第六章 机器学习

机器学习 — 概述

- 参考书：

  《Machine Learning》,Tom M. Mitchell, 1997,（机械出版社）
- 什么是机器学习？
  - Simon（1983）：学习就是系统中的变化，这种变化使系统比以前更有效地去做同样的工作。
  - Minsky（1985）：学习是在我们头脑中（内心）进行有用的变化。
  - 学习是一种具有多侧面的现象。学习的过程有：获取新的陈述性知识、通过教育或实践发展机械技能和认知能力、将新知识组织成为通用化和有效的表达形式、借助观察和实验发现新的事实和新的理论。

机器学习 — 概述

- 基本形式
  - 知识获取和技能求精。

    学习的本质就是获取新的知识。包括物理系统和行为的描述和建模，构造客观现实的表示。

    ——知识获取
  - 通过实践逐渐改造机制和认知技能。

    例：骑自行车。这些技能包括意识的或机制的协调。这种改进又是通过反复实践和从失败的行为中纠正偏差来进行的。

    ——技能求精

机器学习 — 概述

- 知识获取的本质可能是一个自觉的过程，其结果是产生新的符号知识结构和智力模型。
- 而技能求精则是下意识地借助于反复的实践来实现的。

机器学习 – 概述

- 为什么要研究机器学习？
  - AI主要是为了研究人的智能，模仿其机理并应用于工程的科学。在这个过程中必然会问道：“人类怎样做才能获取这种特殊技能（或知识）？”。

机器学习 — 概述

- 为什么要研究机器学习？
  - 当前AI研究的主要障碍和发展方向之一就是机器学习。
  - 学习的计算理论和构造学习系统。
  - 现在的AI系统还完全没有或仅有很有限的学习能力。系统中的知识由人工编程送入，知识中的错误也不能自动改正。
  - 现有的大多数AI是演绎的、没有归纳推理，因而不能自动获取和生成知识。

机器学习 — 概述

为什么要研究机器学习？

- 未来的计算机将有自动获取知识的能力，它们直接由书本学习，通过与人谈话学习，通过观察学习。
- 通过实践自我完善，克服人的存储少、效率低、注意力分散、难以传送所获取的知识等局限性。
- 一台计算机获取的知识很容易复制给其它机器。

机器学习 — 概述

- 实现的困难：
  - 预测难：学习后知识库发生了什么变化，系统功能变化的预测。
  - 归纳推理：现有的归纳推理只保证假，不保证真。演绎推理保真。而且，归纳的结论是无限多的，其中相当多是假的，给生成的知识带来不可靠性。
  - 机器目前很难观察什么重要、什么有意义。

机器学习 — 概述

- 发展历史
  - 神经系统模型和决策理论
    - 50's。一般称为神经网络或自组织系统
    - 由于当时计算机技术状态，多停留在理论和硬件上。这些元件类似于神经元，他们实现简单的逻辑功能。

    ……….

机器学习 — 概述

- 发展历史
  - 神经系统模型和决策理论

    ………

    - 1965年左右，神经网络→模式识别、决策论
    - 从给定的一组经过选择的例子中获得判断函数，有线性的、多项式的、或相关的形式。
    - Samuel (1959-1963) 的跳棋程序是最著名的成功的学习系统之一。达到了跳棋大师的水平。

机器学习 — 概述

– 符号概念获取

- 1975年左右提出的。
- 通过分析一些概念的正例和反例构造出这些概念的符号表示。
- 表示形式一般是逻辑表达式、决策树、产生式规则或语义网络。代表有Winston的ARCH。

机器学习 — 概述

– 知识加强和论域专用学习

- 70’s中期，沿着符号主义路线进行的。
- 加强重于专业的专用性。
- 强调使用面向任务的知识和它对学习过程的引导作用。
- 系统包括预先确定的概念、知识结构、论域约束、启发式规则和论域有关的变换。系统在开始并不具有所有的属性或概念，在学习过程中系统应得到一些新的属性或概念。

没有绝对的学习方法。许多系统体现出上述途径的组合。

机器学习 — 概述

- 机器学习进入新阶段的重要表现：（近十年）（1）
  - 机器学习已成为新的边缘科学并在高校形成一门课程。它综合应用心理学、生物学和神经生理学以及数学、自动化和计算机科学形成机器学习理论基础。

机器学习 — 概述

- 机器学习进入新阶段的重要表现：（近十年）（2）
  - 结合各种学习方法，取长补短的多种形式的集成学习系统的研究正在兴起。
  - 特别是连接学习，符号学习的耦合可以更好地解决连续性信号处理中知识与技能的获取与求精问题而受到重视。

机器学习 — 概述

– 机器学习进入新阶段的重要表现：（近十年）（3）
  - 机器学习与AI各种基础问题的统一性观点正在形成。
  - 例如：学习与问题求解结合进行，知识表达便于学习的观点产生了通用智能系统的组块学习。
  - 类比学习与问题求解结合的基于案例学习已成为经验学习的重要方向。

机器学习 — 概述

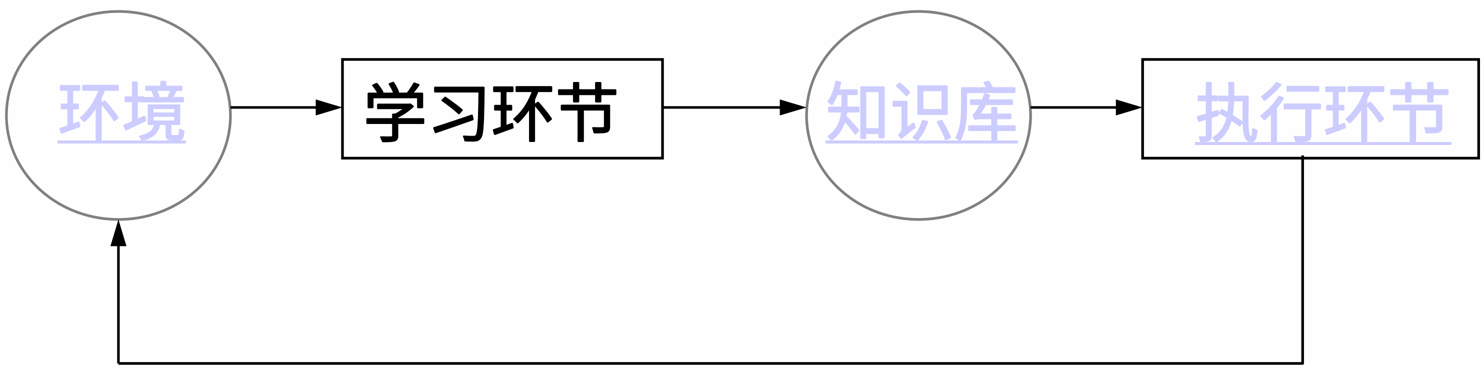
- 机器学习进入新阶段的重要表现：（近十年）(4)
  - 各种学习方法的应用范围不断扩大，一部分已形成商品。
  - 归纳学习的知识获取工具已在诊断分类性专家系统中广泛应用。
  - 连接学习在声图文识别中占优势。
  - 分析学习用于设计综合性专家系统。
  - 遗传算法与强化学习在工程控制中有较好的应用前景。
  - 与符号系统耦合的神经网络连接学习将在企业的智能管理与智能机器人运动规划中发挥作用。

机器学习 — 概述

- 机器学习进入新阶段的重要表现：（近十年）（5）
  - 与机器学习有关的学术活动空前活跃。国际上除每年一次的机器学习研究会外，还有计算机学习理论会议及遗传算法会议。

机器学习 — 概述

- 机器学习模型
  - 学习是建立理论、形成假设和进行归纳推理的过程。



  - 整个过程包括：信息的存储、知识的处理两部分

机器学习 — 概述

- 分类：按学习策略
  - 机械式学习，直接输入新知识（记忆学习）
    学习者不需要进行任何推理或知识转换，将知识直接装进机器中。
  - 根据示教学习（传授学习、指点学习）
    从老师或其它有结构的事物获取知识。要求学习者将输入语言的知识转换成它本身的内部表示形式。并把新的信息和它原有的知识有机地结合为一体。

机器学习 — 概述

…………

- 通过类推学习（演绎学习）

  学习者找出现有知识中所要产生的新概念或技能十分类似的部分。将它们转换或扩大成适合新情况的形式，从而取得新的事实或技能。

- 从例子中学习（归纳学习）

  给学习者提供某一概念的一组正例和反例，学习者归纳出一个总的概念描述，使它适合于所有的正例且排除所有的反例。（目前研究较多的一种方法）

机器学习 — 概述

……

– 类比学习

演绎学习与归纳学习的组合。匹配不同论域的描述、确定公共的结构。以此作为类比映射的基础。寻找公共子结构是归纳推理，而实现类比映射是演绎推理。

机器学习 — 概述

- 研究目的
  - 希望得到通用的算法
  - 研究了解学习知识的模型、认知模型
  - 解决实际问题的知识库与系统，达到工程目标
- 研究特点
  - 不可预测性

决策树学习

- 决策树(Decision Tree)

  一种描述概念空间的有效的归纳推理方法。可进行不相关的多概念学习，简单快捷，已在各领域广泛应用。

决策树学习（概述）

- 是以实例为基础的归纳学习。
- 从一类无序、无规则的事物（概念）中推理出决策树表示的分类规则。
- 概念分类学习算法：来源于
  - Hunt,Marin和Stone 于1966年研制的CLS学习系统，用于学习单个概念。
  - 1979年，J.R. Quinlan 给出ID3算法，并在1983年和1986年对ID3 进行了总结和简化，使其成为决策树学习算法的典型。
  - Schlimmer 和Fisher 于1986年对ID3进行改造，在每个可能的决策树节点创建缓冲区，使决策树可以递增式生成，得到ID4算法。
  - 1988年，Utgoff 在ID4基础上提出了ID5学习算法，进一步提高了效率。
  - 1993年，Quinlan 进一步发展了ID3算法，改进成C4.5算法。
  - 另一类决策树算法为CART，与C4.5不同的是，CART的决策树由二元逻辑问题生成，每个树节点只有两个分枝，分别包括学习实例的正例与反例。
- 基本思想是以信息熵为度量构造一棵熵值下降最快的树，到叶子节点处的熵值为零，此时每个叶节点中的实例都属于同一类。

决策树学习（概述）

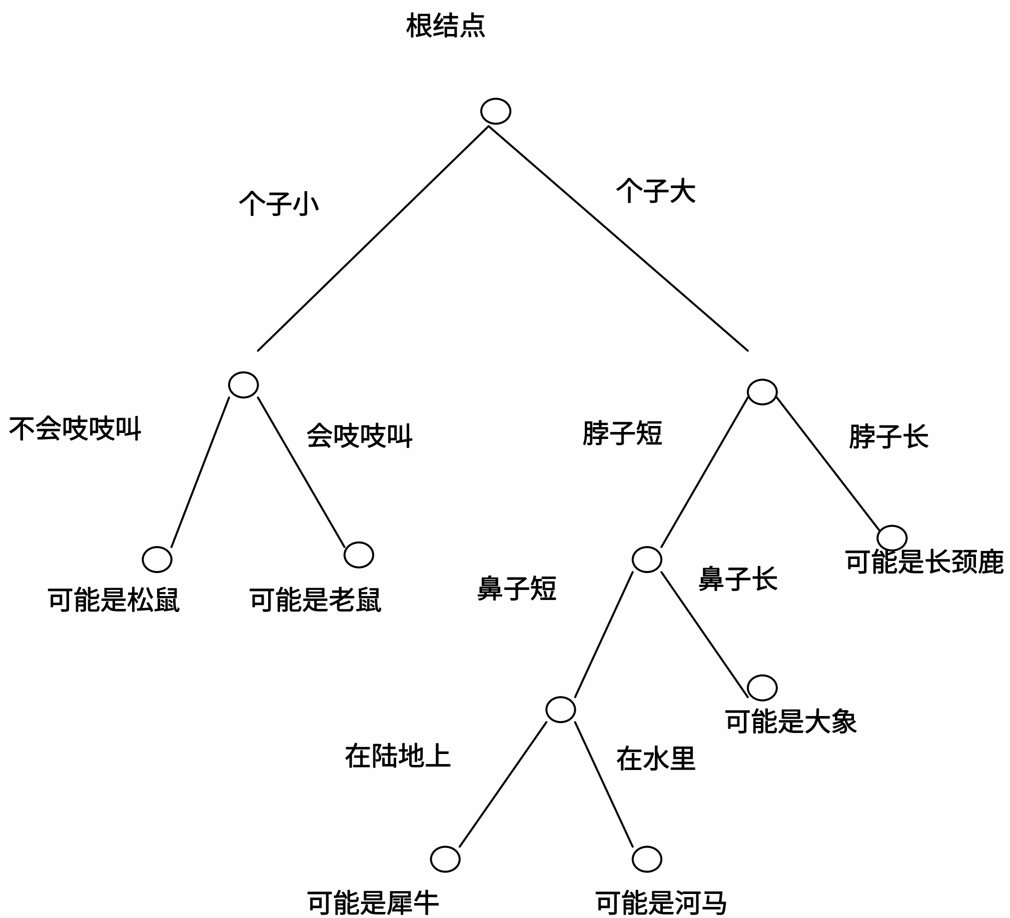
- 当前包括C4.5和CART的各种算法得到进一步改进。
- 比较引人注目的有斜超平面分割的多变决策树(Multi-Variance Decision Tree， MDT)算法；将遗传算法、神经元网络和C4.5相结合的GA-NN-C4.5算法；SVM决策树算法。
- 这些改进算法旨在结合各种方案的优势，取得更合理的分类效果，总结出更通用的规则。

决策树学习（概述）

- 采用自顶向下的递归方法。
- 每层节点依照某一属性值向下分为子节点，实例在每一节点处与相应属性值进行比较，根据结果向下扩展，这一过程在到达决策树的叶节点时结束，此时得到结论。
- 从根到叶节点的每一条路径都对应着一条合理的规则，规则间各部分是合取关系。整个决策树对应着一组析取的规则。
- 算法的最大优点是：可自学习。无需使用者了解过多背景知识，只需对训练例子进行较好的标注；若在应用中发现不符合规则的实例，程序会询问用户该实例的正确分类，从而生成新的分枝和叶子，并添加到树中。

决策树学习（决策树）

- 树是由节点和分枝组成的层次数据结构。
- 节点用于存贮信息或知识，分枝用于连接各个节点。
- 决策树是描述分类过程的一种数据结构，从根节点开始，各种分类原则被引进，将节点的数据集划分为子集，这一划分过程直到某种约束条件满足而结束。



决策树学习（决策树）

- 决策树的内部结点包含学习的实例，每层分枝代表了实例的一个属性的可能取值，叶节点是最终划分成的类。
- 如果判定是二元的，那么构造的将是一棵二叉树，在树中每回答一个问题就降到树的下一层，这类树一般称为CART（Classification And Regression Tree）。
- 判定结构可以机械的转变成产生式规则。可以通过对结构进行广度优先搜索，并在每个节点生成“IF…THEN”规则来实现。

IF “个子大” THEN

IF “脖子短” THEN

IF “鼻子长” THEN 可能是大象

形式化表示成

$$\text{个子大} \wedge \text{脖子短} \wedge \text{鼻子长} \Rightarrow \text{可能是大象}$$

决策树学习（决策树）

- 构造决策树的四个问题：
  - 收集待分类的数据，其所有属性应完全标注。
  - 设计分类原则，即哪些属性可用来分类、如何将属性量化；
  - 分类原则的选择，即每一步选择哪一准则使最终的树更令人满意。
  - 设计分类停止条件，实际应用中属性很多，真正有分类意义的属性往往是有限几个，因此在必要的时候应该停止数据集分裂：
    - 该节点包含的数据太少不足以分裂，
    - 继续分裂数据集对树生成的目标(例如ID3中的熵下降准则)没有贡献，
    - 树的深度过大不宜再分。
- 通用的决策树分裂目标是整棵树的熵总量最小，每一步分裂时，选择使熵减小最大的准则，这种方案使最具有分类潜力的准则最先被提取出来

决策树学习（性质）

- 证据由属性值对表示
  - 证据由固定的的属性和其值表示，如属性（温度），值（热）最简单的学习情况时每个属性拥有少量的不相关的值。
- 目标函数有离散输出值
  - 决策树分配一个二值的树，很容易扩展成为多于两个的输出值。
- 需要不相关的描述
  - 决策树原则上是表述不相关的表示
- 容忍训练数据的错误
  - 对训练样本和表述样本的属性值的错误都有较强的鲁棒性。
- 训练数据可以缺少值
  - 可以采用缺少属性值的样本学习。（不是所有样本都有）

决策树学习（应用）

- 根据病情对病人分类
- 根据起因对故障分类
- 根据付款信用情况对贷款申请者分类

这些都是将输入样本分类成可能离散集

分类问题

决策树学习（学习）

Shannon信息熵

– 自信息量

设信源X发出a_i的概率$p(a_i)$，a_i的自信息量$I(a_i)$为：$I(a_i) = -\log p(a_i)$。

– 信息熵

信息熵用来度量整个信源的不确定性，定义为：

$$H(X) = p(a_1)I(a_1) + p(a_2)I(a2) + \ldots + p(a_r)I(a_r) = -\sum_{i=1}^{r} p(a_i)\log p(a_i)$$

– 条件熵

设信源为X，收信者收到信息Y，用条件熵H(X|Y)来描述收信者在收到Y后对X的不确定性估计。$p(a_i|b_j)$为当Y为b_j时，X为a_i的概率，则有：

$$H(X|Y) = -\sum_{j=1}^{s} p(b_j)\sum_{i=1}^{r} p(a_i|b_j)\log p(a_i|b_j)$$

– 平均互信息量

用平均互信息量$I(X, Y)$来表示信号Y所能提供的关于X的信息量的大小：

$$I(X,Y) = H(X) - H(X|Y)$$

A typical decision table--classification

Positive Examples

Negative Examples

ID	size	hair	eyes	group
1	Low	Gold	Blue	1
2	High	Red	Blue	1
3	High	Gold	Blue	1
4	Low	Gold	Gray	1
5/-1	High	Gold	Black	2
6/-2	Low	Black	Blue	2
7/-3	High	Black	Blue	2
8/-4	High	Black	Gray	2
9/-5	Low	Gold	Black	2

Concept Learning System/Decision Tree

决策表(decision table)是一个四元组:

$$L = (U, A, F, V), U = \{x_1, \cdots x_n\}, A = C \cup D,$$
$$F : U \times A \to V$$

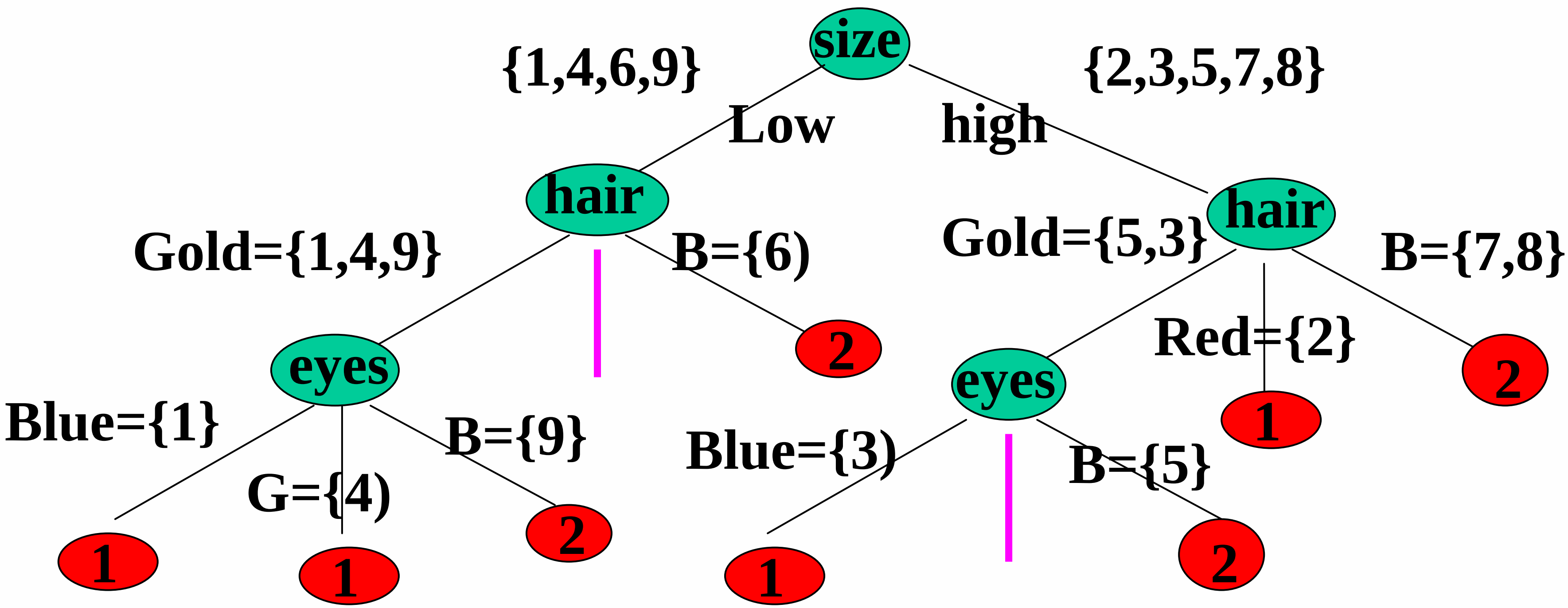
U：论域；

A：由两个互不相交的子集C和D构成的属性集,C称为条件属性；D称为决策属性；

V：属性域，表中每个元素对应于一对：属性—值；

一个 CLS是生成决策树的迭代过程.

A Decision tree for the "Grouping" problem,with attribute Size as the root



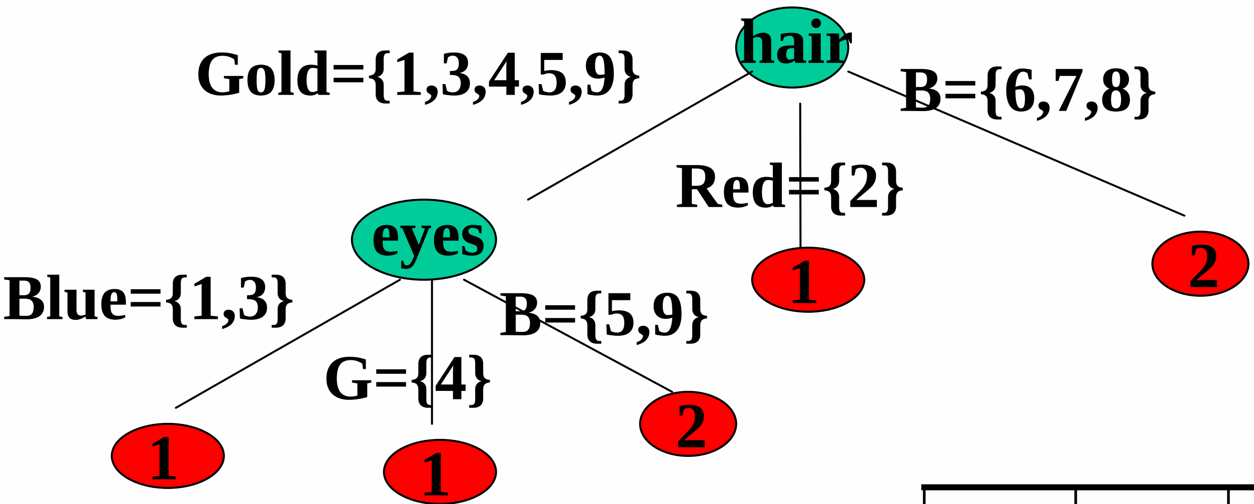
ID	size	hair	eyes	group
1	Low	Gold	Blue	1
2	High	Red	Blue	1
3	High	Gold	Blue	1
4	Low	Gold	Gray	1
5/-1	High	Gold	Black	2
6/-2	Low	Black	Blue	2
7/-3	High	Black	Blue	2
8/-4	High	Black	Gray	2
9/-5	Low	Gold	Black	2

Rule set generated by tracing back from a leaf to the root:

8 rules, in total;

1. **if [size=low] ∧[hair=Gold]∧[eyes=Blue], then Group=1；**
2. **If [size=low] ∧[hair=Gold]∧[eyes=Gray], then Group=1;**
3. **If [size=low] ∧[hair=Gold]∧[eyes=Black], then Group=2;**
4. **If …..to be finished by students!**
5. **If…..**
6. **If**
7. **If**
8. **If**

A smaller Decision tree for the “Grouping” problem,with attribute Hair as the root



5 rules; Length=8;

ID	size	hair	eyes	group
1	Low	Gold	Blue	1
2	High	Red	Blue	1
3	High	Gold	Blue	1
4	Low	Gold	Gray	1
5/-1	High	Gold	Black	2
6/-2	Low	Black	Blue	2
7/-3	High	Black	Blue	2
8/-4	High	Black	Gray	2
9/-5	Low	Gold	Black	2

Rule set generated by tracing back from a leaf to the root:

5 rules, Length=8;

(against:covering method,6 rules, Length=10)

(something new observed??)

1. **if [hair=Gold]∧[eyes=Blue], then Group=1**
2. **If [hair=Gold]∧[eyes=Gray], then Group=1**
3. **If [hair=Gold]∧[eyes=Black], then Group=2**
4. **If [hair=Red] ,then Group=1**
5. **If[hair=Black], then Group=2**

设决策属性D将论域U分成m类，记为:

$$U / ind(D) = \{Y_1, Y_2, \cdots, Y_m\}$$

D的信息熵定义为:

$$\boxed{Entropy(D, U) = -\sum_{j=1}^{m} p(Y_j) \log_2 p(Y_j) = Entropy(D)}$$

令$R \subseteq C, U / ind(R) = \{X_1, X_2, \cdots, X_n\}$,则D关于R的条件熵定义为：

$$\boxed{Entropy(D \mid R) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log_2 (p(Y_j \mid X_i)) = \sum_{i=1}^{n} p(X_i) Entropy(D, X_i)}$$

where

$$\begin{cases} p(X_i\} = \dfrac{|X_i|}{|U_i|} \\ 0 \le p(Y_j \mid X_i\} = \dfrac{|Y_j \cap X_i|}{|X_i|} \le 1 \end{cases}$$

$$Entropy\ (D \mid R) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log_2 (p(Y_j \mid X_i)) = \sum_{i=1}^{n} p(X_i) Entropy\ (D, X_i)$$

$$\text{if}\ \ X_i \subseteq Y_k\ (\text{for some}\ \ k),\ \text{then}\ \ p(Y_k \mid X_i\} = \frac{|Y_k \cap X_i|}{|X_i|} = 1;$$

$$\Rightarrow \sum_{j=1}^{m} p(Y_j \mid X_i) \log_2 (p(Y_j \mid X_i) = 0$$

$$\text{if for each}\ \ X_i, \exists Y_k \Rightarrow X_i \subseteq Y_k (\text{ some}\ \ k),\ \text{then}\ \ p(Y_k \mid X_i\} = \frac{|Y_k \cap X_i|}{|X_i|} = 1;$$

$$\Rightarrow Entropy\ (D \mid R) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j \mid X_i) \log_2 (p(Y_j \mid X_i)) = 0$$

令 Info = *Entropy(D)* - *Entropy(D|R)*

Info 称为“信息增益”或“平均互信息量”;

在经典决策树算法中，R取为单个属性，称为测试属性

如果

$$Entropy(D|R) = 0$$

或等价地

$$\text{Info} = Entropy(D) - Entropy(D|R) = Entropy(D)$$

- 表明由测试属性R的值所确定的每一类均被整个包含在决策属性所确定的某一类中；意味着相对于决策属性此时的测试属性是“纯”的．该测试属性提供了最大的信息增益，
- 可以通过计算每个条件属性相对于决策属性的平均互信息量来确定其“纯度”，帮助选择决策树的节点构成顺序

ID3 for the "Grouping"

ID3-Tree

$$U/IND\{hair\} = (X1 = \{1,3,4,5,9\}, X2 = \{2\}, X3 = \{6,7,8\})$$

$$U/IND\{size\}) = (X1 = \{1,4,6,9\}, X2 = \{2,3,5,7,8\})$$

$$U/IND\{eye\}) = (X1 = \{1,2,3,6,7\}, X2 = \{4,8\}, X3 = \{5,9\})$$

ID	size	hair	eyes	group
1	Low	Gold	Blue	1
2	High	Red	Blue	1
3	High	Gold	Blue	1
4	Low	Gold	Gray	1
5/-1	High	Gold	Black	2
6/-2	Low	Black	Blue	2
7/-3	High	Black	Blue	2
8/-4	High	Black	Gray	2
9/-5	Low	Gold	Black	2

$$E(D|\{hair\}) = -\frac{5}{9}(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}) - \frac{1}{9}\log 1 - \frac{3}{9}\log 1$$

$$= \underleftrightarrow{-\frac{5}{9}(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5})}$$

$$E(D|\{size\}) = -\frac{4}{9}(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}) - \underleftrightarrow{\frac{5}{9}(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5})}$$

$$E(D|\{eye\}) = -\frac{5}{9}(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}) - \frac{2}{9}(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}) - \frac{2}{9}\log 1$$

$$= \underleftrightarrow{-\frac{5}{9}(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5})} - \frac{2}{9}(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2})$$

$$\begin{cases} p(X_i\} = \dfrac{|X_i|}{|U_i|} \\ 0 \le p(Y_j | X_i\} = \dfrac{|Y_j \cap X_i|}{|X_i|} \le 1 \end{cases}$$

$$E(D|\{hair\}) = Minimum \Rightarrow gain(hair) = Max$$

$$\boxed{Entropy(D|R) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j | X_i) \log_2(p(Y_j | X_i)) = 0}$$

防雹部门在雹灾来临前决定采取何种方案以减轻灾害，

序号	属性值				决策
	x_1	x_2	x_3	x_4	μ
1	1	1	1	1	3
2	1	1	1	2	2
3	1	1	2	1	3
4	1	1	2	2	1
5	1	2	1	1	3
6	1	2	1	2	2
7	1	2	2	1	3
8	1	2	2	2	1
9	2	1	1	1	3
10	2	1	1	2	2
11	2	1	2	1	3
12	2	1	2	2	1

序号	属性值				决策
	x_1	x_2	x_3	x_4	μ
13	2	2	1	1	3
14	2	2	1	2	2
15	2	2	2	1	3
16	2	2	2	2	3
17	3	1	1	1	3
18	3	1	1	2	3
19	3	1	2	1	3
20	3	1	2	2	1
21	3	2	1	1	3
22	3	2	1	2	2
23	3	2	2	1	3
24	3	2	2	2	3

决策类：

μ_1：用大口径防雹炮

μ_2：用小口径防雹炮

μ_3：不能用防雹炮

属性：

x_1：云层性质；雨云（1）、浓积云（2）、雨层云（3）

x_2：风速；微风（1）、大风（2）

x_3：季节；夏秋（1）、春冬（2）

x_4：地理条件；向风坡（1）、背风坡（2）

$U / IND\{x_1\} =$

$(X1 = \{1,2,3,4,5,6,7,8\},$

$X2 = \{9,10,11,12,13,14,15,16\},$

$X3 = \{17,18,19,20,21,22,23,24\}); |X1| = 8, |X2| = 8, |X3| = 8;$

$U / IND\{x_2\}) =$

$(X1 = \{1,2,3,4,9,10,11,12,17,18,19,20\},$

$X2 = \{5,6,7,8,13,14,15,16,21,22,23,24\}); |X1| = 12, |X2| = 12;$

$U / IND\{x_3\}) =$

$(X1 = \{1,2,5,6,9,10,13,14,17,18,21,22\},$

$X2 = \{3,4,7,8,11,12,15,16,19,20,23,24\}); |X1| = 12, |X2| = 12;$

$U / IND\{x_4\}) =$

$(X1 = \{1,3,5,7,9,11,13,15,17,19,21,23\},$

$X2 = \{2,4,6,8,10,12,14,16,18,20,22,24\}); |X1| = 12, |X2| = 12;$

$$Entropy(D|R) = -\sum_{i=1}^{n} p(X_i)\sum_{j=1}^{m} p(Y_j|X_i)\log_2(p(Y_j|X_i))$$

$$\begin{cases} p(X_i\} = \dfrac{|X_i|}{|U_i|} \\ 0 \le p(Y_j|X_i\} = \dfrac{|Y_j \cap X_i|}{|X_i|} \le 1 \end{cases}$$

$$\begin{aligned} E(D|\{x_1\}) = &-\frac{8}{24}\left(\frac{2}{8}\log\frac{2}{8}+\frac{2}{8}\log\frac{2}{8}+\frac{4}{8}\log\frac{4}{8}\right) \\ &-\frac{8}{24}\left(\frac{1}{8}\log\frac{1}{8}+\frac{2}{8}\log\frac{2}{8}+\frac{5}{8}\log\frac{5}{8}\right) \\ &-\frac{8}{24}\left(\frac{1}{8}\log\frac{1}{8}+\frac{1}{8}\log\frac{1}{8}+\frac{6}{8}\log\frac{6}{8}\right) \end{aligned}$$

$$= \frac{8}{24}\times 1.5 + \frac{8}{24}\times 1.2988 + \frac{8}{24}\times 1.0613$$

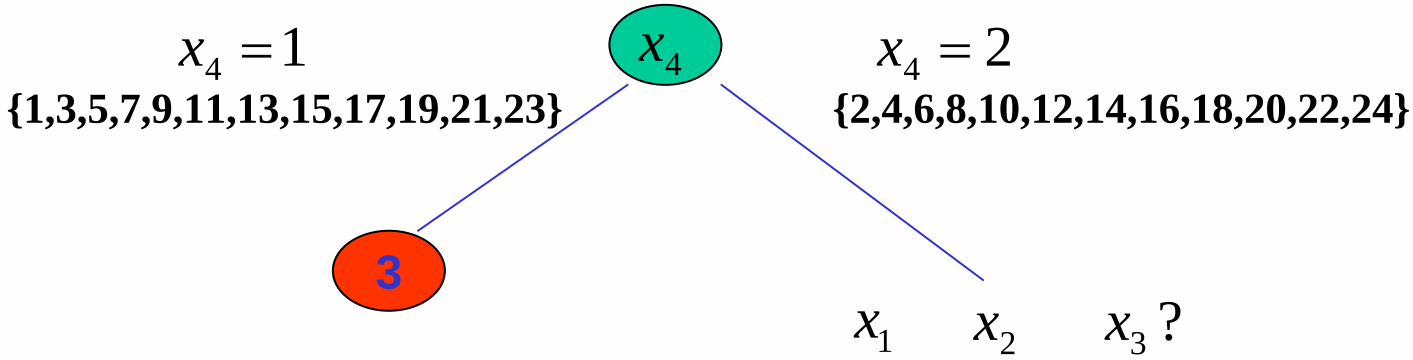
$$= 1.2867$$

$$E(D|\{x_2\}) = 1.2866$$

$$E(D|\{x_3\}) = 0.9491$$

$$E(D|\{x_4\}) = 0.7773$$

$$\boxed{E(D|\{x_4\}) \longrightarrow \text{条件熵}\min \longrightarrow \text{信息增量}\max}$$


$x_4 = 1$
{1,3,5,7,9,11,13,15,17,19,21,23}
x_4
$x_4 = 2$
{2,4,6,8,10,12,14,16,18,20,22,24}
3
x_1 x_2 x_3 ?

$$U_1 = U \setminus \{1,3,5,7,9,11,13,15,1,719,21,23\}$$
$$= \{2,4,6,8,10,12,14,16,18,20,22,24\}$$
$$U_1 / IND\{x_1\} =$$
$$(X1 = \{2,4,6,8\},$$
$$X2 = \{10,12,14,16\},$$
$$X3 = \{18,20,22,24\}); |X1| = 4, |X2| = 4, |X3| = 4;$$
$$U_1 / IND\{x_2\}) =$$
$$(X1 = \{2,4,10,12,18,20\},$$
$$X2 = \{6,8,14,16,22,24\}); |X1| = 6, |X2| = 6;$$
$$U_1 / IND\{x_3\}) =$$
$$(X1 = \{2,6,10,14,18,22\},$$
$$X2 = \{4,8,12,16,20,24\}); |X1| = 6, |X2| = 6;$$

$$Entropy(D|R) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j|X_i) \log_2(p(Y_j|X_i))$$

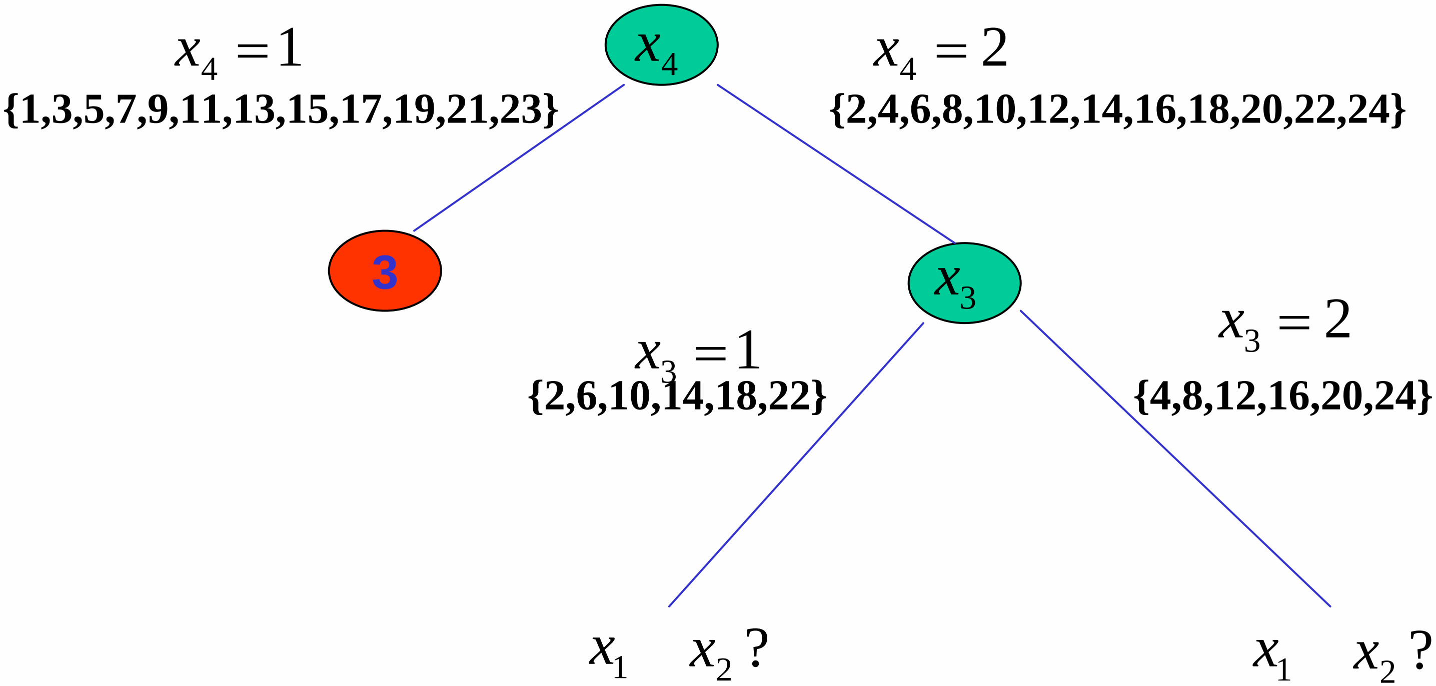
$$\begin{cases} p(X_i\} = \dfrac{|X_i|}{|U_i|} \\ 0 \le p(Y_j|X_i\} = \dfrac{|Y_j \cap X_i|}{|X_i|} \le 1 \end{cases}$$

$$\begin{aligned} E(D|\{x_1\}) &= -\frac{4}{12}(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}) \\ &\quad -\frac{4}{12}(\frac{1}{4}\log_2\frac{1}{4} + \frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4}) \\ &\quad -\frac{4}{12}(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{2}{4}\log_2\frac{2}{4}) \\ &= \frac{4}{12}\times 1 + \frac{4}{12}\times 1.5 + \frac{4}{12}\times 1.5 = \frac{4}{3} \end{aligned}$$

$$\begin{aligned} E(D|\{x_3\}) &= -\frac{6}{12}(0 + \frac{5}{6}\log_2\frac{5}{6} + \frac{1}{6}\log_2\frac{1}{6}) \\ &\quad -\frac{6}{12}(\frac{4}{6}\log_2\frac{4}{6} + 0 + \frac{2}{6}\log_2\frac{2}{6}) \\ &= \frac{6}{12}\times 0.65 + \frac{6}{12}\times 0.9183 = 0.7841 \end{aligned}$$

$$\begin{aligned} E(D|\{x_2\}) &= -\frac{6}{12}(\frac{3}{6}\log_2\frac{3}{6} + \frac{2}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{1}{6}) \\ &\quad -\frac{6}{12}(\frac{1}{6}\log_2\frac{1}{6} + \frac{3}{6}\log_2\frac{3}{6} + \frac{2}{6}\log_2\frac{2}{6}) \\ &= \frac{6}{12}\times 1.4591 + \frac{6}{12}\times 1.4591 = 1.4591 \end{aligned}$$

$E(D|\{x_3\}) \longrightarrow$ 条件熵min $\longrightarrow$ 信息增量max


$x_4 = 1$
{1,3,5,7,9,11,13,15,17,19,21,23}
x_4
$x_4 = 2$
{2,4,6,8,10,12,14,16,18,20,22,24}
3
x_3
$x_3 = 1$
{2,6,10,14,18,22}
$x_3 = 2$
{4,8,12,16,20,24}
x_1 x_2 ?
x_1 x_2 ?

第六章 机器学习

The End.