MACHINE LEARNING

机器学习

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

机器学习—引出

傍晚小街路面上沁出微雨后的湿润，和熙的细风吹来，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!

发现这里涉及很多基于经验做出的预判:

- 为什么看到微温路面、感到和风、看到晚霞，就认为明天是好天呢?
  因为，我们在生活中已经遇见过很多类似情况，头一天观察到”微温路面、感到和风、看到晚霞”，第二天天气通常会很好.
- 为什么色泽青绿、根蒂蜷缩、敲声浊晌，就能判断出是正熟的好瓜?
  因为我们吃过、看过很多西瓜，所以基于色泽、根蒂、敲声这几个特征我们就可以做出相当好的判断。
- 小明从以往学习经验知道，下足了工夫、弄清了概念、做好了作业，自然会取得好成绩.

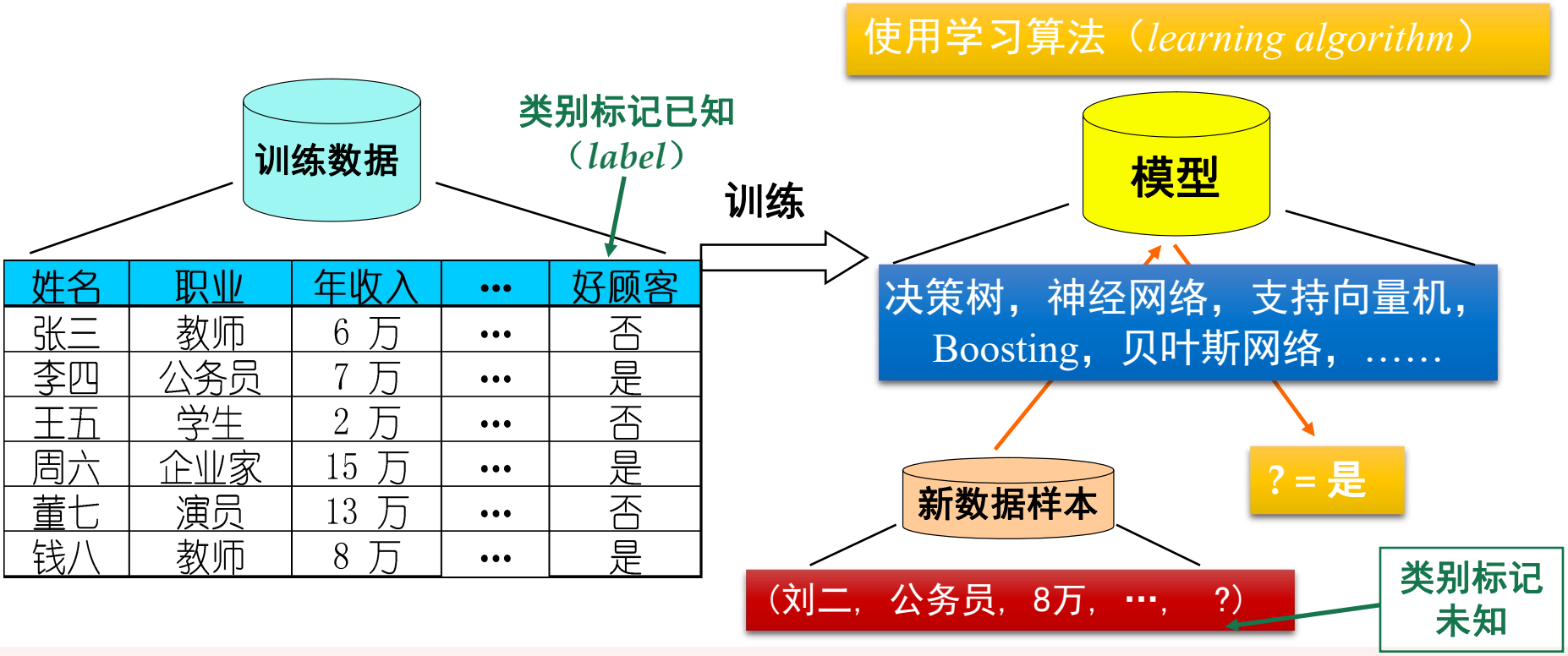
➢ 小明为什么能做出有效的预判?因为，小明通过对已积累经验的利用，对新情况做出有效的决策.

机器学习—定义、过程

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”，并使用该模型对新的情况给出判断。

模型，泛指从数据中学得的结果。



大纲

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

基本术语—数据

- 要进行机器学习，先要有数据。假定已收集了一批关于西瓜的数据：

特征 / 属性: 反映事件或对象在某方面的表现或性质的事项。
属性值: 属性上的取值，如青绿

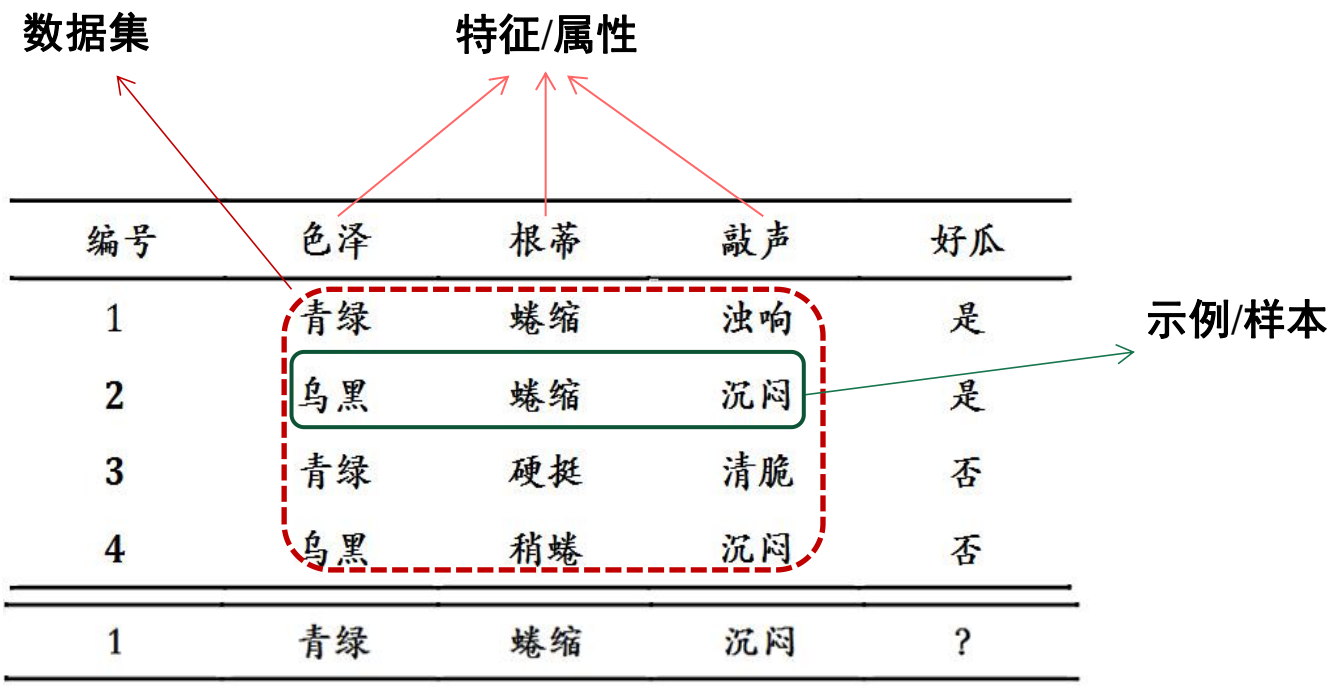
数据集: 记录的集合

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	?

示例 / 样本：每条记录是关于一个事件/对象（西瓜）的描述

- **属性空间 / 样本空间 / 输入空间**: 属性张成的空间.
  - ✓ 如把“色泽” “根蒂” “敲声”作为三个坐标轴，张成一个用于描述西瓜的三维空间，每个西瓜都可在这个空间中找到自己的坐标位置.
  - ✓ 由于空间中的每个点对应一个坐标向量，因此一个**样本**称为一个**"特征向量"**

基本术语——数据



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	?

- 样本空间 $\mathbf{X}=\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ 是包含 m=4 个示例的数据集
- 每个 示例有 d=3 个属性

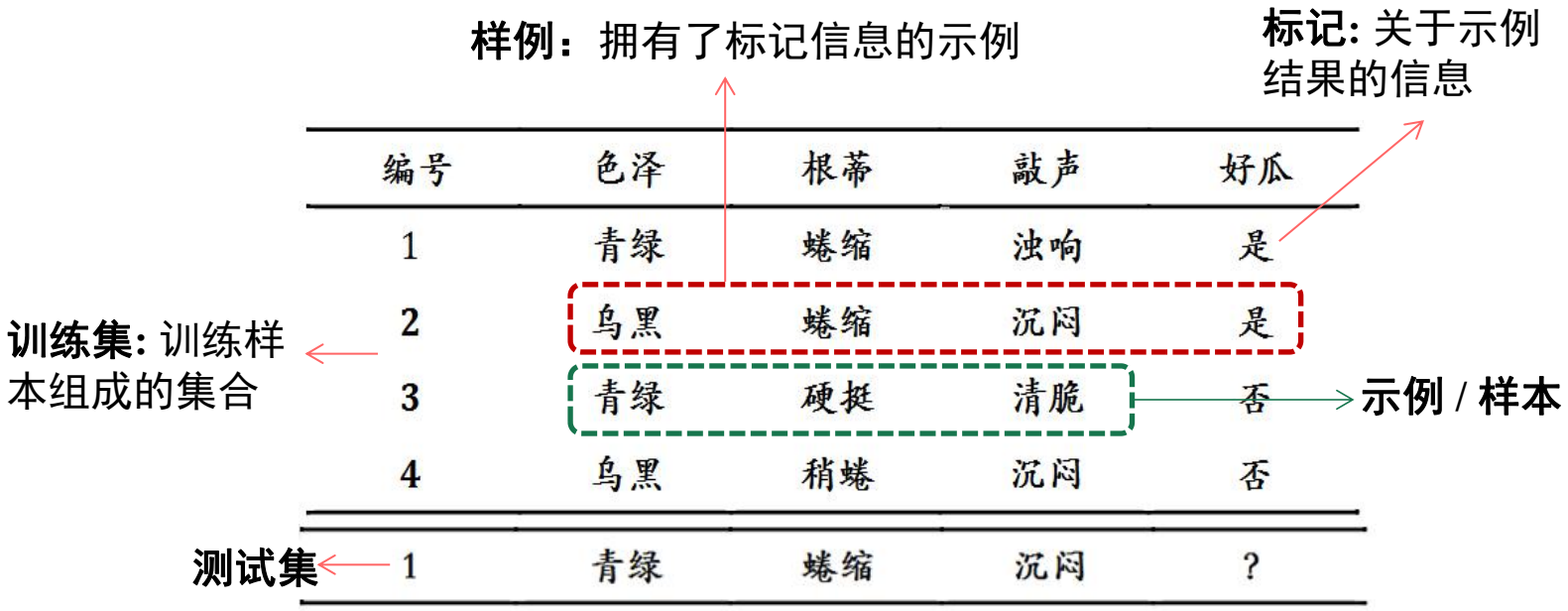
每个 示例 $\boldsymbol{x}_i$ $(x_{i,1}, x_{i,2}, \ldots x_{i,d})$ 是 d=3 维样本空间 $\mathbf{X}$ 中的一个向量， $\boldsymbol{x}_i \in \mathbf{X}$

其中，$x_{i,j}$ 是 $\boldsymbol{x}_i$ 在第 j 个属性上的取值(e.g., 第3个西瓜在第2个属性上的值是"硬挺")

d 称为样本$\boldsymbol{x}_i$的"维数" (dimensionality).

基本术语—学习

- 学习 / 训练，即从数据中学得模型的过程，这个过程通过执行某个学习算法来完成
- 下面给出了学习中相关的概念：



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
测试集 1	青绿	蜷缩	沉闷	?

- (x_i, y_i) 表示第 i 个样例，其中 y_i 是示例 x_i 的标记，Y 是所有标记的集合，亦称“标记空间 或 输出空间”。

基本术语—学习

- 学习 / 训练，即从数据中学得模型的过程，这个过程通过执行某个学习算法来完成

- 通过建立“预测” 模型，可以获得训练样本的“结果”信息；

  例如 (色泽=青绿；根蒂=蜷缩；敲声=浊响) 对应了’好瓜’的标记信息

- 通过学习所获得的模型，对应了数据的某种潜在的规律，因此所获得的模型亦称**“假设”** ；

- 这种潜在的自身规律，则称为**“真相 / 真实“** ，学习过程就是为了找出或逼近真相。

基本术语—任务

- 根据预测目标的不同，学习任务可以分为 3 类：
  - **分类**：预测值的是离散值
    - 二分类: 正类（好瓜）；反类（坏瓜）
    - 多分类: 冬瓜;南瓜;西瓜
  - **回归**：预测值的是连续值

    e.g.西瓜成熟度0.95 、0.37

- 分类和回归的预测任务是希望通过对训练集$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_m, y_m)\}$进行学习，得到一个模型，即建立一个从输入空间**X** 到输出空间**Y**的映射 f：**X->Y.**
- 使用学得的模型（或映射函数f），对测试样本进行测试，可得到测试样本的预测标记 $y = f(\mathrm{x})$。

基本术语—任务

- 根据预测目标的不同，学习任务可以分为3类：
  - **分类**：预测值的是离散值
    - 二分类: 正类（好瓜）；反类（坏瓜）
    - 多分类: 冬瓜;南瓜;西瓜
  - **回归**：预测值的是连续值

    e.g.西瓜成熟度0.95 、0.37
  - **聚类**：无预测值，即没有标记信息
    - 聚类任务，即将训练集中的样本（如西瓜）分成若干组（称为 簇）。
    - 这些自动形成的簇可能对应一些潜在的概念划分，例如“浅色瓜” “深色瓜”

      需说明的是，在聚类学习中，“浅色瓜” “深色瓜”这样的概念，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.
    - 这样的学习过程，有助于我们了解数据内在的规律，利于更深入地分析数据.

基本术语—任务

- 根据预测目标的不同，学习任务可以分为3类：
  - **分类**：预测值的是离散值
  - **回归**：预测值的是连续值
  - **聚类**：无预测值，即没有标记信息

- 根据训练数据是否拥有标记信息，学习任务又可以分为3类：
  - **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：**使得学到的模型

  既能很好地适用于“新样本”，**又能**很好地适用于训练集；

  e.g. 聚类（无监督学习）：希望学得的簇划分，能适用于没在训练集中出现的样本。

- **简化学习**

  通常，假设训练集“**独立同分布**” $(i.i.d)$，即样本空间中的每个训练样本服从一个未知分布D，且每个训练样本都从这个分布中独立采样。

基本术语—泛化能力

- **机器学习的目标：**使得学到的模型

  既能很好地适用于“新样本”，**又能**很好地适用于训练集；

- **泛化 (generalization) 能力：**模型适用于新样本的能力。
  - ✓ 一般，训练样本越多，越有可能通过学习获得强泛化能力的模型；
  - ✓ 具有强泛化能力的模型，能很好地适用于整个样本空间；
  - ✓ 训练集通常只是样本空间上的一个很小的采样，对应训练误差；
  - ✓ 希望训练集上学得的模型，能很好地反映出样本空间的特性，在整个样本空间上都工作得很好，从而泛化能力比较强。

大纲

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

假设空间---获得手段：归纳与演绎

- 获得假设（模型）空间，通常使用两大基本手段：
  - 归纳
    - 从特殊到一般的"泛化"过程，即从具体的事实归结出一般性规律；

      e. g. "从样例中学习"显然是一个归纳的过程，因此亦称"归纳学习"

      1. 广义的归纳学习：从样例中学习
      2. 狭义的归纳学习：从训练数据中学得概念，称"概念学习"

         要学得泛化性能好且语义明确的概念很困难，因此概念学习技术的研究、应用都比较少。
  - 演绎
    - 从一般到特殊的"特化"过程，即从基础原理推演出具体状况.

      e. g. 在数学公理系统中，基于一组公理和推理规则推导出与之相洽的定理，或使用公理解决实际例题。

假设空间---学习目标

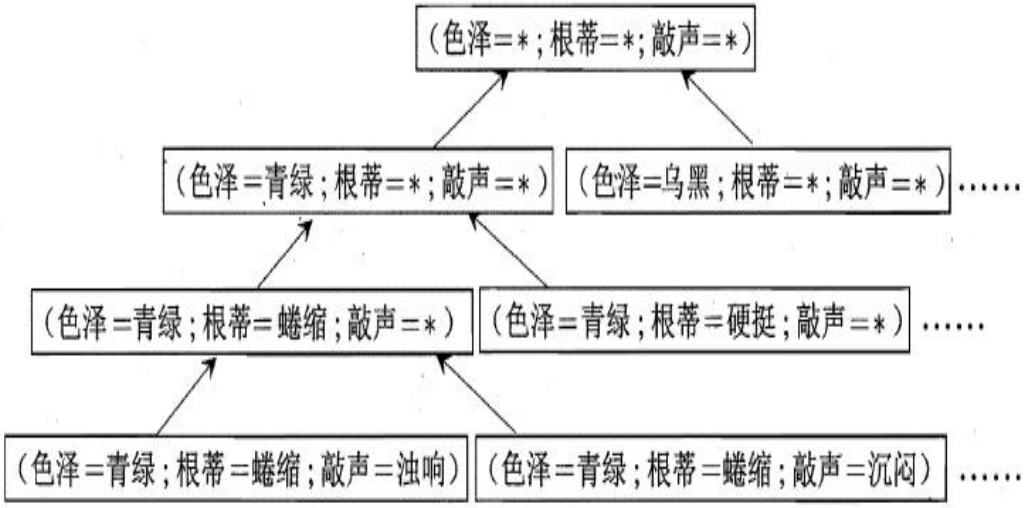
编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

- 学习的目标是“好瓜”，“好瓜”由“色泽”、“根蒂”、"敲声"三个属性完全确定
- 我们学得的将是“好瓜是某种色泽、某种根蒂、某种敲声的瓜”这样的一个概念. **布尔表达式：(色泽=?)∧(根蒂=?)∧(敲声=?)↔好瓜**
- **学习目标**
  - ✓ 通过对表中训练集进行学习，把“?”确定下来
  - ✓ 即通过对训练集中瓜的学习，以获得对没见过的瓜进行判断的能力

假设空间---学习过程

- 学习过程：在所有假设（模型）组成的空间中进行搜索的过程
- 搜索目标：找到与训练集“匹配”的假设，即能够将训练集中的瓜判断正确的假设(模型).
- 假设的表示一旦确定，假设空间及其规模大小就确定了.

  色泽, 根蒂, 敲声分别有3, 3, 3种取值，假设空间大小（3+1)*(3+1)*(3+1)+1=65



西瓜问题的假设空间

假设空间---学习过程

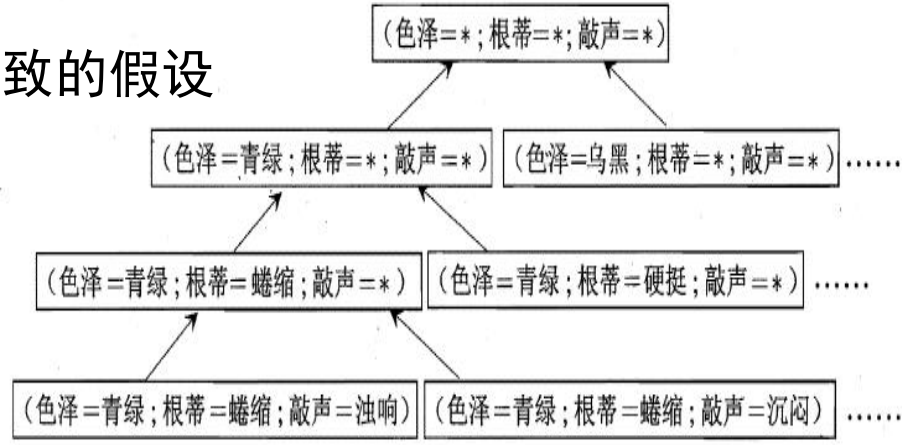
- 学习过程：在所有假设（模型）组成的空间中进行搜索的过程
- 搜索目标：找到与训练集“匹配”的假设，即能够将训练集中的瓜判断正确的假设(模型).

- 搜索策略：

  自顶向下---从一般到特殊，自底向上---从特殊到一般
- 搜索过程：

  删除与正例不一致的假设、与反例一致的假设
- 目标：

  获得与训练集一致的假设（模型），即，对所有训练样本都能正确判断



西瓜问题的假设空间

大纲

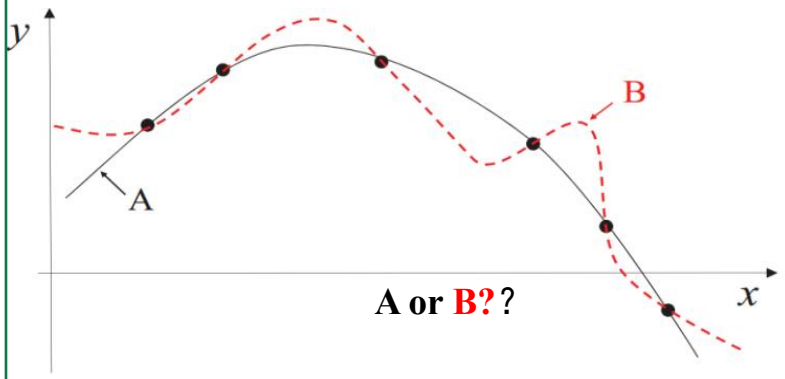
- 引言
- 基本术语
- 假设空间
- **归纳偏好**
- 发展历程
- 应用现状

归纳偏好

- 每个训练样本对应图中的一个点(x,y)
- **目标**：学得一个与训练集一致的模型，即 找到一条穿过所有训练样本点的曲线
- 对于训练集，存在着很多条曲线与其一致，e.g. **A** 黑色，**B** 红色
- **A or B ?**

  学习算法必须有某种**偏好**，才能产出它认为“正确”的模型（假设）.

- 若学习算法认为：相似的样本应有相似的输出
  - e.g. 各种属性上都很相像的西瓜，成熟程度应该比较接近
- 学习算法可能**偏好** 比较“平滑”的曲线**A**，而**不是**比较"崎岖"的曲线B
- 与**B** 相比，A与训练集外的样本更一致；换言之，A 的泛化能力比B 强.



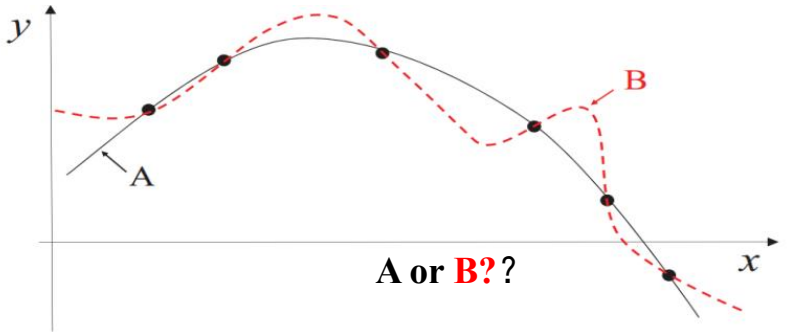
存在多条曲线与有限样本训练集一致

归纳偏好

- **目标**：学得一个与训练集一致的模型，即找到一条穿过所有训练样本点的曲线
- 学习算法必须有某种**偏好**，才能产出它认为“正确”的模型（假设）.

- 如果**没有偏好**，
  - ✓ 西瓜学习算法产生的2个假设（模型）e.g. 曲线 A、B
  - ✓ 每次对新瓜(色泽=青绿;根蒂=蜷缩;敲声=沉闷)进行**预测**时，随机抽选假设（e.g. 曲线 A、B）
  - ✓ 模型，时而预测是好瓜、时而预测是坏瓜，这样的学习结果显然没有意义

选取哪个假设作为学习模型？



存在多条曲线与有限样本训练集一致

- **归纳偏好**：学习过程中，对某种类型假设（e.g. 曲线 A、B）的偏好

归纳偏好——奥卡姆剃刀

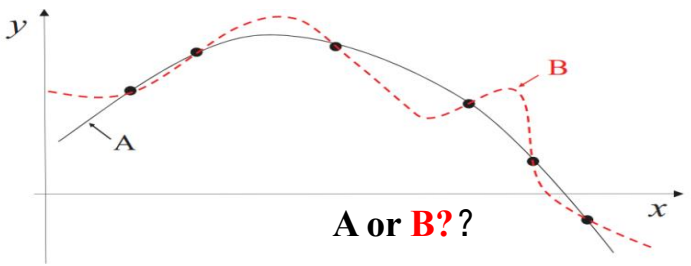
- 归纳偏好可看作学习算法自身在一个可能很庞大的假设空间中，对假设进行选择的启发式或“价值观”.

有没有一般性的原则来引导算法确立"正确的"偏好呢?

- “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”.
  - ✓ 奥卡姆剃刀并非唯一可行的原则
  - ✓ 奥卡姆剃刀本身存在不同的诠释
- 归纳偏好对应了学习算法本身所做出的关于"什么样的模型更好"的假设.
- 具体的现实问题中，学习算法本身所做的假设是否成立，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.

归纳偏好——没有免费的午餐，NFL定理

- 假设学习算法 ζ_a 基于某种归纳偏好产生了对应于曲线 A 的模型，学习算法 ζ_b 基于另一种归纳偏好产生了对应于曲线 B 的模型.
- 基于前面讨论的平滑曲线的某种“描述简单性”我们满怀信心地期待算法 ζ_a 比 ζ_b 更好。
- 事实上与B相比，A与训练集外的样本更一致，A的泛化能力比B强.
- 遗憾的是，实际应用中，“一个模型肯定比另一个模型具有更强的泛化能力”的这种情况是不存在的。
- 一个算法ζ_a如果在某些问题上比另一个算法ζ_b好，必然存在另一些问题ζ_b比ζ_a好，即没有免费的午餐定理。



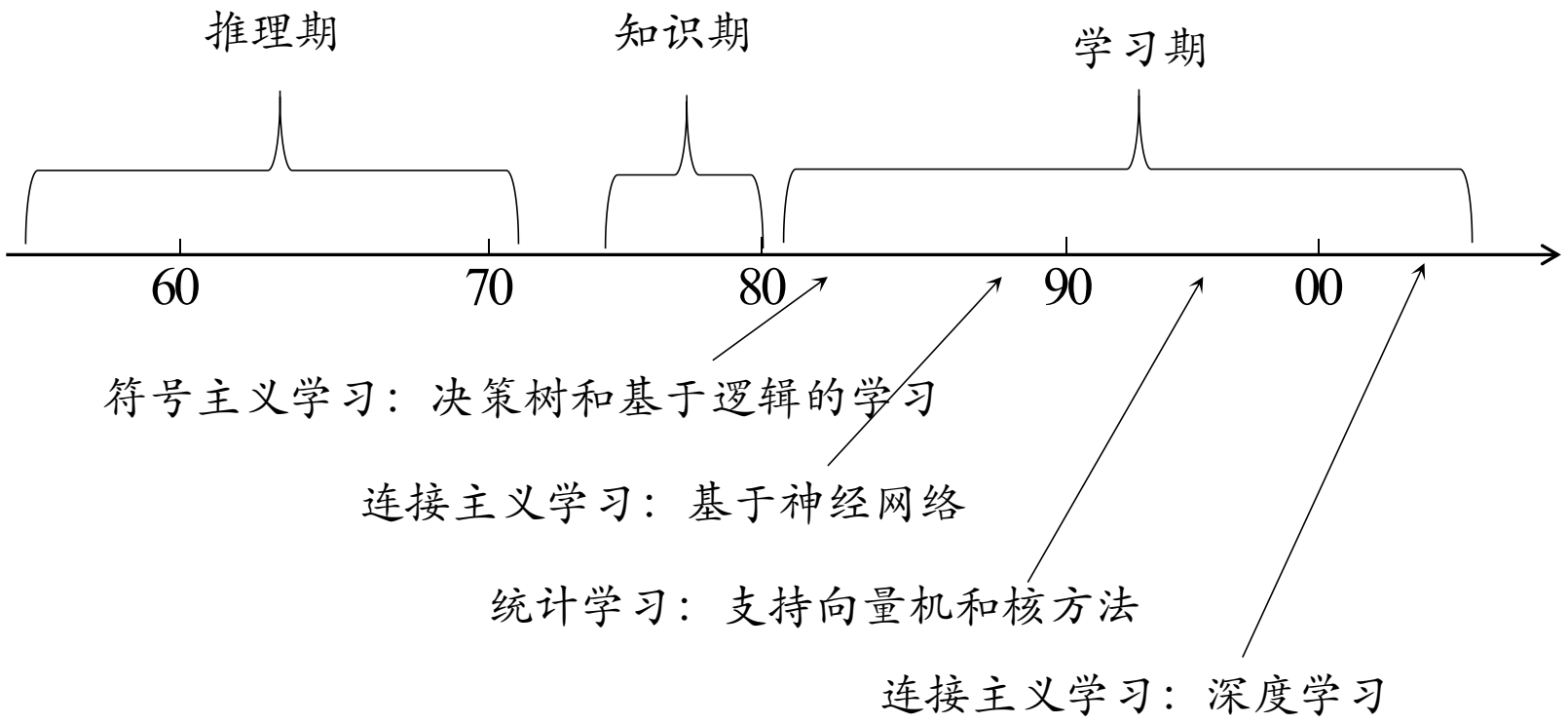
存在多条曲线与有限样本训练集一致

- 实际问题中，脱离具体问题，空谈“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
  - ✓ 在某些问题上表现好的学习算法，在另一些问题上却可能不尽人意
  - ✓ 学习算法自身的归纳偏好与问题是否相配，往往会起到决定性的作用

大纲

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

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**

应用现状

- 计算机领域最活跃的研究分支之一：

  在多媒体、图形学、网络通信、软件工程、体系结构、芯片设计、计算机视觉、自然语言处理、生物信息学、都能找到机器学习技术的身影

- 与普通人的生活密切相关：
  - **天气预报、能源勘探、环境监测**
    对卫星和传感器的数据进行分析，提高预报和检测准确性
  - **商业营销**
    对销售数据、客户信息进行分析，不仅可帮助商家优化库存降低成本，还有助于针对用户群设计特殊营销策略
  - **搜索引擎**
    很多人已习惯于在出行前通过互联网搜索来了解目的地信息、寻找合适的酒店、餐馆等。互联网搜索是通过分析网络上的数据来找到用户所需的信息，在这个过程中，用户查询是输入、搜索结果是输出，而要建立输入与输出之间的联系，必然需要机器学习技术
  - **自动驾驶汽车**
    把车载传感器接收到的信息作为输入，把方向、刹车、油门的控制行为作为输出，其关键问题可抽象为一个机器学习任务

应用现状

- 计算机领域最活跃的研究分支之一：

  在多媒体、图形学、网络通信、软件工程、体系结构、芯片设计、计算机视觉、自然语言处理、生物信息学、都能找到机器学习技术的身影

- 影响到人类社会的政治生活：

  2012美国大选期间，奥巴马的机器学习团队，对社交网络等各类数据进行分析，为奥巴马提示下一步的竞选行动。

  - 机器学习模型分析出，某电影明星对某地区某年龄段的特定人群很有吸引力。而这个群体很愿意出高价与该明星及奥巴马共进晚餐。果然，这样一次筹资晚宴成功募集到1500万美元
  - 机器学习模型通过对不同群体选民进行分析，建议购买了一些冷门节目的广告时段，而没有采用在昂贵的黄金时段购买广告的传统做法，使得广告资金效率相比2008 年竞选提高了14%;
  - 胜选后，《时代》周报报道了奥巴马的半监督学习研究专家

应用现状

- 计算机领域最活跃的研究分支之一：

  在多媒体、图形学、网络通信、软件工程、体系结构、芯片设计、计算机视觉、自然语言处理、生物信息学、都能找到机器学习技术的身影

- 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”
  - **生物信息学**，试图利用信息技术来研究生命现象和规律，而基因组计划的实施和基因药物的美好愿景让人们为之心潮薛湃。生物信息学研究涉及从"生命现象"到"规律发现"的整个过程，其间必然包括数据获取、数据管理、数据分析、仿真实验等环节，而"数据分析"恰是机器学习技术的舞台

大纲

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

知识总结：第1章 绪论

1. **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”，并使用该模型对新的情况给出判断。

2. 根据**预测目标的不同**，学习任务可以分为3类：
   - **分类**：预测值的是离散值
     - ✓ **二分类**：正类（好瓜）；反类（坏瓜）；**多分类**：冬瓜;南瓜;西瓜
   - **回归**：预测值的是连续值　e.g.西瓜成熟度0.95、0.37
   - **聚类**：无预测值，即没有标记信息

3. 根据**训练数据是否拥有标记信息**，学习任务又可以分为3类：

   监督学习：分类、回归；**无监督学习**：聚类；**半监督学习**：两者结合

4. **泛化**：机器学习的目标是使学得的模型能很好地适用于“新样本”，该模型适用于新样本的能力，称为泛化。
   - ✓ 泛化 能力，指模型在训练集上表现出的性能。
   - ✓ 训练样本越多，越有可能通过学习获得强泛化能力的模型

知识总结：第1章 绪论

1. 归纳：从特殊到一般的"泛化"过程，即从具体的事实归结出一般性规律
2. 演绎：从一般到特殊的"特化"过程，即从基础原理推演出具体状况.
3. 学习过程中对某种类型假设的偏好称作归纳偏好
   - ✓ “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”.
   - ✓ 奥卡姆剃刀，并非唯一可行的原则
   - ✓ 奥卡姆剃刀，本身存在不同的诠释
4. 一个算法ζ_a如果在某些问题上比另一个算法ζ_b好，必然存在另一些问题，ζ_b比ζ_a好，即没有免费的午餐定理。