第一章 数据科学与机器学习基本概念

1. 指导学习是什么？

1.1 什么是指导学习：映射、函数、泛化

1.2 指导学习的类型：搜索、生成、模型

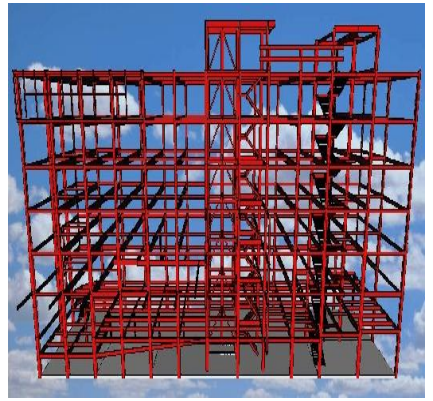
1.3 概念和概念学习

2. 基本术语：

2.1 数据集、属性、样本空间

2.2 假设、假设空间、偏序，归纳学习

2.3 版本空间和泛化学习



什么是机器学习？

- 计算机中，“经验”通常是以“数据”形式储存下来，因此机器学习所研究的主要内容是关于如何从观测数据中**不断学习**和总结**经验**，从数据中产生“**模型**”（model）帮助计算机做出准确判断的自动化技术，这种技术称为算法，也称为“learning algorithm”.
- 两个注释：
  - “模型”是算法的结果，通过经验提升自身的性能。
  - “经验”可以从哲学、社会、人文等多个角度进行解读。归纳起来“经验”是人与客观事物接触过程中，通过感官获得的关于客观事物的现象和外部联系的认识
    practical contact with and observation of facts or events.
- 通过观察客观现象与结果之间的联系所发现的规律就是经验。
- 机器学习的特点：输入数据，产出模型，并能动态自我更新的算法。

什么是大数据？Big data the next frontier for innovation, competition and productivity

（观察、归纳和判断的模式化）

4V特性(Variable,Variety,Variance,Visulization)

体量Volume	*非结构化数据*的超大规模和增长 总数据量的80~90% 比结构化数据增长快10倍到50倍 是传统数据仓库的10倍到50倍
多样性Variety	大数据的异构和多样性 很多不同形式（文本、图像、视频、机器数据） 无模式或者模式不明显 不连贯的语法或句义
价值密度Value	大量的不相关信息 对未来趋势与模式的可预测分析 深度复杂分析（机器学习、人工智能Vs传统商务智能(咨询、报告等）
速度Velocity	*实时分析*而非批量式分析 数据输入、处理与丢弃 立竿见影而非事后见效

大数据分析中统计学的作用

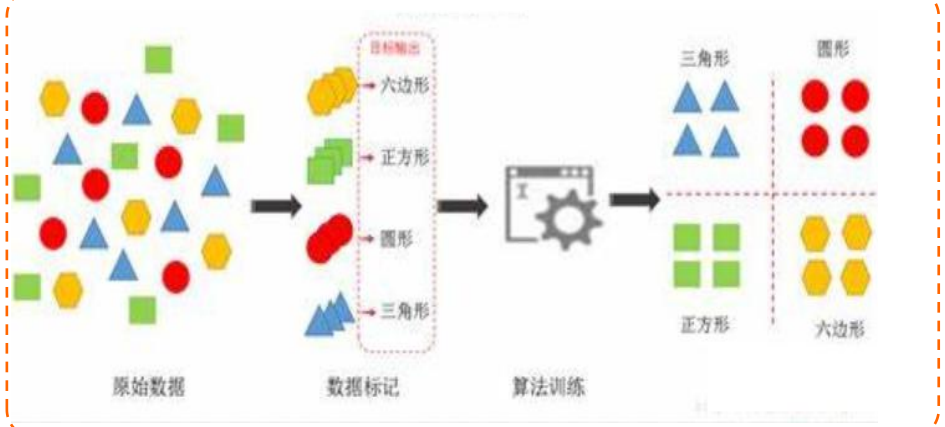
- “人工智能”、“大数据”、“云计算”和“5G”是统计学和数据科学案例“好问题”和“好数据”的基础来源，提出新的社会现实，具有许多例外的特征。

- 统计学深度服务于数据科学案例的养分：
  - 系统性呈现证据；
  - 发现新见解的涌现；
  - “数据耳目一新”理论的构建；
  - 在原始数据和结论之间形成牢固证据链方面空间潜力巨大

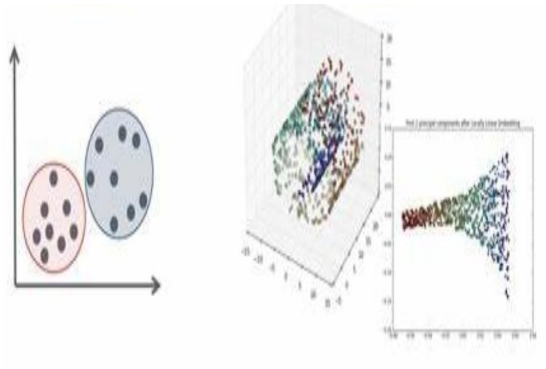
- 稀缺的“宝藏”信息的特征：以微弱信号的状态点滴积累
- 统计学家参与到研究设计和审核环节可增强其可及性(touchable)，将研究者有限的注意力指向与任务最相关的要务信息，迅速找到见微知著的出口，以利于做出正确的判断。

学习的分类

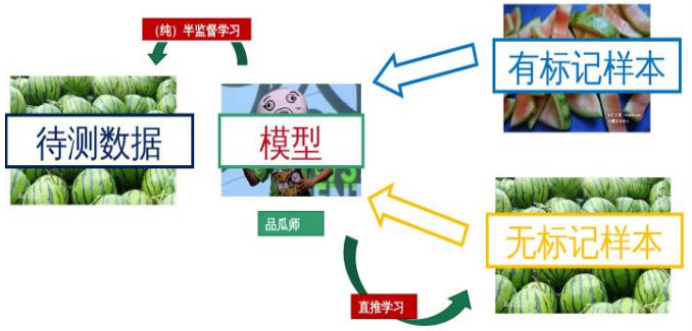
有指导学习



无指导学习



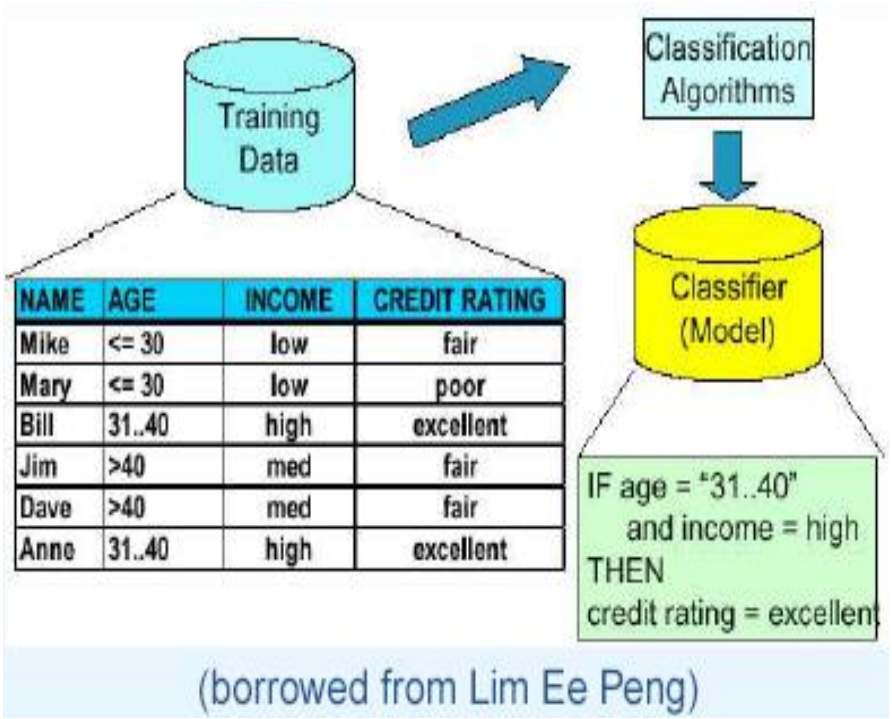
半指导学习

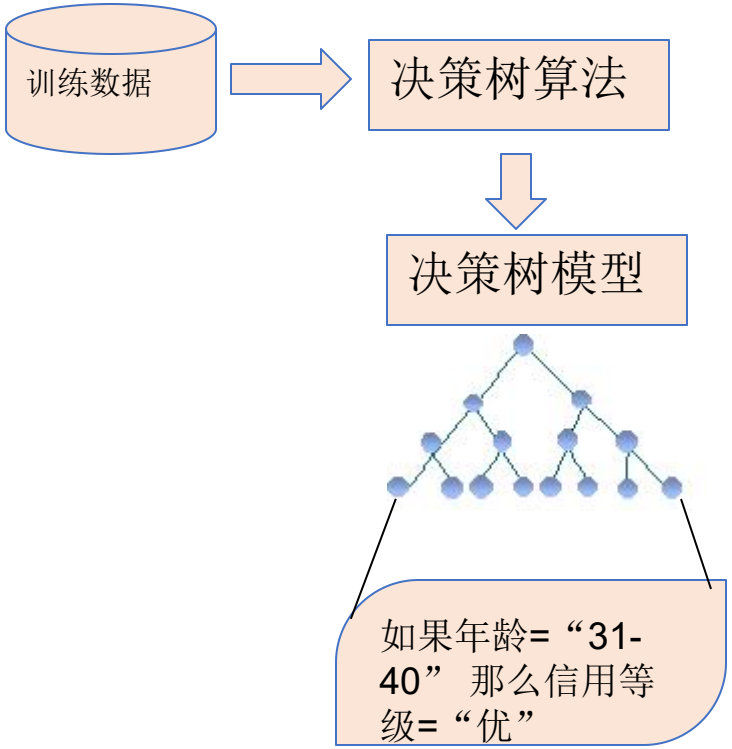


增强学习

有指导的学习(Supervised Learning):

用一组输入变量 (predictors, inputs, features, independents)
对输出变量 (responses, outputs , dependent) 产生预测





学习问题的一般表示

- $\mathcal{X}$ 输入空间（ $\mathcal{X} \subseteq \mathbb{R}^d$ ），每个元样例 $x_i = \{x_{i1}, \cdots, x_{id}\}$.
- $\mathcal{Y}$ 输出空间每个元样例 y_i.

  分类问题（分类的输出）：$\mathcal{Y} = \{c_1, \cdots, c_k\}$;

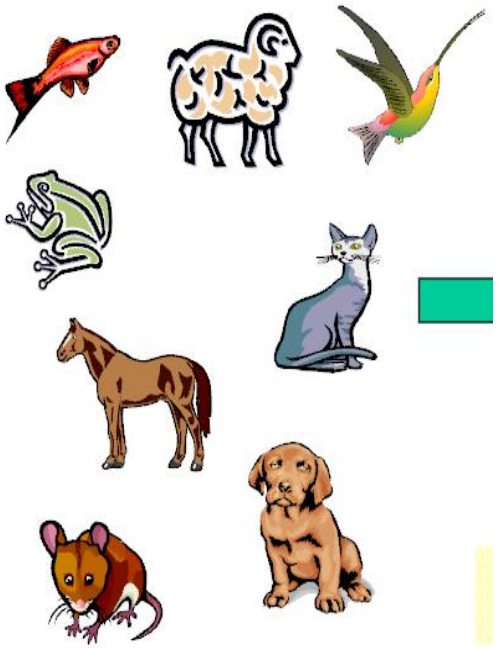
  回归问题（连续的输出）：$\mathcal{Y} \subseteq \mathbb{R}$
- $S = \{(x_i, y_i)\}_{i=1}^m$：训练样本

有监督的学习模型问题就是要计算出一个最优的函数，该函数可以恰当地描述输入和输出之间的关系。

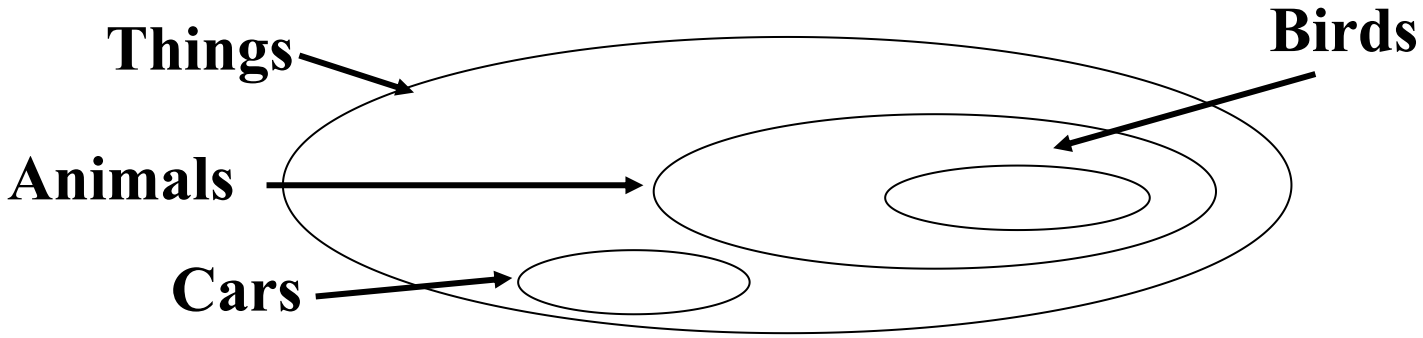
指导学习的定义和类型

- **定义：指导**学习的目标是学习输入到输出的映射关系，其中正确值已部分地由指导者通过训练数据给出。
- **类型**:
  - 概念学习：0-1学习
    - 特点1：将学习问题转化为一个搜索问题;
    - 特点2：强调假设空间的性质、搜索算法和评价准则;
  - 生成式学习:
  - 模型学习（统计学习）：回归
    - 特点1：将学习问题转化为一个估计问题，特别是分布的特征估计问题;
    - 特点2：强调分布选择，估计的性质和模型的解释;

概念是什么

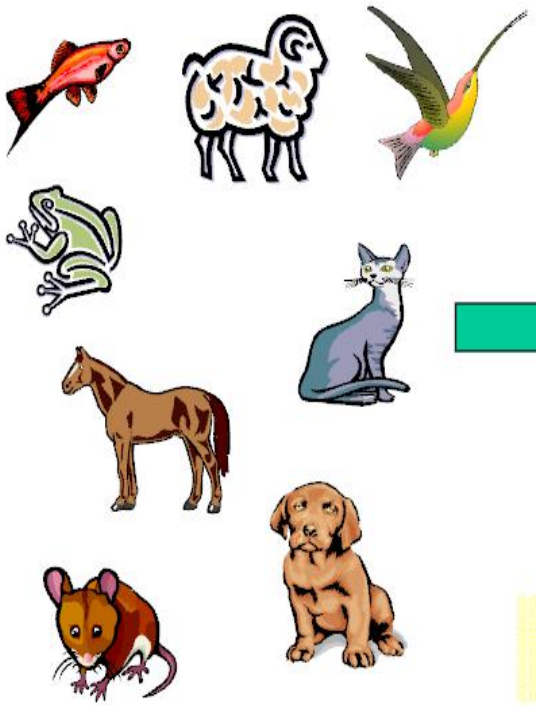


- A Concept is a subset of objects or events defined over a larger set [Example: The concept of a *bird* is the subset of all objects (i.e., the set of all *things* or all *animals*) that belong to the category of bird.]概念是一组研究对象或事件的集合。它是较大集合中选取的子集，或在较大集合中定义的布尔函数



- Alternatively, a concept is a boolean-valued function defined over this larger set [Example: a function defined over all *animals* whose value is *true* for birds and *false* for every other animal].

概念学习



- 概念学习：是指从有关某个布尔函数的输入输出训练样例中推断出该布尔函数。
- 另一种定义：给定一样例集合以及每个样例是否属于某一概念的标注，怎样自动推断出该概念的一般定义。这一问题被称为概念学习。
- “一个搜索过程，它在预定义的假设空间中按着某种搜索策略进行搜索，使学到的概念与训练实例有最佳的拟合度。”- Tom Michell

2.1 数据集、属性维度到样本空间

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

- 以样本的属性为坐标轴张成的多维空间，也叫属性空间、输入空间。
- 上例中，每行样本包含三个属性：色泽、根蒂、敲声，则可以以这三个属性为坐标轴，生成一个三维空间，每个西瓜（只要用这三种属性描述）都能在该空间中找到其对应的坐标位置。
- 通常样本空间中的全体样本服从未知分布，采样越多，对样本空间分布的认知越多。

一般意义上的概念学习

Concept: "days on which my friend Tom enjoys his favourite water sports"

Task: predict the value of "Enjoy Sport" for an arbitrary day based on the values of the other attributes

attributes

example

Sky	Temp	Humid	Wind	Water	Fore-cast	Enjoy Sport
Sunny	Warm	Norr		Warm	Same	Yes
Sunny	Warm	High		Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

概念的要素构成学习例1

Database:

Day	*Sky*	*AirTemp*	*Humidity*	*Wind*	*Water*	*Forecast*	*WaterSport*
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

class

假设的形式:

合取**Conjunction of constraints on each attribute** where:

- "?" means "any value is acceptable"
- "0" means "no value is acceptable"

Example of a hypothesis: <?,Cold,High,?,?,?>

(If the air temperature is cold and the humidity high then it is a good day for water sports)

补充：合取范式

- 析取范式disjunctive normal form (DNF)和合取范式conjunctive normal form (CNF) 是命题逻辑等值演算中的重要内容，其目的是为了标准化命题公式。
  - 合取 (Conjunctive) 指代的是与操作 (∧)。仅有与运算符连接而成的布尔表达式称为合取子句 (Conjunctive clause)。如 (A∧B)或(A∧B∧¬C)
  - 析取 (Disjunctive) 指代的是或操作 (∨)。仅由或运算符连接而成的布尔表达式称为析取子句 (Disjunctive clause)。如 (A∨B∨¬C)
- 合取范式：析取子句用合取操作连接起来：(A1∨A2)∧(A3∨A4)
  - 例(Age=[30,39)VAge=[50,60))∧ income=[40,49))V(income=[100,500)) =>credit=”good”
- 析取范式：合取子句用析取操作连接起来(A1∧A2)∨(A3∧A4)
  - 例如：Age=[20,29)∨income=[0,20) =>credit=”bad”
- A,¬A既是一个简单析取式，又是一个简单合取式
- A∧(¬C∨A)∧(B∨C) 合取范式

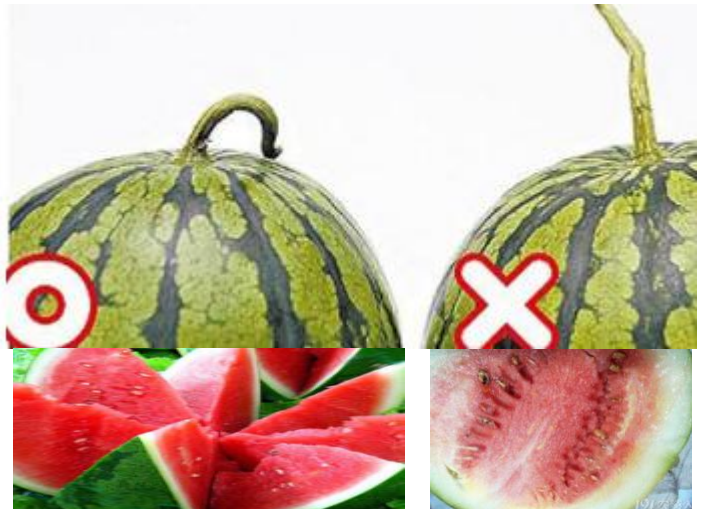
西瓜集上的训练集、泛化

- 任务Task T: 识别好瓜
- 评价Performance measure P:正确识别好瓜的比例
- 训练样例Training experience E: 3个好瓜和坏瓜的样例。

表 1.1 西瓜数据集

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

- 目标概念：待学习的概念和函数c:X->{0,1}
- 训练集属性色泽、根蒂、敲声分别有2、3、3种可能取值;
- 假设空间：采用属性合取式描述假设空间假设空间由形如“(色泽=？) ∧ (根蒂=？) ∧ (敲声=？)”的所有假设组成;
- 泛化：通过对训练集中“好瓜”的经验归纳出对没有见过的瓜进行判断的能力。



- 参考周志华《机器学习》，2016

假设空间

表 1.1 西瓜数据集

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

表1.1的训练数据集对应的假设空间应该如下：

1 色泽 = * , 根蒂 = * , 敲声 = *

2 色泽 = 青绿 , 根蒂 = * , 敲声 = *

3 色泽 = 乌黑 , 根蒂 = * , 敲声 = *

4 色泽 = * , 根蒂 = 蜷缩 , 敲声 = *

5 色泽 = * , 根蒂 = 硬挺 , 敲声 = *

6 色泽 = * , 根蒂 = 稍蜷 , 敲声 = *

7 色泽 = * , 根蒂 = * , 敲声 = 浊响

8 色泽 = * , 根蒂 = * , 敲声 = 清脆

9 色泽 = * , 根蒂 = * , 敲声 = 沉闷

10 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = *

11 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = *

12 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = *

13 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = *

14 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = *

15 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = *

16 色泽 = 青绿 , 根蒂 = * , 敲声 = 浊响

17 色泽 = 青绿 , 根蒂 = * , 敲声 = 清脆

18 色泽 = 青绿 , 根蒂 = * , 敲声 = 沉闷

19 色泽 = 乌黑 , 根蒂 = * , 敲声 = 浊响

20 色泽 = 乌黑 , 根蒂 = * , 敲声 = 清脆

21 色泽 = 乌黑 , 根蒂 = * , 敲声 = 沉闷

22 色泽 = * , 根蒂 = 蜷缩 , 敲声 = 浊响

23 色泽 = * , 根蒂 = 蜷缩 , 敲声 = 清脆

24 色泽 = * , 根蒂 = 蜷缩 , 敲声 = 沉闷

25 色泽 = * , 根蒂 = 硬挺 , 敲声 = 浊响

26 色泽 = * , 根蒂 = 硬挺 , 敲声 = 清脆

27 色泽 = * , 根蒂 = 硬挺 , 敲声 = 沉闷

28 色泽 = * , 根蒂 = 稍蜷 , 敲声 = 浊响

29 色泽 = * , 根蒂 = 稍蜷 , 敲声 = 清脆

30 色泽 = * , 根蒂 = 稍蜷 , 敲声 = 沉闷

31 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = 浊响

32 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = 清脆

33 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = 沉闷

34 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = 浊响

35 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = 清脆

36 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = 沉闷

37 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = 浊响

38 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = 清脆

39 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = 沉闷

40 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = 浊响

41 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = 清脆

42 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = 沉闷

43 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = 浊响

44 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = 清脆

45 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = 沉闷

46 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = 浊响

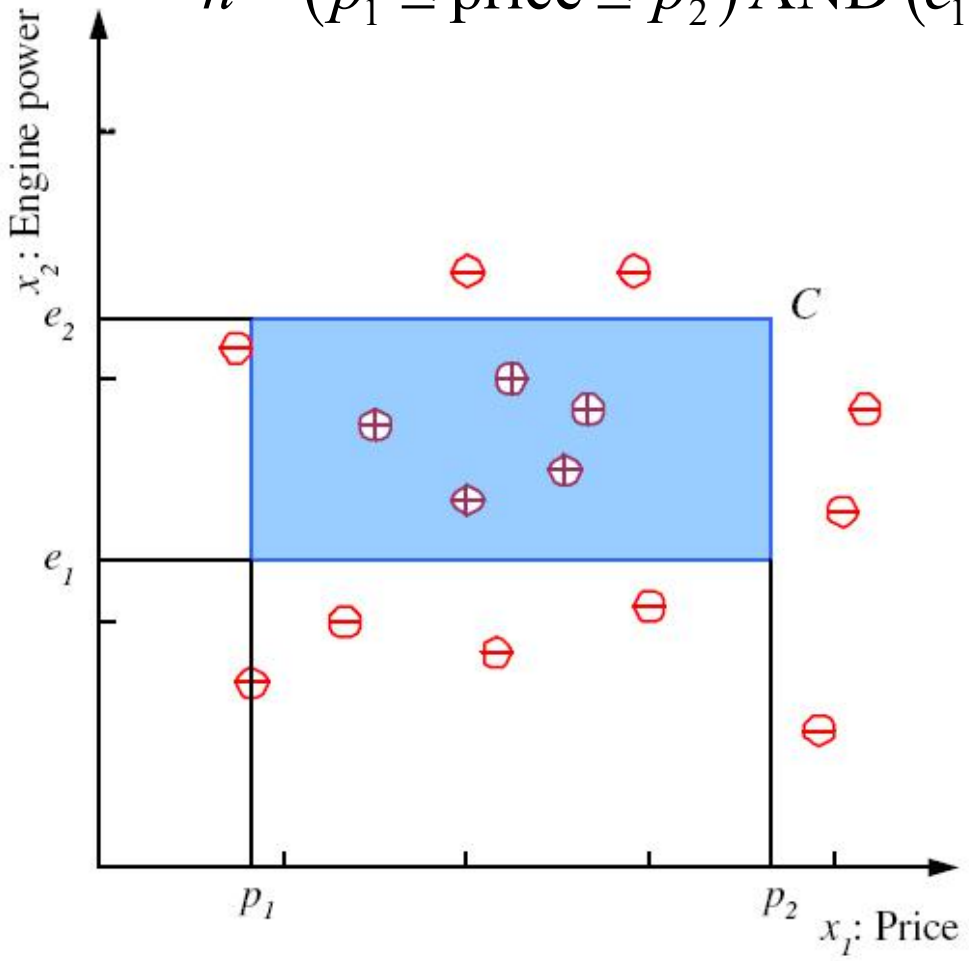
47 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = 清脆

48 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = 沉闷

49 Ø

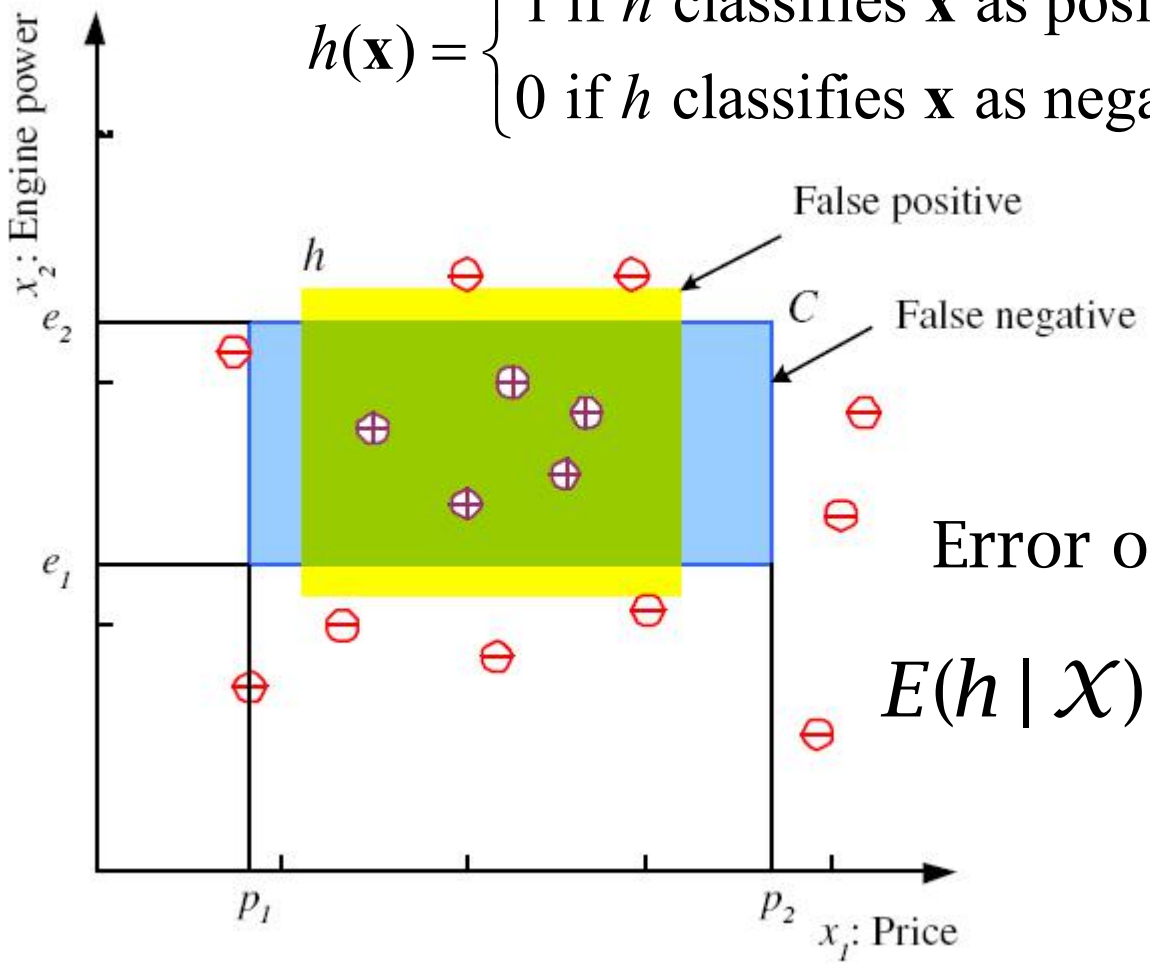
怎么学习？能学会吗？概念学习连续的问题：真实的概念Class C

$$h = (p_1 \le \text{price} \le p_2) \text{ AND } (e_1 \le \text{engine power} \le e_2)$$



假设空间 Hypothesis class $\mathcal{H}$

$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ classifies } \mathbf{x} \text{ as positive} \\ 0 \text{ if } h \text{ classifies } \mathbf{x} \text{ as negative} \end{cases}$$



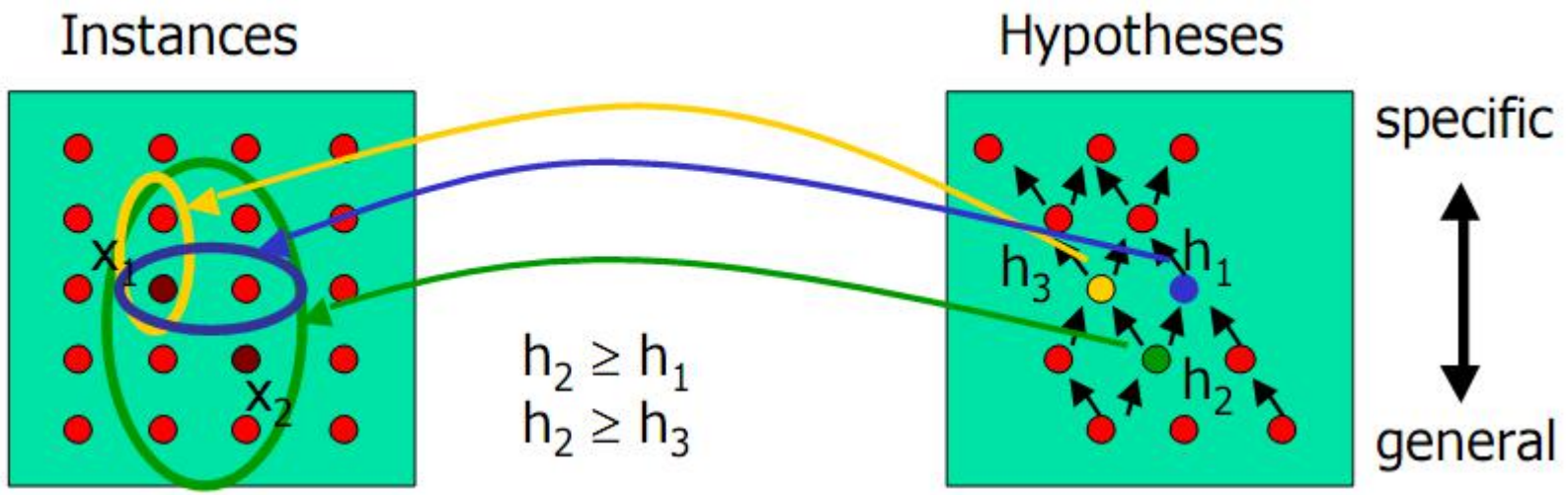
Error of h on $\mathcal{H}$

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

假设的一般到特殊序：偏序,很多假设空间的假设存在序结构

x1=<青绿,蜷缩,浊响>

x3=<乌黑,蜷缩,浊响>



h3=<？,蜷缩,浊响>

h1=< 青绿,?,浊响 >

h2=< ？ ,？ ,浊响 >

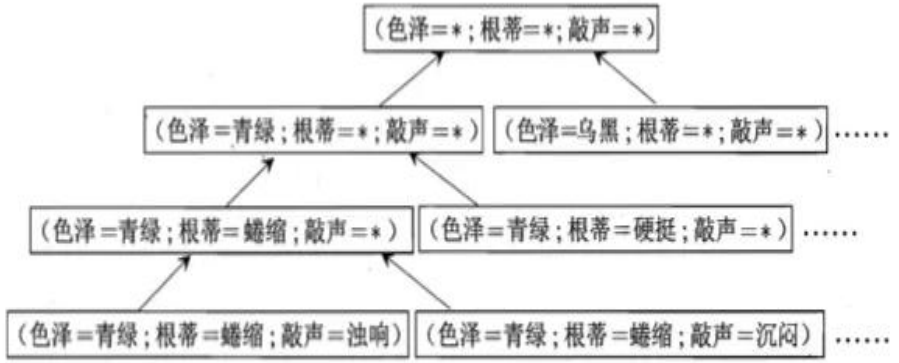
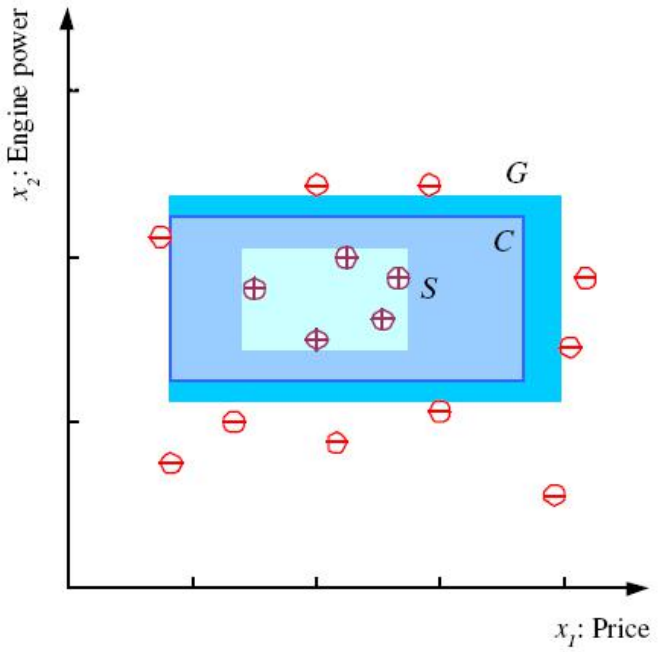


图 1.1 西瓜问题的假设空间

版本空间和一致性

最特殊的假设，最一般的假设



涵盖所有正例不包括任何负例的最小的假设

<∅, ∅, ∅, ∅,…, ∅ >

most general hypothesis, G

<?,?,?,?,…,?>涵盖所有正例不包括任何负例的最大的假设

h ∈ $\mathcal{H}$, between S and G is 满足一质性 and make up the 版本空间(Mitchell, 1997)

一致性定义:一个假设h和训练样本集称为一致,当且仅当对D中每个样例(x,C(x)),有C(x)=h(x)

$$\text{consistent}\quad (h, D) = \{\forall (x, c(x)) \in D, h(x) = C(x)\}$$

2.2.版本空间基本概念

- Any $h \in H$ between S and G
- Consisting of valid hypotheses with no error (consistent with the training set)

1.**版本空间定义**: 假设空间H和训练数据集D的版本空间是H中每个与训练样本D一致的假设构成的子集

$$VS_{H,D} = \{h \in H, Consistent(h, D)\}$$

2.关于假设空间H和训练数据集D的**一般边界**(General Boundary)

$$G = \{g \in H, Consistent(g, D) \wedge [\neg \exists ((g' > g) \wedge Consistent(g', D))]\}$$

3.关于假设空间H和训练数据集D的**特殊边界**(Specific Boundary)

$$S = \{s \in H, Consistent(s, D) \wedge [\neg \exists ((s > s') \wedge Consistent(s', D))]\}$$

归纳学习（Inductive Learning）

- 机器学习的目标是从假设空间H中找到h(x):X →{0,1}，使得h(x)=c(x).
- **归纳学习的基本假设：**任何一个假设如果在足够大的训练样例中很好的逼近目标函数，那么也可能在未见的实例中更好地逼近目标函数。

Given:

- Instance Space X : Possible days decribed by the attributes Sky, Temp, Humidity, Wind, Water, Forecast
- Target function c: EnjoySport X → {0,1}
- Hypothese Space H: conjunction of literals e.g.

  < Sunny ? ? Strong ? Same >
- Training examples D : positive and negative examples of the target function: $<x_1,c(x_1)>,\ldots, <x_n,c(x_n)>$

Determine:

- A hypothesis h in H such that h(x)=c(x) for all x in D.

FIND-S学习算法

表 2-3　FIND-S 算法

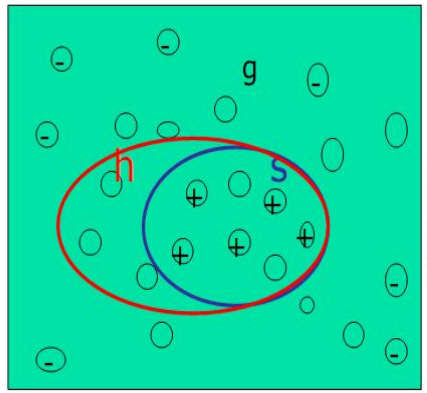
1. 将 h 初始化为 H 中最特殊假设
2. 对每个正例 x
   - 对 h 的每个属性约束 a_i
     - 如果 x 满足 a_i
     - 那么不做任何处理
     - 否则将 h 中 a_i 替换为 x 满足的另一个更一般约束
3. 输出假设 h

表 1.1　西瓜数据集

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

- 属性合取式的假设空间
- 保证输出为H中与正例一致的最大的特殊假设
- 只要正确的目标概念在H中，训练数据正确，也可能得到最好的假设

Find-S的不足：

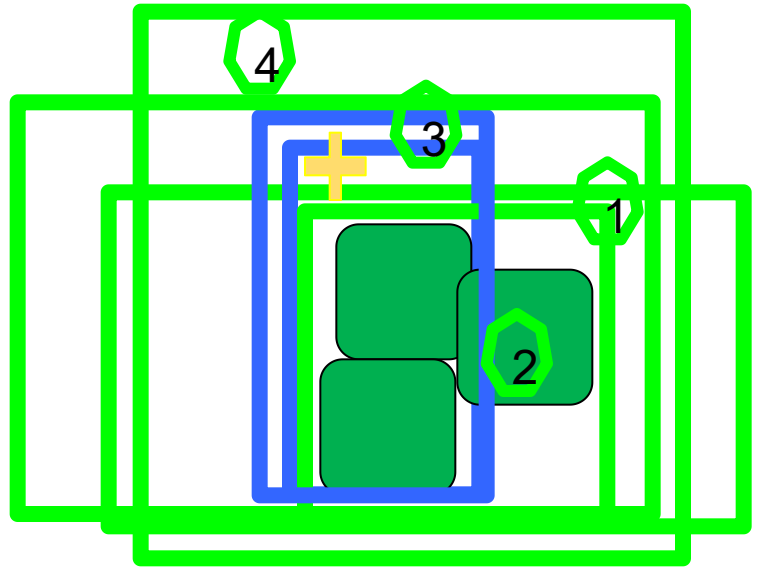


h is consistent with D, then h>s;

用于概念学习的 Find-S 算法是机器学习最基本的算法之一，它的不足和缺点十分明显：

1. 无法确定最终假设（Find-S 找出的）是否是唯一一个与数据一致（consistent）的假设。也**不允许多个假设共存；无法评价所得假设与目标的接近程度。**
2. **忽略一般性：**不一致（inconsistent）的训练实例会误导 Find-S 算法；
3. 忽略“负”例。**Find-s找到的假设仅是H中最特殊的一种，要保证得到的假设与所有的训练数据集一致，则必须考虑反例如何从空间中剔除（也就是寻找极大一般边界）。**一个能检测训练数据不一致的算法才是更好的选择。
4. 一个好的概念学习算法应该能够回溯对找到的假设的选择，以便能够逐步改进所得到的假设。但不幸的是，Find-S 不能提供这样的方法。
5. 许多局限性都可以通过一个被称之为概念学习的最重要的算法来消除，它便是：候选删除算法（candidate elimination algorithm)。

改进版：版本空间的候选消除算法



- 将G中初始化为H中极大一般假设
- 将S中初始化为H中极小特殊假设
- 对每个样例d,进行以下操作:

如果d是正例

（1）从G中移出所有与d不一致的假设

（2）对S中每个与d不一致的假设s

从S中移除s

把s中所有极小泛化h加入到S中，其中h满足与 d一致，而且G的每个成员比h更一般

（3）从S中移除所有的假设：它比S中另一假设更一般

如果d是负例

（1）从S中移出所有与d不一致的假设

（2）对G中每个与d不一致的假设g

从G中移除g

把g中所有极小特殊h加入到G中，其中h满足与 d一致，而且S的每个成员比h更特殊

（3） 从G中移除所有的假设：它比G中另一假设更特殊

在西瓜问题中，如何根据训练集求所对应的版本空间

- ①写出假设空间：先列出所有可能的样本点（即特征向量）（即每个属性都取到所有的属性值）
- ②对应着给出已知数据集，将与正样本不一致的、与负样本一致的假设删除。
- 即可得出与训练集一致的假设集合，也就是版本空间了。

表 1.1 西瓜数据集

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

表1.1的训练数据集对应的假设空间应该如下：

1 色泽 = * , 根蒂 = * , 敲声 = *

2 色泽 = 青绿 , 根蒂 = * , 敲声 = *

3 色泽 = 乌黑 , 根蒂 = * , 敲声 = *

4 色泽 = * , 根蒂 = 蜷缩 , 敲声 = *

5 色泽 = * , 根蒂 = 硬挺 , 敲声 = *

6 色泽 = * , 根蒂 = 稍蜷 , 敲声 = *

7 色泽 = * , 根蒂 = * , 敲声 = 浊响

8 色泽 = * , 根蒂 = * , 敲声 = 清脆

9 色泽 = * , 根蒂 = * , 敲声 = 沉闷

10 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = *

11 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = *

12 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = *

13 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = *

14 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = *

15 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = *

16 色泽 = 青绿 , 根蒂 = * , 敲声 = 浊响

17 色泽 = 青绿 , 根蒂 = * , 敲声 = 清脆

18 色泽 = 青绿 , 根蒂 = * , 敲声 = 沉闷

19 色泽 = 乌黑 , 根蒂 = * , 敲声 = 浊响

20 色泽 = 乌黑 , 根蒂 = * , 敲声 = 清脆

21 色泽 = 乌黑 , 根蒂 = * , 敲声 = 沉闷

22 色泽 = * , 根蒂 = 蜷缩 , 敲声 = 浊响

23 色泽 = * , 根蒂 = 蜷缩 , 敲声 = 清脆

24 色泽 = * , 根蒂 = 蜷缩 , 敲声 = 沉闷

25 色泽 = * , 根蒂 = 硬挺 , 敲声 = 浊响

26 色泽 = * , 根蒂 = 硬挺 , 敲声 = 清脆

27 色泽 = * , 根蒂 = 硬挺 , 敲声 = 沉闷

28 色泽 = * , 根蒂 = 稍蜷 , 敲声 = 浊响

29 色泽 = * , 根蒂 = 稍蜷 , 敲声 = 清脆

30 色泽 = * , 根蒂 = 稍蜷 , 敲声 = 沉闷

31 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = 浊响

32 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = 清脆

33 色泽 = 青绿 , 根蒂 = 蜷缩 , 敲声 = 沉闷

34 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = 浊响

35 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = 清脆

36 色泽 = 青绿 , 根蒂 = 硬挺 , 敲声 = 沉闷

37 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = 浊响

38 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = 清脆

39 色泽 = 青绿 , 根蒂 = 稍蜷 , 敲声 = 沉闷

40 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = 浊响

41 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = 清脆

42 色泽 = 乌黑 , 根蒂 = 蜷缩 , 敲声 = 沉闷

43 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = 浊响

44 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = 清脆

45 色泽 = 乌黑 , 根蒂 = 硬挺 , 敲声 = 沉闷

46 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = 浊响

47 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = 清脆

48 色泽 = 乌黑 , 根蒂 = 稍蜷 , 敲声 = 沉闷

49 Ø

Candidate-Elimination Algorithm

$S_0 = \{\langle \varnothing, \varnothing, \varnothing \rangle\}$
$G_0 = \{\langle ?, ?, ? \rangle\}$

$S_1 = \{\langle$青绿, 蜷缩, 浊响 $\rangle\}$
$G_1 = \{\langle ?, ?, ? \rangle\}$

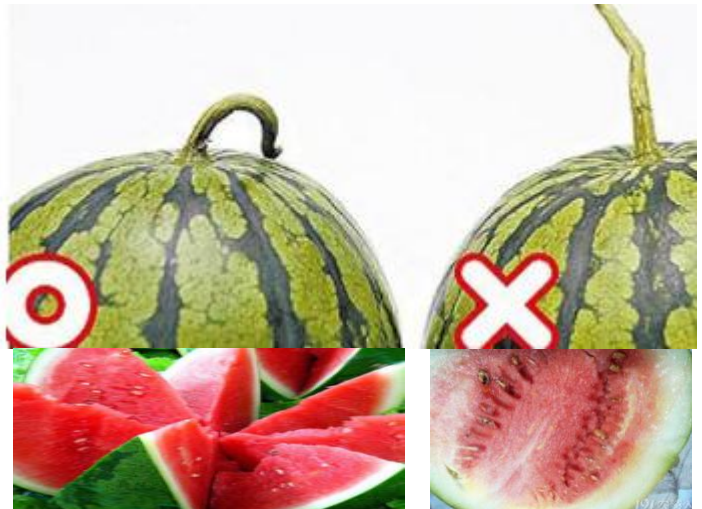
$S_2 = \{\langle ?,$ 蜷缩, 浊响$\rangle\}$
$G_2 = \{\langle ?, ?, ? \rangle\}$

$S_3 = \{\langle ?,$ 蜷缩, 浊响$\rangle\}$
$G_3 = \{\langle ?,$蜷缩, $? \rangle, \langle ?, ?,$浊响$\rangle\}$

$S_4 = \{\langle ?,$ 蜷缩, 浊响$\rangle\}$
$G_4 = \{\langle ?,$蜷缩, $? \rangle, \langle ?, ?,$浊响$\rangle\}$

表 1.1 西瓜数据集

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



Candidate-EliminationAlgorithm(收敛性)

- 候选消去算法的特点：寻找与训练样例一致的假设；
- 原理：通过S泛化和G的特殊化不断缩小版本空间，实现对一致假设的搜索。
- The version space will converge toward the correct target concepts if:
  - H contains the correct target concept
  - H中包含了描述目标概念的正确假设(可知学习)
  - There are no errors in the training examples
    在训练样本中没有错误（完全学习）
- A training instance to be requested next should discriminate among the alternative hypotheses in the current version space。理想的训练样例是对S和G都有作用，于是可以使边界单调移动，从而有效地推动搜索进程。
- Partially learned concept can be used to classify new instances using the majority rule.不完全学习仍然可以用于预测

学习过后剩余的假设为：

- 4 色泽 = *，根蒂 = 蜷缩，敲声 = *
- 7 色泽 = *，根蒂 = *，敲声 = 浊响
- 22 色泽 = *，根蒂 = 蜷缩，敲声 = 浊响

这就是最后的“假设集合”，也就是“版本空间”。