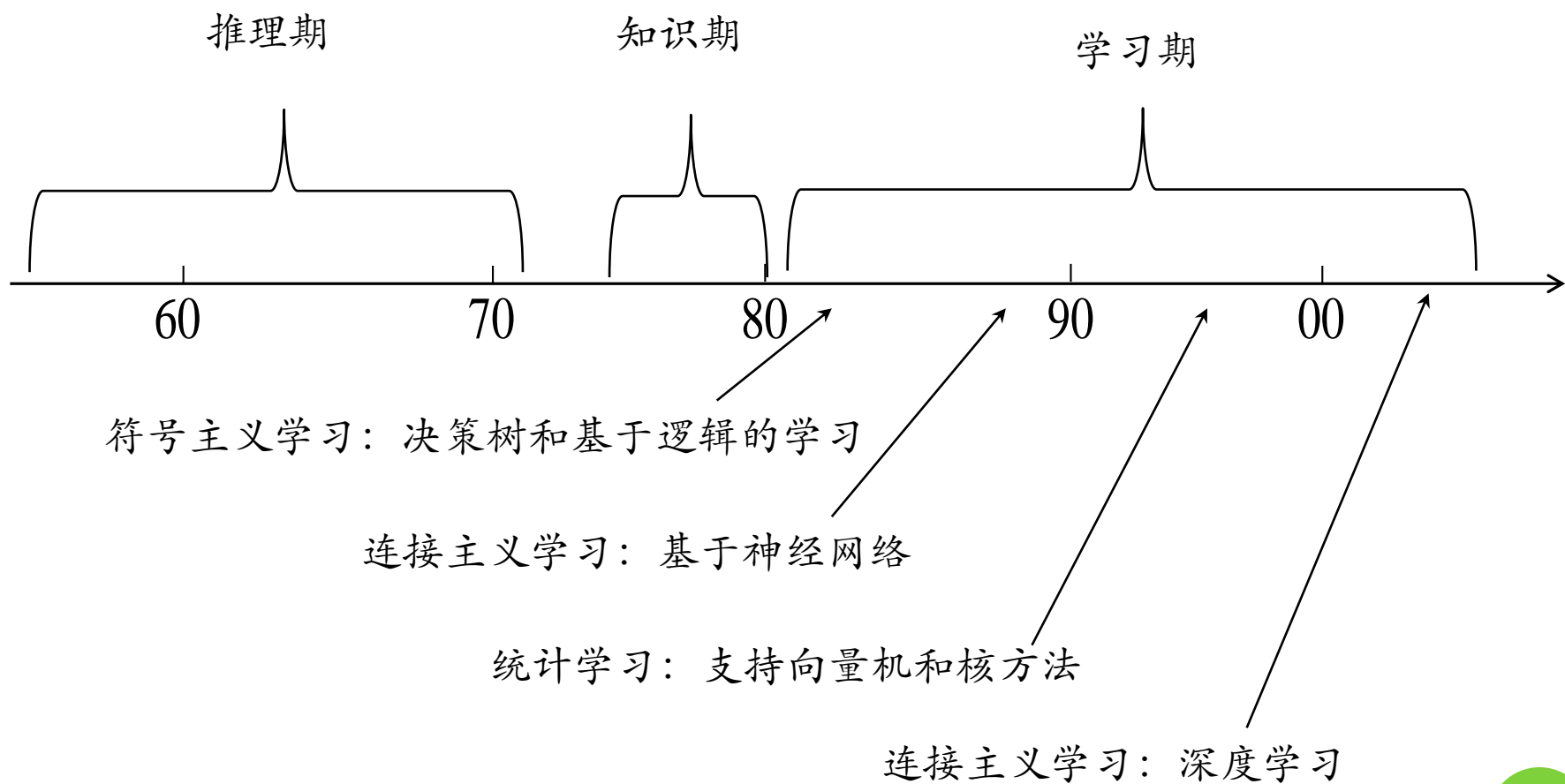


# 发展历程

## □ 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



# 目录

---

- 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。

2006年9月，Facebook全面开放，用户数量爆炸式增长，在年底达到1200万。这一过程恰好有利地推升了奥巴马的知名度。此后，在克里斯的辅佐下，奥巴马掀起了一系列的网络活动，在Facebook、MySpace等社交网站上发表公开演讲、推广施政理念，赢得大量网民支持，募集到5亿多美元的竞选经费。

最终，“黑人平民”战胜了实力雄厚的对手，成为美国历史上第一位黑人总统，之后，在第二次的选举中更获得连任。此次选举被认为是美国民主的巨大进步，而互联网则提供了前所未有的实施手段，其中尤以Facebook代表的社交网站最为突出，以至于有人戏称之为“Facebook之选”



# 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”





# 目录

---

- 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 Machine Learning Research 和 Machine Learning.
- ❑ 国内不少书记包含机器学习方面的内容, 例如 [陆汝钊, 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).

# 第一章 绪论

## what | 什么是

定义：机器学习是用数据或以往的经验，以此优化计算机程序的性能标准

示例/样本：一行数据

属性/特征：描述某一性质，就是列名

属性值：属性的取值

属性空间/样本空间：所有属性构成的空间

特征向量：一个示例

维数：属性个数

标记：示例结果的信息，即要求解的y

术语：假设：模型对应了数据的某种规律

训练集：训练模型使用的数据叫训练集

测试集：用训练集训练处模型后，被预测的样本叫测试集/测试样本

分类：学习任务是预测离散值

回归：学习任务是预测连续值

监督学习：有标记信息

无监督学习：无标记信息

泛化能力：最终得到的模型适用于新样本的能力

## how good | 如何得到模型

假设空间：和训练集一致的假设集合

归纳偏好：所有的机器学习算法都有其归纳偏好

一般性原则：奥卡姆剃刀：有多个选择的话，选最简单的  
NFL没有免费的午餐：总误差和学习算法无关

## how | 怎么用

应用场景

数据挖掘

自然语言处理NLP

计算机视觉CV

机器人决策

发展历程