# 人工智能：不确定知识表示和推理 I


饶洋辉

计算机学院,

中山大学

raoyangh@mail.sysu.edu.cn

http://cse.sysu.edu.cn/node/2471


课件来源：中山大学刘咏梅教授；多伦多大学Sheila McIlraith教授；浙江大学吴飞教授；海军工程大学贲可荣教授等

# 目录

- 背景
- 贝叶斯网络的表示
- 贝叶斯网络的构建
- 贝叶斯网络的参数学习
- 贝叶斯网络中的D-分离
- 贝叶斯网络的推理

# 背景

- 人工智能系统的表示与推理过程实际上就是一种思维过程。其中，推理是从已知事实出发，通过运用相关的知识逐步推出某个结论的过程。人们常常对原因和结果的推理很感兴趣。比如，我感冒症状减轻了，是不是因为服用了维生素C片导致的？但是，由于事件都带有随机性，导致看似直截了当的问题，却不容易回答。例如，我可能仅仅喝白开水，感冒也会自己消失。
  - 已知事实（证据），用以指出推理的出发点及推理时应使用的知识；
  - 知识是推理得以向前推进，并逐步达到最终目标的依据。
- 按照所用知识的确定性，可以分为确定性和不确定性两种类别。
  - 确定性推理是建立在经典逻辑基础上的，经典逻辑的基础之一就是集合论。这在很多实际情况中是很难做到的，如高、矮、胖、瘦就很难精确地分开；
  - 不确定性推理就是从不确定性初始证据出发，通过运用不确定性的知识，最终推出具有一定程度的不确定性但却是合理或者近乎合理的结论的思维过程。

# 背景

- 常识（common sense）具有不确定性。
  - 一个常识可能有众多的例外，一个常识可能是一种尚无理论依据或者缺乏充分验证的经验。
- 常识往往对环境有极强的依存性。
  - “鸟是会飞的”
- 把指示确定性程度的数据附加到推理规则，并由此研究不确定强度的表示和计算问题。
- 处理数据的不精确和知识的不确定所需要的一些工具和方法，包括：
  - **基于Bayes理论的概率推理**
  - 基于信任测度函数的证据理论
  - 基于模糊集合论的模糊推理等

# 背景

- 不确定知识表示与推理是人工智能的核心模块之一，其理论基础包括概率论和可能性理论等。
  - 概率论处理的是由随机性引起的不确定性；
  - 可能性理论处理的是由模糊性引起的不确定性。
- 李德毅院士在统一主观认知和客观现象中的随机性和模糊性方面提出了不确定性人工智能的研究问题。不确定性人工智能认为，随机性和模糊性常常是联系在一起的，在人类思维和智能行为中难以区分并独立存在，研究不确定性需要研究随机性和模糊性之间的关联性。
- 本讲主要以概率论为基础，涵盖贝叶斯网络的表示、构建和参数学习，以及D-分离等内容。

# 背景

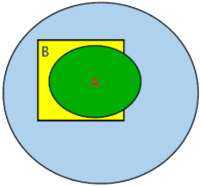
- **频率论学派**
  - 事件的概率是当我们无限次重复试验时，事件发生次数的比值
  - 投掷硬币、掷骰子等

- **贝叶斯学派**
  - 将事件的概率视为一种主观置信度
  - 我认为明天下雨的概率是30%；他认为明天下雨的概率是80%

Thomas Bayes（约1701-1761），英国数学家。约1701年出生于伦敦，1742年成为英国皇家学会会员。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。

# 背景

- **条件概率**
  - 假设变量$A$表示一个事件的集合，且$Pr(A) > 0$
  - 在变量$A$发生的前提下，另一个事件的集合（用变量$B$表示）发生的概率记为条件概率$Pr(B|A)$




- 假设$B$覆盖了整个事件集合空间的30%，且覆盖了$A$的80%，那么$Pr(B) = 30\%$，且$Pr(B|A) = 80\%$
- 如果$Pr(B|A) = Pr(B)$，则$B$和$A$独立
- 如果$Pr(B|A,C) = Pr(B|A)$，则给定$A$的前提下，$B$和$C$条件独立

# 背景

- **乘法/链式法则**

联合概率$Pr(A,B)=Pr(A)Pr(B|A)=Pr(B,A)=Pr(B)Pr(A|B)$

$Pr(A,B_1,B_2,B_3)=Pr(A)Pr(B_1|A)Pr(B_2|A,B_1)Pr(B_3|A,B_1,B_2)$

$Pr(\text{Grade} = \text{A} \mid \text{Student} = \text{Smart}) = 0.6$

$Pr(\text{Grade} = \text{A}) = 0.2$

$Pr(\text{Student} = \text{Smart}) = 0.3$

$Pr(\text{Student} = \text{Smart} \mid \text{Grade} = \text{A}) = ?$

# 背景

- **乘法/链式法则**

联合概率$Pr(A,B)=Pr(A)Pr(B|A)=Pr(B,A)=Pr(B)Pr(A|B)$

$Pr(A,B_1,B_2,B_3)=Pr(A)Pr(B_1|A)Pr(B_2|A,B_1)Pr(B_3|A,B_1,B_2)$

$Pr(\text{Grade} = \text{A} \mid \text{Student} = \text{Smart}) = 0.6$

$Pr(\text{Grade} = \text{A}) = 0.2$

$Pr(\text{Student} = \text{Smart}) = 0.3$

$Pr(\text{Student} = \text{Smart} \mid \text{Grade} = \text{A}) = 0.9$

# 背景

- **加法法则**

$Pr(A) = Pr(A,B) + Pr(A,B^c)$

$=Pr(B)Pr(A|B) + Pr(B^c)Pr(A|B^c)$

$$Pr(A) = \sum_B Pr(A,B) = \sum_{i=1}^{n} Pr(A,B_i)$$

$$= \sum_{i=1}^{n} Pr(A|B_i)Pr(B_i)$$

# 背景

- **加法法则**

$Pr(A) = Pr(A,B) + Pr(A,B^c)$

$=Pr(B)Pr(A|B) + Pr(B^c)Pr(A|B^c)$

$$Pr(A) = \sum_B Pr(A,B) = \sum_{i=1}^{n} Pr(A,B_i)$$

$$= \sum_{i=1}^{n} Pr(A|B_i)Pr(B_i)$$

- **贝叶斯定理**

$$Pr(B|A) = \frac{Pr(A,B)}{Pr(A)}$$

$$= \frac{Pr(B)Pr(A|B)}{Pr(A)}$$

$$= \frac{Pr(B)Pr(A|B)}{Pr(A,B) + Pr(A,B^c)}$$

# 背景

- **加法法则**

$Pr(A) = Pr(A,B) + Pr(A,B^c)$

$= Pr(B)Pr(A|B) + Pr(B^c)Pr(A|B^c)$

$$Pr(A) = \sum_B Pr(A,B) = \sum_{i=1}^{n} Pr(A,B_i)$$

$$= \sum_{i=1}^{n} Pr(A|B_i)Pr(B_i)$$

- **贝叶斯定理**

后验概率

$$Pr(B|A) = \frac{Pr(A,B)}{Pr(A)}$$

先验概率 似然度

$$= \frac{\boxed{Pr(B)}\boxed{Pr(A|B)}}{\boxed{Pr(A)}}$$

标准化常量

$$= \frac{Pr(B)Pr(A|B)}{Pr(A,B) + Pr(A,B^c)}$$

# 背景

- 假设有一盒骰子，里面有4面的（点数为1、2、3、4），6面的、8面的、12面的、20面的均匀骰子各1个。如果我随机从盒子中选一个骰子，投掷它得到了点数5。那么我选中的骰子为4面、6面、8面、12面、20面的概率各是多少?

# 背景

- 假设有一盒骰子，里面有4面的（点数为1、2、3、4），6面的、8面的、12面的、20面的均匀骰子各1个。如果我随机从盒子中选一个骰子，投掷它得到了点数5。那么我选中的骰子为4面、6面、8面、12面、20面的概率各是多少?

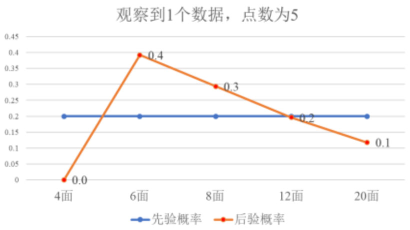
$$Pr(\text{骰子}=4 \mid \text{点数}=5) = \frac{Pr(\text{骰子}=4)Pr(\text{点数}=5 \mid \text{骰子}=4)}{Pr(\text{点数}=5)} = \frac{0.2\times 0}{Pr(\text{点数}=5)}$$

$$Pr(\text{骰子}=6 \mid \text{点数}=5) = \frac{Pr(\text{骰子}=6)Pr(\text{点数}=5 \mid \text{骰子}=6)}{Pr(\text{点数}=5)} = \frac{0.2\times \frac{1}{6}}{Pr(\text{点数}=5)}$$

…

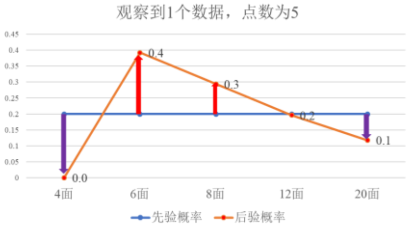
# 背景

- 假设有一盒骰子，里面有4面的（点数为1、2、3、4），6面的、8面的、12面的、20面的均匀骰子各1个。如果我随机从盒子中选一个骰子，投掷它得到了点数5。那么我选中的骰子为4面、6面、8面、12面、20面的概率各是多少?



# 背景

- 假设有一盒骰子，里面有4面的（点数为1、2、3、4），6面的、8面的、12面的、20面的均匀骰子各1个。如果我随机从盒子中选一个骰子，投掷它得到了点数5。那么我选中的骰子为4面、6面、8面、12面、20面的概率各是多少?




置信度发生改变!

# 小结

- 概率推理是一类在不确定性知识的基础上，基于概率方法进行推理、或近似推理的技术。
- 频率论学派（Frequentists） → 频率论推理
- 贝叶斯学派（Bayesians）→ 贝叶斯推理
- 频率论学派认为频率论是获得可靠推理的一种统计方法，称其为统计推理（statistical inference）。该学派提出了两种密切相关的方法：内曼-皮尔森（Neyman-Pearson）的显著性检验理论（Theory of significance tests）和费雪（Fisher）的*p*值（*p*-values）。
- 贝叶斯推理基于贝叶斯定理，是现代概率推理的基础。
- 简单的因果关系可采用贝叶斯推理，而复杂的关系则需要基于贝叶斯网络。

# 目录

- 背景
- 贝叶斯网络的表示
- 贝叶斯网络的构建
- 贝叶斯网络的参数学习
- 贝叶斯网络中的D-分离
- 贝叶斯网络的推理

# 贝叶斯网络的表示

- 贝叶斯网络（Bayesian Network，简称BN）是不确定知识表示与推理的一种有效方法，它由一个有向无环图（表达了变量之间的有向依赖关系）和一系列条件概率表（衡量了上述关系的强度）组成。

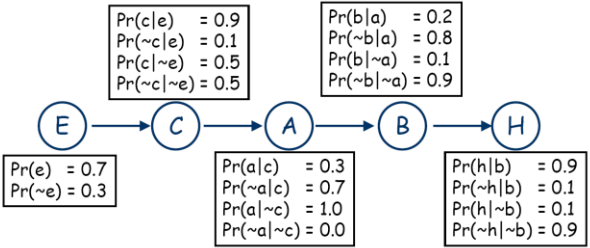
有向无环图指的是一个无回路的有向图，即从图中任意一个节点出发经过任意条边，均无法回到该节点。有向无环图刻画了图中所有节点之间的依赖关系。

- 1986年，Judea Pearl提出贝叶斯网络，用于描述不确定知识表示与推理问题。贝叶斯网络让机器可以回答问题——给出一个从非洲回来的发烧且身体疼痛的病人，最有可能的解释是疟疾。

Judea Pearl（朱迪亚 · 佩尔），贝叶斯网络之父，加州大学洛杉矶分校计算机科学学院教授、认知系统实验室主任。2011年，因人工智能概率方法和因果推理算法获得图灵奖。

# 贝叶斯网络的表示

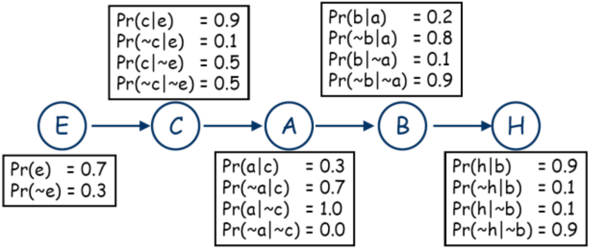
- 示例:




E - Craig woke too early  C - Craig needs coffee  A - Craig is angry
B - Craig burst a blood vessel  H - Craig hospitalized

# 贝叶斯网络的表示

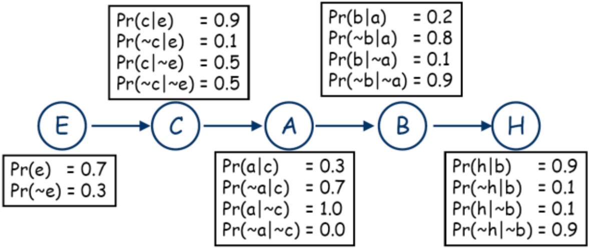
- 示例:




从上图中我们可以得知，Craig doesn’t get sent to the hospital because he’s angry, he gets sent because he had a burst blood vessel.

# 贝叶斯网络的表示

- 示例：

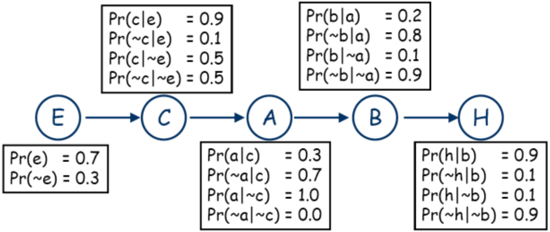



$Pr(H,B,A,C,E) = Pr(H|B,A,C,E)Pr(B|A,C,E)Pr(A|C,E)Pr(C|E)P(E)$ 31个参数

$Pr(H,B,A,C,E) = Pr(H|B)Pr(B|A)Pr(A|C)Pr(C|E)P(E)$ 9个参数

# 贝叶斯网络的表示

- 示例：




$Pr(c) = Pr(c|e)Pr(e) + Pr(c|\sim e)Pr(\sim e)$

$\quad = 0.9 * 0.7 + 0.5 * 0.3 = 0.78$

$Pr(\sim c) = Pr(\sim c|e)Pr(e) + Pr(\sim c|\sim e)Pr(\sim e) = 0.22$

- $Pr(\sim c) = 1 - Pr(c)$, as well

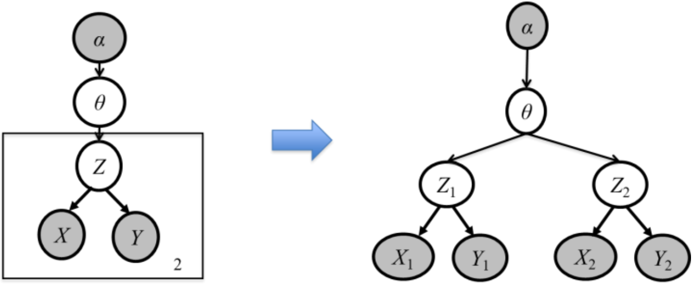
$Pr(a) = Pr(a|c)Pr(c) + Pr(a|\sim c)Pr(\sim c)$

$\quad = 0.3 * 0.78 + 1.0 * 0.22 = 0.454$

$Pr(\sim a) = 1 - Pr(a) = 0.546$

# 贝叶斯网络的表示

- 盘式记法（plate notation）
  - 贝叶斯网络的一种简洁的表示方法，其将相互独立的、由相同机制生成的多个变量放在一个方框（盘）内，并在方框中标出类似变量重复出现的个数。此外，方框可以嵌套，且通常用阴影标注出可观察到的变量。
  - 示例：



# 目录

- 背景
- 贝叶斯网络的表示
- 贝叶斯网络的构建
- 贝叶斯网络的参数学习
- 贝叶斯网络中的D-分离
- 贝叶斯网络的推理

# 贝叶斯网络的构建

- 给定变量$\{X_1,\ldots,X_n\}$，构建一个贝叶斯网络的步骤如下：
  - 步骤1：在某种变量顺序下，对所有变量的联合概率应用链式法则

    $$Pr(X_1,\ldots,X_n) = Pr(X_n|X_1,\ldots, X_{n-1})Pr(X_{n-1}|X_1,\ldots, X_{n-2})\ldots Pr(X_1)$$

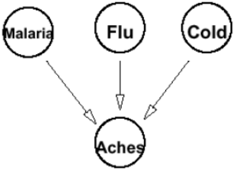
  - 步骤2：对于每个变量$X_i$，考虑该变量的条件集合$X_1,\ldots,X_{i-1}$，采用如下方法递归地判断条件集合中的每个变量$X_j$是否可以删除：如果给定其余变量的集合，$X_i$和$X_j$是条件独立的，则将$X_j$从$X_i$的条件集合中删除。经过这一步骤，可以得到下式

    $$Pr(X_1,\ldots,X_n) = Pr(X_n|Par(X_n))Pr(X_{n-1}|Par(X_{n-1}))\ldots Pr(X_1)$$

  - 步骤3：基于上述公式，构建一个有向无环图。其中，对于每个用节点表示的变量$X_i$，其父节点为$Par(X_i)$中的变量集合。
  - 步骤4：为每个家庭（即变量及其父节点集合）确定条件概率表的取值。

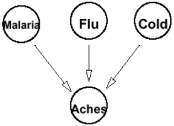
# 贝叶斯网络的构建

- 对于疟疾（M）、流感（F）、感冒（C）和疼痛（A）这四个变量，假设我们知道疟疾（M）、流感（F）和感冒（C）都会导致疼痛（A），那么在构建贝叶斯网络时，我们可以按照下述顺序完成步骤1：$Pr(M,F,C,A) = Pr(A|M,F,C)Pr(C|M,F)Pr(F|M)Pr(M)$
- 由于$M$、$F$和$C$都会导致$A$，所以变量$A$的条件集合不能删减。此外，我们通常认为这三种疾病（即$M$、$F$和$C$）的发生是独立的，因此，$Pr(C|M,F) = Pr(C)$，且$Pr(F|M) = Pr(F)$。据此构建的贝叶斯网络的有向无环图如下所示：



# 贝叶斯网络的构建

- $Pr(M,F,C,A) = Pr(A|M,F,C)Pr(C|M,F)Pr(F|M)Pr(M)$




| M |
|---|
| True |

| F |
|---|
| True |

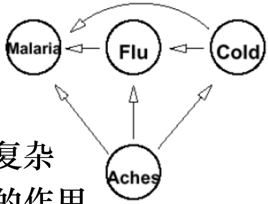
| C |
|---|
| True |

| M | F | C |
|---|---|---|
| True | True | True |
| True | True | False |
| True | False | True |
| True | False | False |
| False | True | True |
| False | True | False |
| False | False | True |
| False | False | False |

- 该贝叶斯网络的参数个数为：$2^3 + 1 + 1 + 1 = 11$

# 贝叶斯网络的构建

- 对于上述例子，如果我们按照另一种顺序完成步骤1：
  $Pr(M,F,C,A) = Pr(M|F,C,A)Pr(F|C,A)Pr(C|A)Pr(A)$
- 由于每个变量的条件集合都不能删减，据此构建的贝叶斯网络的有向无环图为：
- 该贝叶斯网络的参数个数为：
  $2^3 + 2^2 + 2 + 1 = 15$
- 与前面构建的贝叶斯网络相比，更为复杂
- 由此可见因果关系等先验知识在其中的作用



# 贝叶斯网络的构建

- 在分类等特定的任务中，为了能够从有限的训练样本中进行标签预测，有时也会忽略因果关系，人为假定一些条件独立性。
- 基于“属性条件独立性假设”的朴素贝叶斯分类模型是通过这种方式构建的一种特殊结构的贝叶斯网络，它在强（朴素）独立性假设的条件下运用贝叶斯定理来计算每个类别的条件概率。
- 虽然朴素贝叶斯分类模型的条件独立性假设太强，但在实际应用中，该模型在很多任务上也能得到很好的结果，并且模型简单，可以有效防止过拟合。

# 贝叶斯网络的构建

| 年龄 (A) | 收入 (I) | 学生 (S)? | 信用等级 (C)? | 是否买电脑 (B)? |
| --- | --- | --- | --- | --- |
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |
| <=30 | medium | yes | fair | ? |

后验概率$Pr(B = \text{yes} \mid A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair}) = ?$

# 贝叶斯网络的构建

| 年龄 (A) | 收入 (I) | 学生 (S)? | 信用等级 (C)? | 是否买电脑 (B)? |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |
| <=30 | medium | yes | fair | ? |

后验概率$Pr(B = \text{yes} \mid A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair}) = ?$

先验概率$Pr(B = \text{yes}) = 9/14 \approx 0.64$

似然度$Pr(A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair} \mid B = \text{yes}) = 0$

先验概率$Pr(B = \text{no}) = 5/14 \approx 0.36$

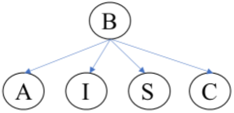
似然度$Pr(A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair} \mid B = \text{no}) = 0$

需要更大规模的训练数据！

# 贝叶斯网络的构建

| 年龄 (A) | 收入 (I) | 学生 (S)? | 信用等级 (C)? | 是否买电脑 (B)? |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |
| <=30 | medium | yes | fair | ? |

假设$Pr(A,I,S,C|B) = Pr(A|B)Pr(I|B)Pr(S|B)Pr(C|B)$，即给定$B$的条件下，$A$、$I$、$S$、$C$相互独立 $\Rightarrow$ **朴素贝叶斯分类**




该贝叶斯网络的参数个数为：1 + 2*2 + 2*2 + 2 + 2 = 13个

无上述假设的网络参数个数：3*3*2*2*2 - 1 = 71个

# 贝叶斯网络的构建

| 年龄 (A) | 收入 (I) | 学生 (S)? | 信用等级 (C)? | 是否买电脑 (B)? |
|---|---|---|---|---|
| <=30 | high | no | fair | no |
| <=30 | high | no | excellent | no |
| 31…40 | high | no | fair | yes |
| >40 | medium | no | fair | yes |
| >40 | low | yes | fair | yes |
| >40 | low | yes | excellent | no |
| 31…40 | low | yes | excellent | yes |
| <=30 | medium | no | fair | no |
| <=30 | low | yes | fair | yes |
| >40 | medium | yes | fair | yes |
| <=30 | medium | yes | excellent | yes |
| 31…40 | medium | no | excellent | yes |
| 31…40 | high | yes | fair | yes |
| >40 | medium | no | excellent | no |
| <=30 | medium | yes | fair | ? |

后验概率$Pr(B = \text{yes} \mid A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair}) = ?$

先验概率$Pr(B = \text{yes}) = 9/14 \approx 0.64$

似然度$Pr(A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair} \mid B = \text{yes}) = Pr(A <= 30 \mid B = \text{yes}) * Pr(I = \text{medium} \mid B = \text{yes}) * Pr(S = \text{yes} \mid B = \text{yes}) * Pr(C = \text{fair} \mid B = \text{yes}) = (2/9) * (4/9) * (6/9) * (6/9)$

归一化后，$Pr(B = \text{yes} \mid A <= 30, I = \text{medium}, S = \text{yes}, C = \text{fair}) = 0.8$ 买电脑的置信度上升！

# 贝叶斯网络的构建

- 在给定一个数据样本集合的前提下，寻找一个与训练样本集匹配最好的网络结构，被称为贝叶斯网络的结构学习。
- 搜索最优的网络结构是一个NP难问题，从数据中学习贝叶斯网络的结构一直是研究的热点。


[1] C. K. Chow, C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3): 462-467, 1968.

[2] N. Wermuth, S. L. Lauritzen. Graphical and recursive models for contingency tables. *Biometrika*, 72: 537-552, 1983.

[3] G. Rebane, J. Pearl. The recovery of causal polytrees from statistical data. *UAI*, 222-228, 1987.

[4] G. F. Cooper, E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9: 309-347, 1992.

[5] X. Zheng, et al. DAGs with no tears: Continuous optimization for structure learning. *NeurIPS*, 9492-9503, 2018.

[6] S. Lachapelle, et al. Gradient-based neural DAG learning. *ICLR*, 2020.

[7] Y. Bengio, et al. A meta-transfer objective for learning to disentangle causal mechanisms. *ICLR*, 2020.

[8] Y. Luo, et al. When causal inference meets deep learning. *Nature Machine Intelligence*, 2: 426-427, 2020.