人工智能
不确定性知识表示与推理

陈 川
中山大学 计算机学院
2024年

中山大學
SUN YAT-SEN UNIVERSITY

案例

垃圾邮件困扰着互联网用户，正确识别技术难度大

- 关键词法：依据特定词语
- 校验码法：计算邮件文本的效验码，与已知的垃圾邮件进行对比
- 识别效果不理想，容易规避。

2002年，Paul Graham提出使用“贝叶斯推断”过滤垃圾邮件

- 1000封垃圾邮件可以过滤掉995封，没有误判
- 具有自我学习功能，根据新收到的邮件，不断调整。收到的垃圾邮件越多，准确率越高

案例

预先提供两组已经识别好的邮件：正常邮件和垃圾邮件各4000封

解析所有的邮件，提取每一个词。计算每个词语在正常邮件和垃圾邮件中出现的频率。

- “prince”在4000封垃圾邮件中，有200封包含，频率为5%；4000封正常邮件中，2封包含，频率为0.05%
- 而如果某个词**只出现**在垃圾邮件中，就假定它在正常邮件的出现频率是1%（经验值，可调整），反之亦然。避免概率为0

案例

收到一封新邮件，假定它是垃圾邮件的概率为50%

- S表示垃圾邮件（spam）H表示正常邮件（healthy）
- P(S)和P(H)的先验概率，均为50%

解析邮件，发现它包含prince这个词，请问这封邮件属于垃圾邮件的概率有多高?

用W表示事件—包含“prince”这个词，问题变成如何计算P(S|W)，即在某个词语（W)已经存在的情况下，垃圾邮件（S）的概率有多大

$$P(S \mid W) = \frac{P(W \mid S)P(S)}{P(W \mid S)P(S) + P(W \mid H)P(H)}$$

$P(W \mid S) = 5\%$ $P(W \mid H) = 0.05\%$

$P(S) = 50\%$ $P(H) = 50\%$

案例

能否得出结论，这封新邮件是垃圾邮件?

不能，一封邮件包含很多词语，一些词语（比如prince）说是垃圾邮件，另一些说不是，以哪个词为准?

选出这封邮件中P(S|W)最高的15个词，计算它们的联合概率（多个事件发生的情况下，另一个事件发生的概率有多大）。比如，W1和W2两个不同的词语，都出现在某封电子邮件中，这封邮件是垃圾邮件的概率，就是联合概率。

在已知W1和W2的情况下，两种结果：垃圾邮件（事件E1）或正常邮件（事件E2）。

事件	W_1	W_2	垃圾邮件
E_1	出现	出现	是的
E_2	出现	出现	不是

案例

事件	W_1	W_2	垃圾邮件
E_1	$P(S\|W_1)$	$P(S\|W_2)$	$P(S)$
E_2	$1-P(S\|W_1)$	$1-P(S\|W_2)$	$1-P(S)$

- 假定所有事件都是独立事件（严格来说，这个假定不成立，但是这里先忽略）

$$P(E_1) = P(S \mid W_1)P(S \mid W_2)P(S)$$

$$P = \frac{P(E_1)}{P(E_1) + P(E_2)}$$

案例

$$P = \frac{P(S \mid W_1)P(S \mid W_2)P(S)}{P(S \mid W_1)P(S \mid W_2)P(S) + (1 - P(S \mid W_1))(1 - P(S \mid W_2))(1 - P(S))}$$

将 (S) 等于 0.5 代入, 得到

$$P = \frac{P(S \mid W_1)P(S \mid W_2)}{P(S \mid W_1)P(S \mid W_2) + (1 - P(S \mid W_1))(1 - P(S \mid W_2))}$$

$$= P1P2/(P1P2 + (1 - P1)(1 - P2))$$

$$P = \frac{P_1 P_2 \cdots P_{15}}{P_1 P_2 \cdots P_{15} + (1 - P_1)(1 - P_2) \cdots (1 - P_{15})}$$

- Paul Graham的门槛值是0.9，概率大于0.9，表示15个词联合认定，这封邮件有90%以上的可能属于垃圾邮件；概率小于0.9，就表示是正常邮件
- 有了这个公式以后，一封正常的邮件即使出现了prince这个词，也不会被认定为垃圾邮件了。

Naïve Bayes 朴素贝叶斯特点

- Robust to isolated noise points / irrelevant attributes 根据后验概率判断，对数据噪音有天然的鲁棒性

- Sensitive to prior probability
（垃圾邮件例子中的P(S)对最终结果影响大）
  - Noninformative prior

- Independence assumption may not hold for some attributes （独立性条件有时候并不成立，如邮件中有些单词一起出现概率大，prince 与 account）
  - Use other techniques such as Bayesian Belief Networks (BBN) 贝叶斯网络or信念网络…

频率学派 vs 贝叶斯学派

贝叶斯公式真正得到重视和广泛应用却是最近二三十年的事，其间被埋没了200多年

原因在于我们有另外一种数学工具—经典统计学，或者叫频率主义统计学，它在200多年的时间里一直表现不错。从理论上它可以揭示一切现象产生的原因，既不需要构建模型，也不需要默认条件，只要进行足够多次的测量，隐藏在数据背后的原因就会自动揭开面纱

- 经典统计学：科学是关于客观事实的研究，我们只要反复观察一个可重复的现象，直到积累了足够多的数据，就能从中推断出有意义的规律
- 贝叶斯方法：像算命先生一样，从主观猜测出发，这显然不符合科学精神。连拉普拉斯后来也放弃了贝叶斯方法这一思路，转向经典统计学。因为他发现，如果数据量足够大，人们完全可以通过直接研究这些样本来推断总体的规律

频率学派 vs 贝叶斯学派

频率学派：

1. 概率必须符合科学的要求，可以用大量重复试验的频率去解释（概率=频率）
2. 其理论与方法的研究基于样本信息，从无到有

贝叶斯学派：

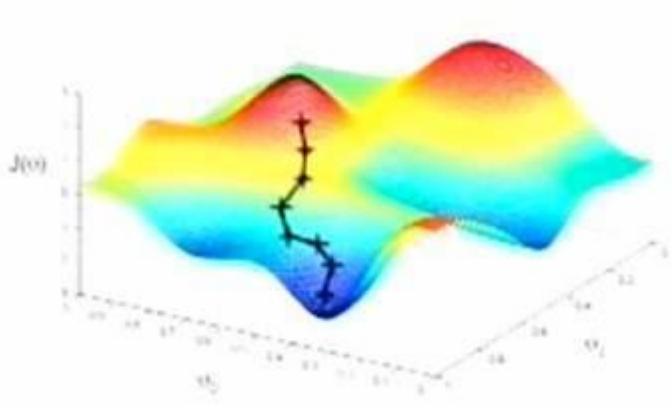
1. 概率是认识主体对事件出现可能性大小的相信程度（概率=知识经验+频率）
2. 其理论与方法的研究基于样本信息和先验信息，从有到优

频率学派 vs 贝叶斯学派

假如想知道某个区域里海拔最低的地方

- 经典统计学的方法是首先进行观测，取得区域内不同地方的海拔数据，然后从中找出最低点。这个数据量必须足够多，以反映区域内地形全貌的特征，这样我们才能相信找到的就是实际上的最低点
- 贝叶斯方法是我不管哪里最低，就凭感觉在区域内随便选个地方开始走，每一步都往下走，虽然中间可能有一些曲折，但相信这样走早晚能够到达最低点

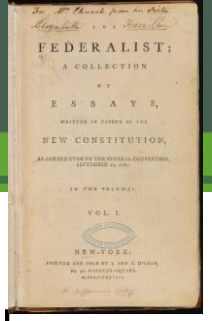
可以看出，贝叶斯方法的关键问题是这个最终到达的低点可能不是真正的最低点，而是某个相对低点，它可能对该区域的地形（碗型、马鞍形等）和最初我们主观选择的出发点有依赖性。这是贝叶斯方法最受经典统计学方法诟病的原因，也是它在过去的200多年被雪藏的原因所在



贝叶斯的崛起

- 另一个原因：频率学派主要使用最优化的方法，在很多时候处理起来要方便很多。而贝叶斯方法中先验后验计算复杂，直到计算机的迅速发展，以及抽样算法的进步（Gibbs sampling…）才重新回归
- 而两个标志性的事件在让学术界开始重视贝叶斯方法上起到了重要作用
  - 联邦党人文集作者公案
  - 天蝎号核潜艇搜救

联邦党人文集作者公案



1787年5月，美国各州（当时为13个）代表在费城召开制宪会议

1787年9月，美国的宪法草案被分发到各州进行讨论。一批反对派以“反联邦主义者”为笔名，发表了大量文章对该草案提出批评

宪法起草人之一**亚历山大·汉密尔**顿着急了，他找到曾任外交国务秘书（即后来的国务卿）的约翰·杰伊，以及纽约市国会议员**麦迪逊**，一同以普布利乌斯（Publius）的笔名发表文章，向公众解释为什么美国需要一部宪法。1788年，他们所写的85篇文章结集出版，这就是美国历史上著名的《联邦党人文集》

《联邦党人文集》出版的时候，汉密尔顿坚持匿名发表，于是，这些文章到底出自谁人之手，成了一桩公案

1810年，汉密尔顿接受了一个政敌的决斗挑战，但出于基督徒的宗教信仰，他决意不向对方开枪。在决斗之前数日，汉密尔顿自知时日不多，他列出了一份《联邦党人文集》的作者名单。1818年，麦迪逊又提出了另一份作者名单。这两份名单并不一致。在85篇文章中，有73篇文章的作者身份较为明确，其余**12篇作者存在争议**

联邦党人文集作者公案

1955年，哈佛大学统计学教授Fredrick Mosteller找到芝加哥大学的年轻统计学家David Wallance，建议他跟自己一起做一个小课题，他想**用统计学的方法，鉴定出《联邦党人文集》的作者身份**。

汉密尔顿 or麦迪逊？这根本不是小课题

- 汉密尔顿和麦迪逊都是文章高手，他们的文风非常接近
- 从已经确定作者身份的文本，汉密尔顿写了9.4万字，麦迪逊写了11.4万字
- 汉密尔顿每个句子的平均长度是34.55字，而麦迪逊是34.59字
- 就写作风格而论，汉密尔顿和麦迪逊简直就是一对双胞胎
- 汉密尔顿和麦迪逊写这些文章，用了大约一年的时间，而Mosteller和Wallance甄别出作者的身份花了10多年的时间

联邦党人文集作者公案

如何分辨两人写作风格的细微差别，并据此判断每篇文章的作者就是问题的关键

以贝叶斯公式为核心的分类算法

- 先挑选一些能够反映作者写作风格的词汇，在已经确定了作者的文本中，对这些特征词汇的出现频率进行统计**（条件概率）**
- 统计这些词汇在那些不确定作者的文本中的出现频率，从而根据词频的差别推断其作者归属

将近100个哈佛大学学生帮助处理数据

- 用打字机把《联邦党人文集》的文本打出来，然后把每个单词剪下来，按照字母表的顺序，把这些单词分门别类地汇集在一起

联邦党人文集作者公案

首先剔除掉用不上的词汇。比如“战争”、“立法权”、“行政权”等，这些词汇是因主题而出现，并不反映不同作者的写作风格。只有**像“in”，“an”，“of”，“upon”这些介词、连词等才能显示出作者风格的微妙差异**

一位历史学家好心地告诉他们，有一篇1916年的论文提到，汉密尔顿总是用while，而麦迪逊则总是用whilst

- 仅仅有这一个线索是不够的。while和whilst在这12篇作者身份待定的文章里出现的次数不够多。况且，汉密尔顿和麦迪逊有时候会合写一篇文章，也保不齐他们会互相改文章，要是汉密尔顿把麦迪逊的whilst都改成了while呢?

联邦党人文集作者公案

当学生们把每个单词的小纸条归类、粘好之后，他们发现，**汉密尔顿的文章里平均每一页纸会出现两次upon，而麦迪逊几乎一次也不用**

汉密尔顿更喜欢用enough，麦迪逊则很少用，其它一些有用的词汇包括：there、on等等

- 1964年，Mosteller和Wallance发表了他们的研究成果。他们的结论是，**这12篇文章的作者很可能都是麦迪逊。他们最拿不准的是第55篇，麦迪逊是作者的概率是240：1**
- 这个研究引起了极大的轰动，但最受震撼的不是宪法研究者，而是统计学家
- 这个研究把贝叶斯公式这个被统计学界禁锢了200年的幽灵从瓶子中释放了出来

天蝎号核潜艇搜救

1968年5月，美国海军的天蝎号核潜艇在大西洋亚速海海域突然失踪，潜艇和艇上的99名海军官兵全部杳无音信

事后调查报告：这艘潜艇上的一枚奇怪的鱼雷，发射出去后竟然敌我不分，扭头射向自己，让潜艇中弹爆炸

为了寻找天蝎号，美国政府调集了多位专家前往现场，包括了**John Craven数学家，美国海军特别计划部首席科学家**

早在1966年，Craven就帮忙找到一颗“不小心”丢失的氢弹….

天蝎号核潜艇搜救

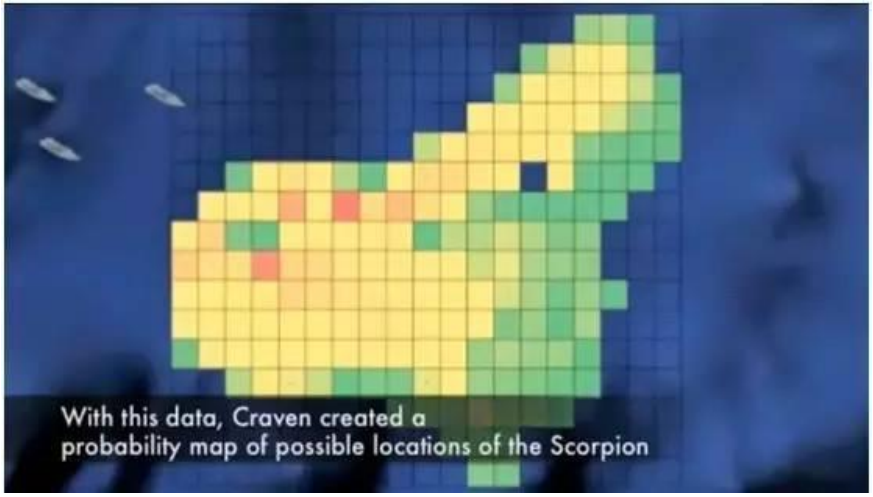
- Craven使用了贝叶斯公式。他召集了数学家、潜艇专家、海事搜救等各个领域的专家。每个专家都有自己擅长的领域，但并非通才，没有专家能准确估计到在出事前后潜艇到底发生了什么
- Craven并不要求团队成员互相协商寻求一个共识，而是让各位专家编写了各种可能的“剧本”，让他们按照自己的知识和经验对于情况会向哪一个方向发展进行猜测，并评估每种情境出现的可能性
- 他把一瓶威士忌作为猜中的奖品。于是团队成员开始对潜艇可能遇到的麻烦、潜艇的下沉速度、下沉角度等因素下注

天蝎号核潜艇搜救

这一做法受到了很多同行的质疑

- 因为结果很多是这些专家以猜测、投票甚至可以说赌博的形式得到的，不可能保证准确性
- 因为搜索潜艇的任务紧迫，没有时间进行精确的实验、建立完整可靠的理论

Craven粗略估计了一下，半径20英里的数千英尺深的海底，都是天蝎号核潜艇可能沉睡的地方，Craven把各位专家的意见综合到一起，得到了一张20英里海域的概率图



天蝎号核潜艇搜救

- 每个小格子有两个概率值p和q
  - p: 潜艇躺在这个格子里的概率，
  - q: 如果潜艇在这个格子里，它被搜索到的概率 (主要跟海域的水深有关，深海区漏网可能性更大)
- 如果一个格子被搜索后，没有发现潜艇的踪迹，那么按照贝叶斯公式，这个格子潜艇存在的概率就会降低
  - A:潜艇躺在这个格子里　P(A) = p
  - B: 潜艇在格子里被找到

    $P(\bar{B})=P(A)P(\bar{B}|A)+P(\bar{A})P(\bar{B}|\bar{A})=p(1-q)+(1-p)1$
  - $P(A|\bar{B})=P(\bar{B}|A)P(A)/P(\bar{B})=p(1-q)/p(1-q)+(1-p)$

$$p' = \frac{p(1-q)}{(1-p)+p(1-q)} = p\frac{1-q}{1-pq} < p$$

天蝎号核潜艇搜救

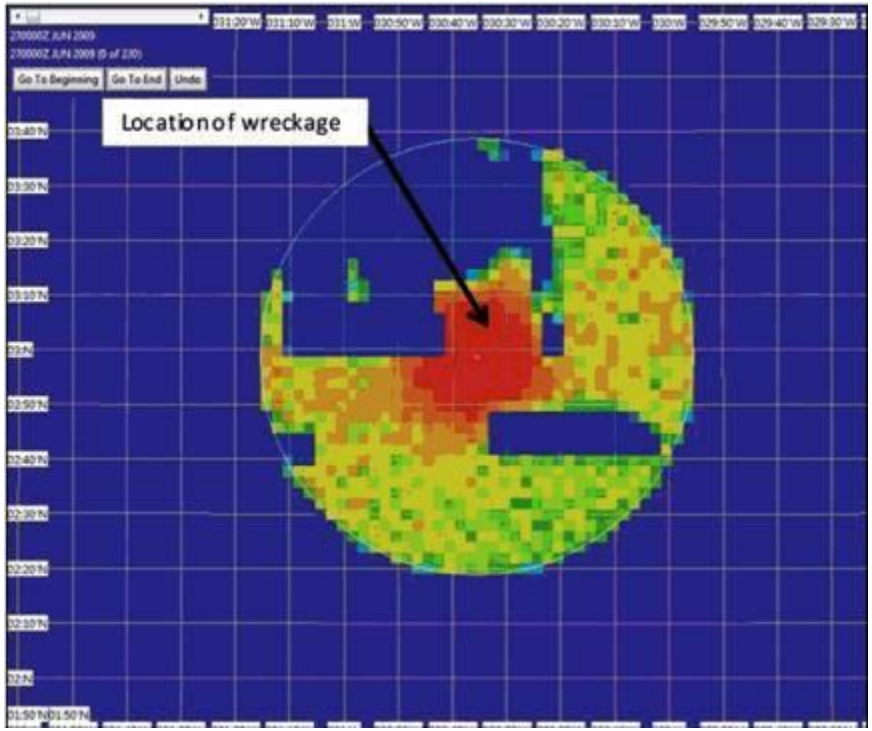
- 由于所有格子概率的总和是1，这时其他格子潜艇存在的概率值就会上升

$$r' = r\frac{1}{1-pq} > r$$

- 先搜索概率值最高的格子，如没发现，重算概率分布图，搜寻船搜索新的最可疑格子
- 最初，海军人员对Craven和其团队的建议嗤之以鼻，他们凭经验估计潜艇是在爆炸点的东侧海底。但几个月一无所获，他们才不得不听从了Craven的建议，按照概率图在爆炸点的西侧寻找，经过几次搜索，潜艇果然在爆炸点西南方被找到

天蝎号核潜艇搜救

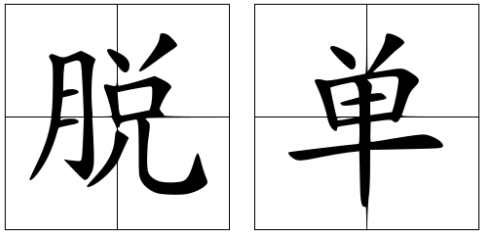
- 由于这种基于贝叶斯公式的方法在后来多次搜救实践中被成功应用，现在已经成为海难空难搜救的通行做法（Bayesian search theory）



2009年法航空难搜救的后验概率分布图：

讨论一个严肃的问题

曾几何时，偶遇心仪的漂亮女孩，从此日思夜想，废寝忘食，开始了漫长的暗恋之旅，等到一日，在无尽的纠结中，终于鼓起勇气向女孩表白，结果女孩一句“我已经有男朋友了”如晴天霹雳，实在难以接受……



为了避免此种尴尬的发生，如何准确判断一位男/女生是否单身就成了一项重要的技能

目标：在对方毫无察觉的情况下，就可以用手头有限的信息判断出他/她的单身概率，不仅如此，死理性派追求的结果一定是量化的，计算出的单身概率还要保留两位小数

一个严肃的问题

- 第一步，多找几个朋友一起秘密观察目标，找的人最好多样化，心事鉴定组、谣言粉碎机、自然控、犯罪法医谜最好都找几个来，**人越多越好，越多样化越好**
- 然后大家根据自己对他/她的印象从各自的角度估计一下目标单身的概率是多少，投一下票，最后的结果一定会有差异了，可能心事鉴定组觉着单身可能性是90%，谣言粉碎机却觉着单身只是谣言。死理性派根据这些人平时一贯的靠谱程度，把每个人说出的概率平均一下，原则是平时比较靠谱的人给出的结果考虑的比重就要大一些，不靠谱的人给的数字就要小一些
- 假设最后得到的结果是:他/她单身概率为65.65%

这个65.65是什么？先验概率

一个严肃的问题

- 像做科学研究一样，先查一下资料，网络上随便一搜就可以找到不少寂寞人士多年潜心研究的简单易用的单身判别标准，比如手机原则（恋爱中的女生手机使用频率会比较高），自习原则（单身的女孩常常和几个女生结伴上自习的话）。之后，在自己身边已经知道是否单身的人群中做一下统计实验，当然样本越大越好了，得到：

① 在单身中, 经常结伴上自习的比例是多少, 独自上自习的比例是多少;
② 在非单身中, 经常结伴上自习的比例是多少, 独自上自习的比例是多少;
③ 在经常结伴上自习中，单身占多少比例，非单身占多少比例;
④ 在独自上自习中，单身占多少比例，非单身占多少比例;
⑤ 学习中的手机使用频率（次/每小时）的概率分布;
⑥ .....
⑦ .....

我们是在计算什么？ （条件概率）

一个严肃的问题

- 对投票出来的概率值65.65%进行修正和优化。依靠的是目标女孩在各项标准上的表现
  - 比如发现目标mm喜欢和朋友一起去上自习，而根据自己的统计研究结果：在已经恋爱的mm中，喜欢和朋友一起去上自习的女孩大约占其中的30%；在没有恋爱的女孩中，喜欢和朋友一起去上自习的女孩大约占其中的60%
  - A：女孩单身
  - B：女孩跟朋友一起上自习
  - $P(A|B)=\frac{P(A)P(B|A)}{P(B)}=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\bar{A})P(B|\bar{A})}$

$$\frac{\text{女孩单身的概率}\times\text{单身女孩中和朋友一起上自习的概率}}{\text{女孩单身的概率}\times\text{单身女孩中和朋友一起上自习的概率}+\text{女孩恋爱的概率}\times\text{恋爱女孩中和朋友一起去上自习的概率}}$$

$$\frac{65.65\%\times 60\%}{65.65\%\times 60\%+34.35\%\times 30\%}=79.26\%$$

一个严肃的问题

- 研究结果还发现，在单身女孩中，手机使用率高于1.2次/小时占其中的20%；在已经恋爱的女孩中，这一数值则是60%。对于目标女孩的观察结果是，她的手机使用率高于每小时1.2次，那么概率结果又要更新了
  - A：女孩单身
  - B：女孩手机使用率高于1.2次/小时
  - $P(A|B)=\frac{P(A)P(B|A)}{P(B)}=\frac{P(A)P(B|A)}{P(A)P(B|A)+P(\bar{A})P(B|\bar{A})}$

$$\frac{79.26\% \times 20\%}{79.26\% \times 20\% + 20.74\% \times 60\%} = 56.02\%$$

一个严肃的问题

- 这回单身的概率又悲剧地降回了56.02%（**后验概率**），死理性派可以去找更多的“评核标准”，做更多的研究，不断更新女孩的单身概率值，让它越来越贴近事实，可以定一个阈值：对方单身概率超过这个阈值，自己就出手表白
- 不管计算次数多少，得到的终归是一个概率值，不是事实，就算经过多次研究，已经可以将目标女孩的单身概率确定到99.9%，
- 马上就准备向她表白了，可是在最后一次对女孩的观察研究中，发现人家和一个男生手挽手的有说有笑，那么，女孩的单身概率值会立刻从99.9%掉到接近于0……　（ **P(B|A)接近于0** ）

免责声明：方法有风险，尝试需谨慎，本课程不负任何责任

一个严肃的问题

- 这回单身的概率又悲剧地降回了56.02%（**后验概率**），死理性派可以去找更多的“评核标准”，做更多的研究，不断更新女孩的单身概率值，让它越来越贴近事实，可以定一个阈值：对方单身概率超过这个阈值，自己就出手表白
- 不管计算次数多少，得到的终归是一个概率值，不是事实，就算经过多次研究，已经可以将目标女孩的单身概率确定到99.9%，
- 马上就准备向她表白了，可是在最后一次对女孩的观察研究中，发现人家和一个男生手挽手的有说有笑，那么，女孩的单身概率值会立刻从99.9%掉到接近于0……　（ **P(B|A)接近于0** ）

- 贝叶斯统计方法：先验概率+新得到的证据=更正后的概率
  - 不受信息量多少的限制，将各种来源的结果，包括主观判断和有限的客观信息，综合到一起，得到最后的结论
- 思考：上述过程存在什么哪些局限性？ **独立性假设**

免责声明：方法有风险，尝试需谨慎，本课程不负任何责任

不确定性知识表示与推理

- 概率统计、独立性、贝叶斯规则
- 贝叶斯推断-朴素贝叶斯方法
- 贝叶斯网络

独立性带给我们什么？

- Suppose Boolean variables $X_1, X_2, \ldots, X_n$ are mutually independent (*i.e.*, every subset is variable independent of every other subset)
- We can specify full joint distribution (probability function over all vectors of values) using only n parameters (linear) instead of $2^n - 1$ (exponential)
  只使用n个参数(线性)来计算完整的联合分布
- Simply specify $Pr(X_1 = true), \ldots, Pr(X_n = true)$ (*i.e.*, $Pr(X_i = true)$ for all i)
- We can easily recover probability of any primitive event, *e.g.*
- $Pr(X_1 \neg X_2 X_3 X_4) = Pr(X_1)(1 - Pr(X_2))Pr(X_3)Pr(X_4)$

The Value of Independence 独立性的价值

- Complete independence reduces both representation and inference from $O(2^n)$ to $O(n)$!
- Unfortunately, such complete mutual independence is rare.

  Most realistic domains do not exhibit this property.

  完全的相互独立性不存在了！
- Fortunately, most domains do exhibit a fair amount of conditional independence. 而条件独立性时常有！
- And we can exploit conditional independence for representation and inference as well.
- Bayesian networks do just this

Exploiting Conditional Independence

想象这么一个故事:

- If Craig woke up too early (E), Craig probably needs coffee (C);
- if C, he's likely angry (A).
- If A, there is an increased chance of a burst blood vessel (B).
- If B, Craig is quite likely to be hospitalized (H).

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

Exploiting Conditional Independence

- If you learned any of E, C, A, or B, your assessment of $Pr(H)$ would change. **对E，C，A，B信息的掌握影响对$Pr(H)$的评估**
  - *e.g.*, if any of these are seen to be true, you would increase $Pr(h)$ and decrease $Pr(\neg h)$.
  - So H is not independent of E, or C, or A, or B.
- But if you knew value of B (true or false), learning value of E, C, or A, would not influence $Pr(H)$. Influence these factors have on H is mediated by their influence on B.
  - Craig doesn't get sent to the hospital because he's angry, he gets sent because he's had a burst blood vessel.
  - So H is independent of E, and C, and A, given B

Exploiting Conditional Independence

- Similarly
  - B is independent of E, and C, given A
  - A is independent of E, given C
- This means that:
  - $Pr(H|B,\{A, C, E\}) = Pr(H|B)$
  - $Pr(B|A,\{C, E\}) = Pr(B|A)$
  - $Pr(A|C,\{E\}) = Pr(A|C)$
  - $Pr(C|E)$ and $Pr(E)$ don't simplify

Exploiting Conditional Independence

By the chain rule (for any instantiation of H…E):

$$P(H,B,A,C,E) = P(H|B,A,C,E)\, P(B|A,C,E)\, P(A|C,E)\, P(C|E)\, P(E)$$

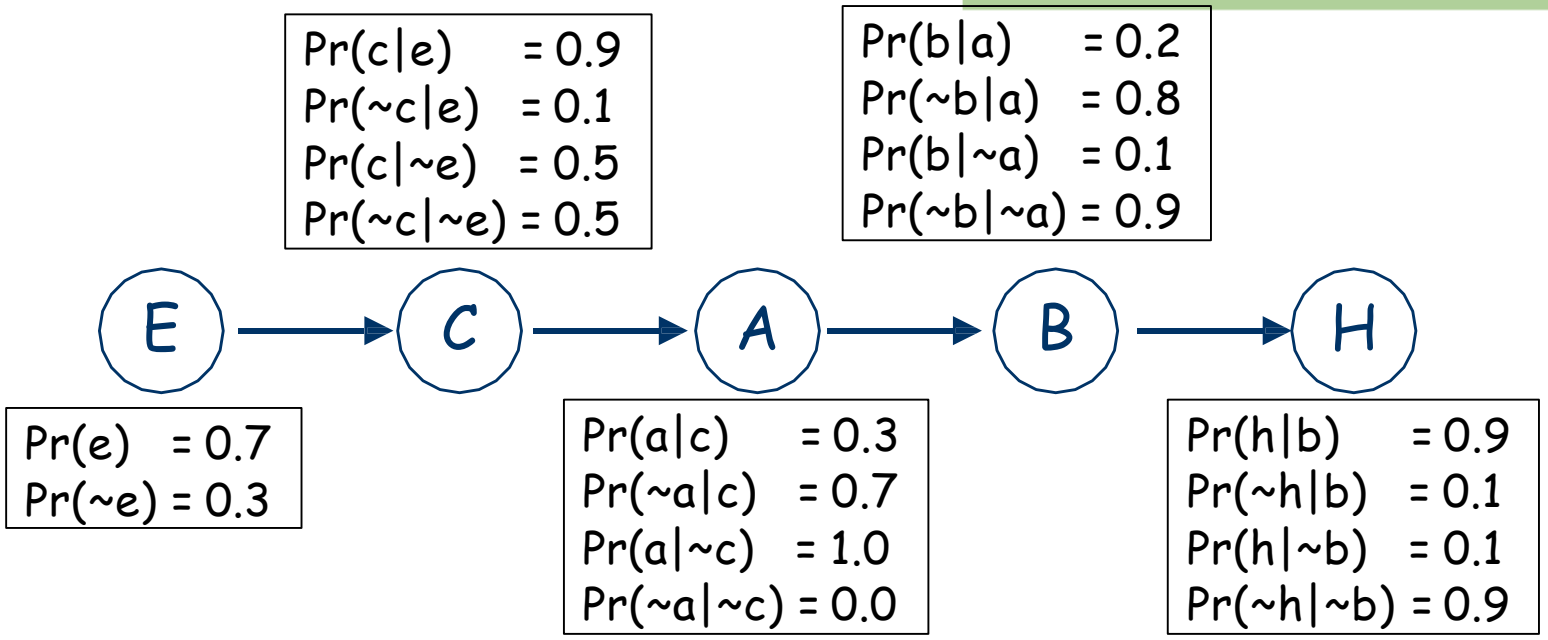
$2^4+2^3+2^2+2^1+2^0 = 2^5-1$

By our conditional independence assumptions:

$$P(H,B,A,C,E) = P(H|B)\, P(B|A)\, P(A|C)\, P(C|E)\, P(E)$$

We can specify the full joint by specifying five local conditional distributions (joints): $P(H|B)$; $P(B|A)$; $P(A|C)$; $P(C|E)$; and $P(E)$

Example quantification

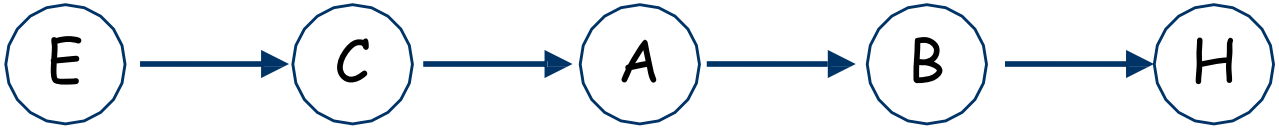


Note that half of these are "1 minus" the others

So specifying the full joint requires only 9(1+2+2+2+2) parameters, instead of 31 for explicit representation

- Linear in number of variables instead of exponential!
- Linear generally if dependence has a chain structure

Inference is Easy



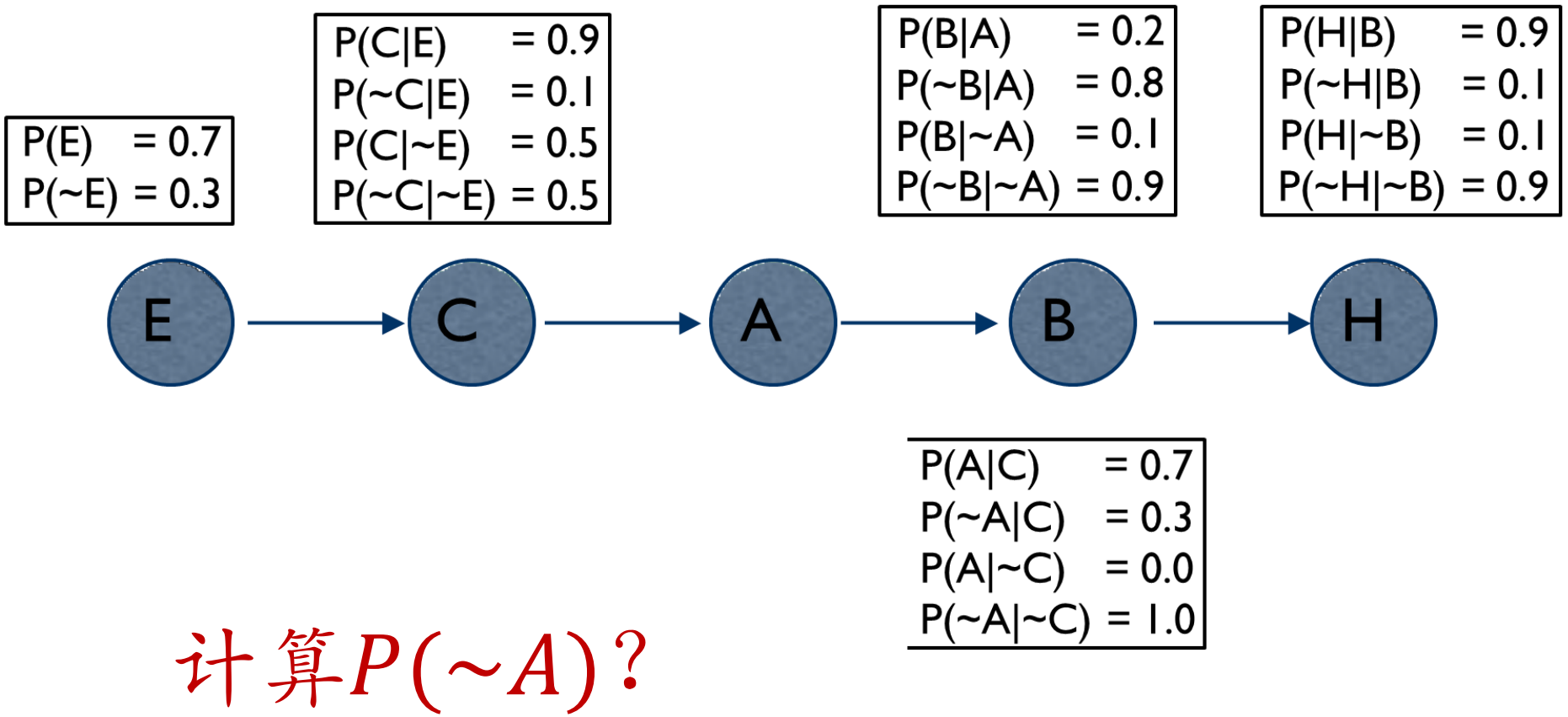
Want to know P(a)? Use summing out rule:

$$\Pr(a) = \sum_{c_i \in Dom(C)} \Pr(a \mid c_i) \Pr(c_i)$$

$$= \sum_{c_i \in Dom(C)} \Pr(a \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i) \Pr(e_i)$$

These are all terms specified in our local distributions!

Inference is Easy



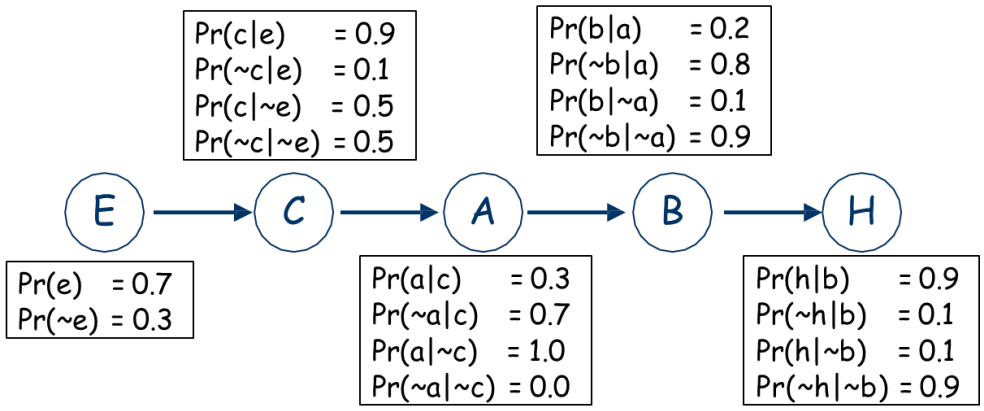
计算$P(\sim A)$?

Bayesian Networks: graph + tables 一图一表一BN

- The structure above is a Bayesian network（贝叶斯网络）.
- A BN is a graphical representation of the direct dependencies over a set of variables, together with a set of conditional probability tables (CPTs) quantifying the strength of those influences.

Bayes nets generalize the above ideas in very interesting ways, leading to effective means of representation and inference under uncertainty.

一图：有向无环图
一表：条件概率表



Bayesian Networks

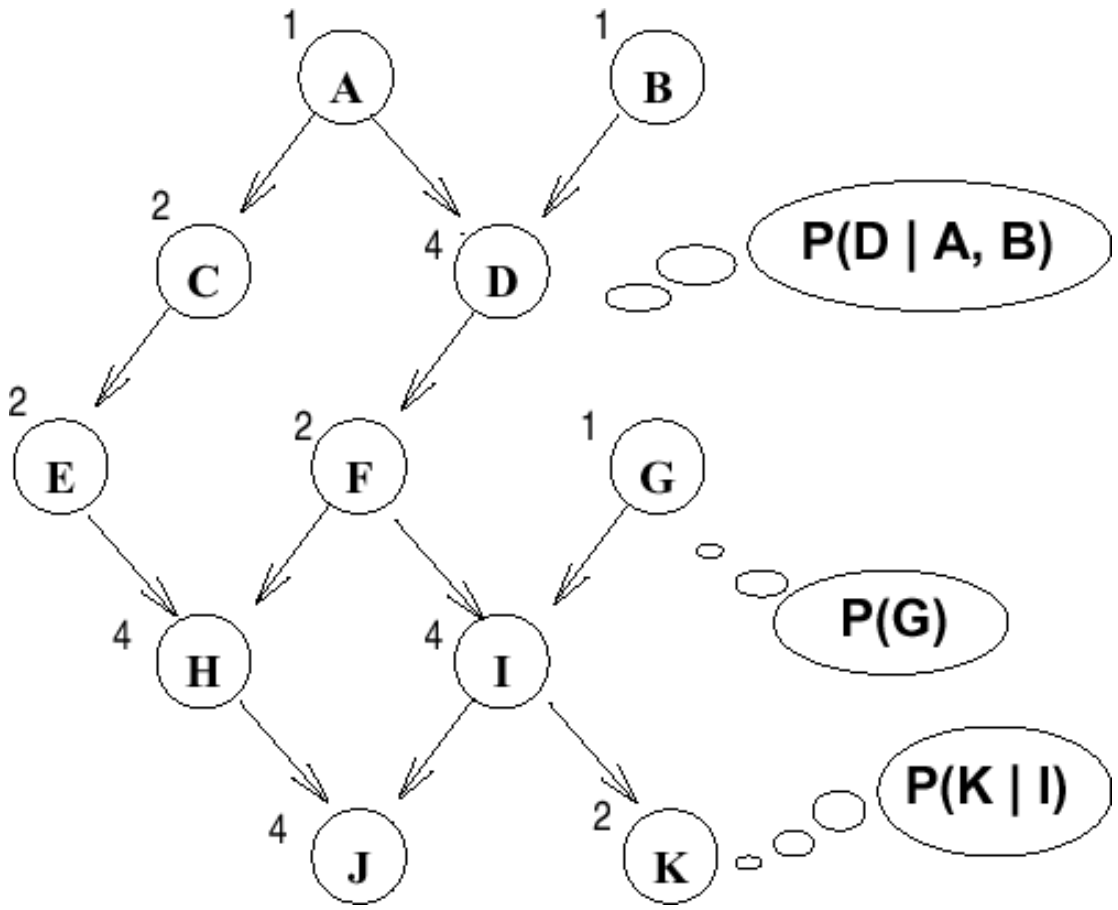
A BN over variables $\{X_1, X_2, \ldots, X_n\}$ consists of:

- a DAG (directed acyclic graph) whose nodes are the variables
- a set of CPTs (conditional probability tables) $Pr(X_i | Par(X_i))$ for each X_i

Key notions

- parents of a node: $Par(X_i)$
- children of a node
- descendents of a node
- ancestors of a node
- family: set of nodes consisting of X_i and its parents

Example (此处变量取二值)



- A couple CPTS are “shown” with 11 variables
- Explicit joint requires **2^{11} -1 =2047 parmtrs!**
- BN requires only **27 parmtrs** (the number of entries for each CPT is listed)

Semantics of Bayes Nets

- A Bayes net specifies that the joint distribution over the variable in the net can be written as the following product decomposition.

$$Pr(X_1, X_2, \ldots, X_n) = Pr(X_n | Par(X_n)) * Pr(X_{n-1} | Par(X_{n-1})) * \ldots * Pr(X_1 | Par(X_1))$$

- This equation holds for any set of values $d_1, d_2, \ldots, d_n$ for the variables $X_1, X_2, \ldots, X_n$.

- *e.g.*, We have X_1, X_2, X_3 each with domain $Dom[X_i] = \{a, b, c\}$ and we have

$$Pr(X_1, X_2, X_3) = P(X_3 | X_2) P(X_2) P(X_1)$$

- Then

$$Pr(X_1 = a, X_2 = a, X_3 = a) = P(X_3 = a | X_2 = a) P(X_2 = a) P(X_1 = a)$$

Another Bayesian Network

A: Aameri gives the lecture

S: It is sunny out

L: The lecturer arrives late

Assume that all instructors may arrive late in bad weather.
Some instructors may be more likely to be late than others.

Another Bayesian Network

A: Aameri gives the lecture

S: It is sunny out

L: The lecturer arrives late

Assume that all instructors may arrive late in bad weather.
Some instructors may be more likely to be late than others.

We'll start by writing down what we know:
$P(S|A) = P(S)$, $P(S) = 0.3$, $P(A) = 0.5$

Lateness is not independent of the weather or the lecturer.
We need to formulate P(L|S,A) for all of the values of S and A.

Another Bayesian Network

A: Aameri gives the lecture

S: It is sunny out

L: The lecturer arrives late

$$P(S|A) = P(S), P(S) = 0.3, P(A) = 0.5$$

$$P(L|S,A) = 0.05, P(L|S,\sim A) = 0.1, P(L|\sim S,A) = 0.1, P(L|\sim S,\sim A) = 0.2$$

Because of conditional independence, we only need 6 values to specify the full joint instead of 7=2^3-1.

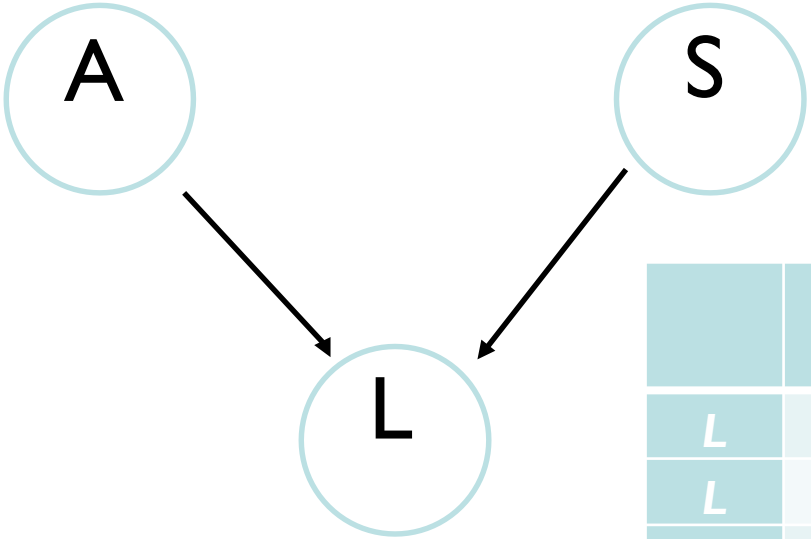
Conditional independence leads to computational savings!

Drawing the Network

$P(S|A) = P(S), P(S) = 0.3, P(A) = 0.5$

$P(L|S,A) = 0.05, P(L|S,\sim A) = 0.1, P(L|\sim S,A) = 0.1, P(L|\sim S,\sim A) = 0.2$

A	*T*	*0.5*
A	*F*	*0.5*

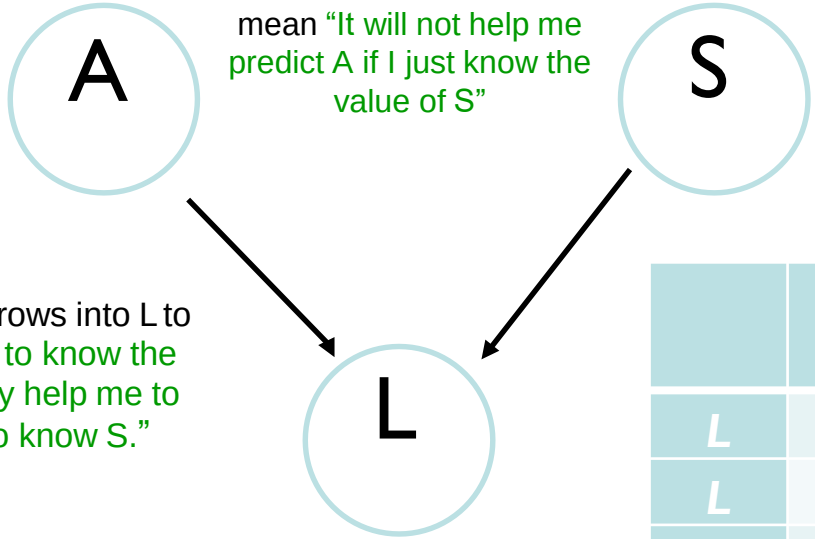


S	*T*	*0.3*
S	*F*	*0.7*

	A	*S*	*P(L=T|A,S)*
L	*T*	*T*	*0.05*
L	*T*	*F*	*0.1*
L	*F*	*T*	*0.1*
L	*F*	*F*	*0.2*

Drawing the Network

A	T	0.5
A	F	0.5



Read the absence of an arrow between S and A to mean "It will not help me predict A if I just know the value of S"

S	T	0.3
S	F	0.7

Read the two arrows into L to mean "If I want to know the value of L it may help me to know A and to know S."

	A	S	P(L=T|A,S)
L	T	T	0.05
L	T	F	0.1
L	F	T	0.1
L	F	F	0.2

Back to the network

Now let's suppose we have these three events:

A: Aameri gives the lecture (~A: Allin gives the lecture)

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

And we know:

- Allin has a higher chance of being late than Aameri. （~A）
- Allin has a higher chance of giving lectures about reasoning with BNs

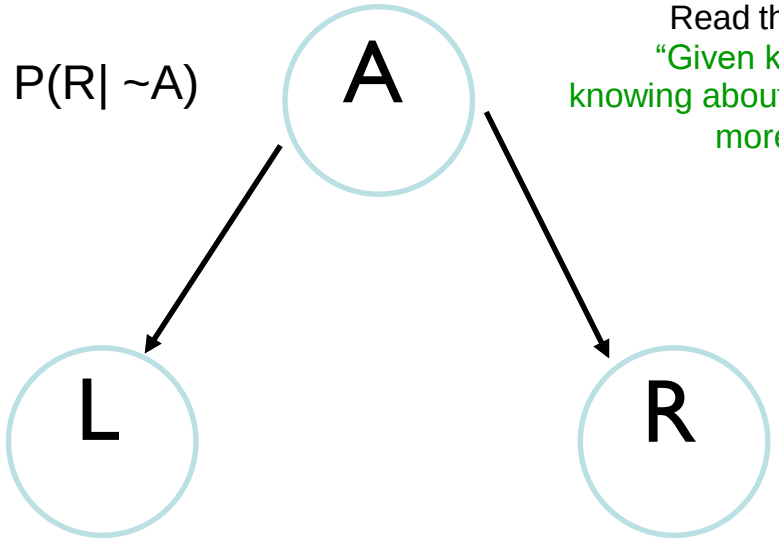
What kind of independences exist in our graph?

Back to the network

Once you know who the lecturer is, then whether they arrive late doesn't affect whether the lecture concerns Reasoning with Bayes' Nets, i.e.:

$P(R|A,L) = P(R|A)$

$P(R|\sim A,L) = P(R|\sim A)$

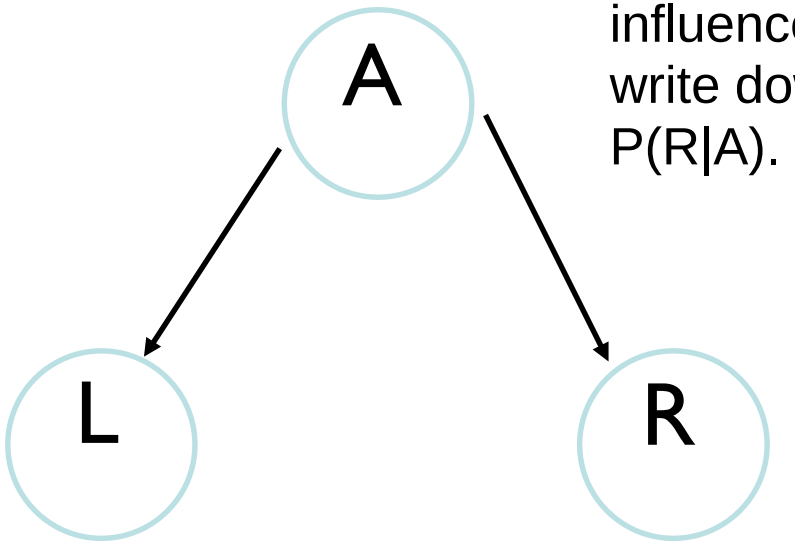


Read this diagram as
"Given knowledge of A,
knowing about L won't tell anything
more about R."

The network reflects conditional independences

To specify CPTs, we first write down P(A). Then, because L and R are only directly influenced by A, we write down P(L|A) and P(R|A).

A	T	0.5



	A	P(L=T|A)
L	T	0.085
L	F	0.17

	A	P(R=T|A)
R	T	0.3
R	F	0.6

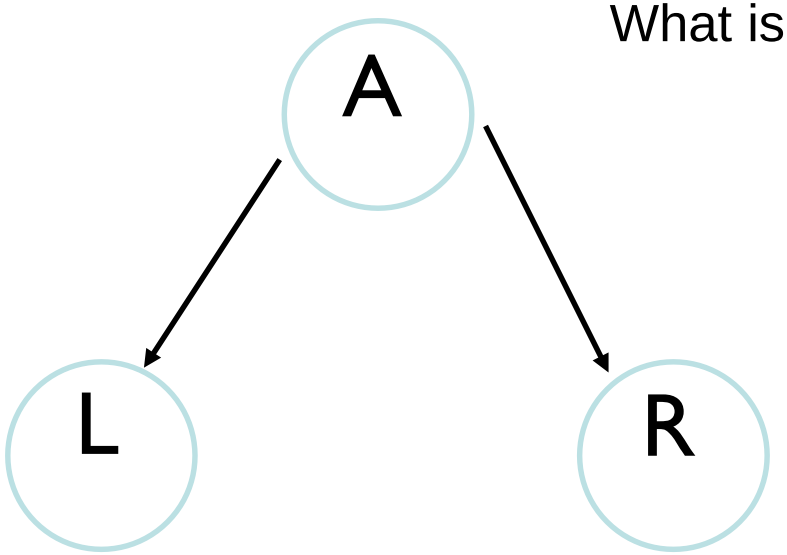
R is conditionally independent of L given A (and vice versa)

The network reflects conditional independences

What is P(L,R,A)?

What is P(~L,R,A)?

A	T	0.5



	A	P(L=T|A)
L	T	0.085
L	F	0.17

	A	P(R=T|A)
R	T	0.3
R	F	0.6

R is conditionally independent of L given A (and vice versa)

Try to Build a Bayes Net

A: Aameri gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: The lecture starts on time.

- T is only directly influenced by L (i.e. T is conditionally independent of R,A,S given L)
- L is only directly influenced by A and S (i.e. L is conditionally independent of R given A & S)
- R is only directly influenced by A (i.e. R is conditionally independent of L,S, given A)
- A and S are independent

Building a Bayes Net

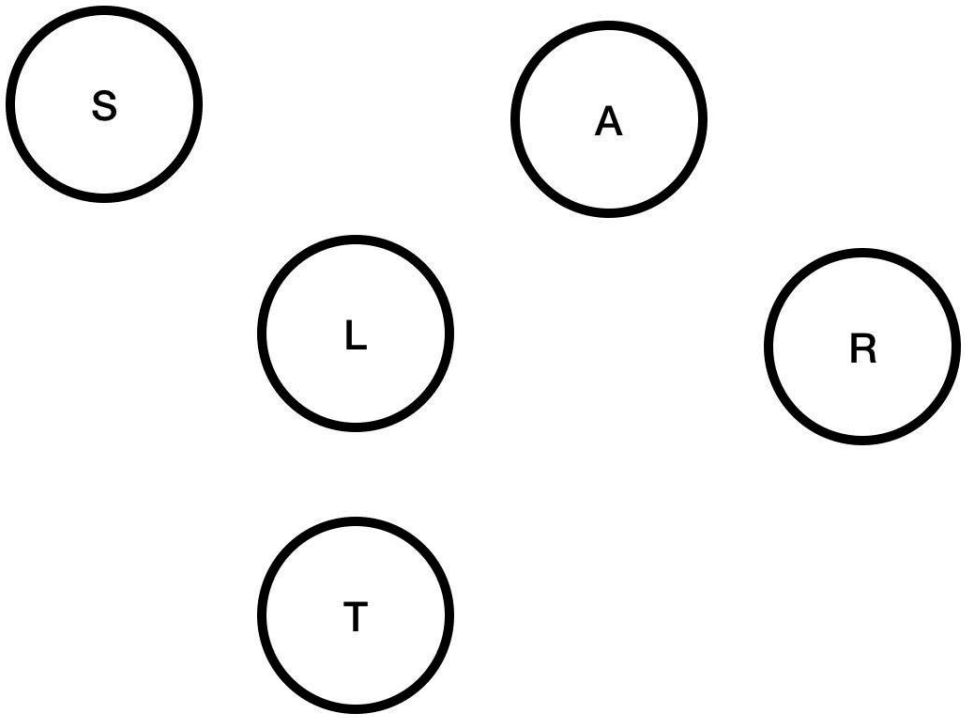
A: Aameri gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: Lecture starts on time.



Step One: Add variables

Building a Bayes Net

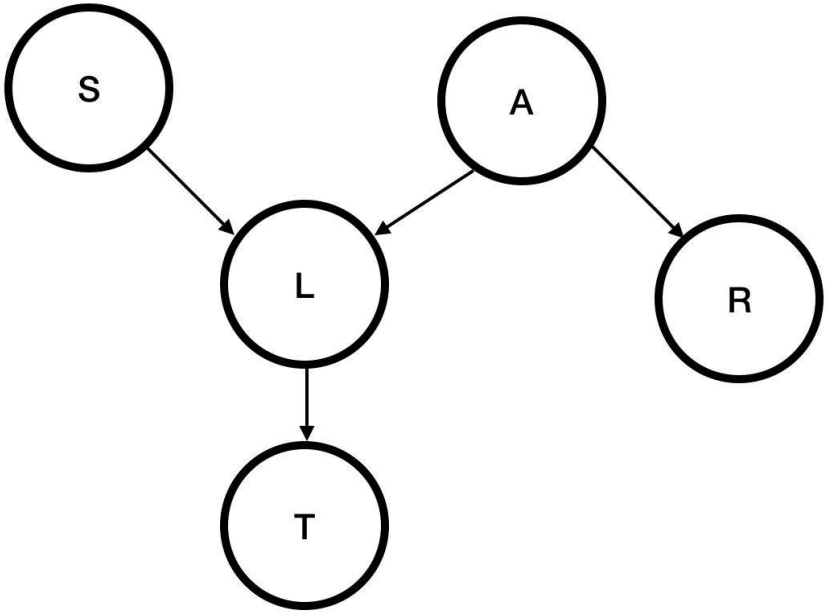
A: Aameri gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: Lecture starts on time.

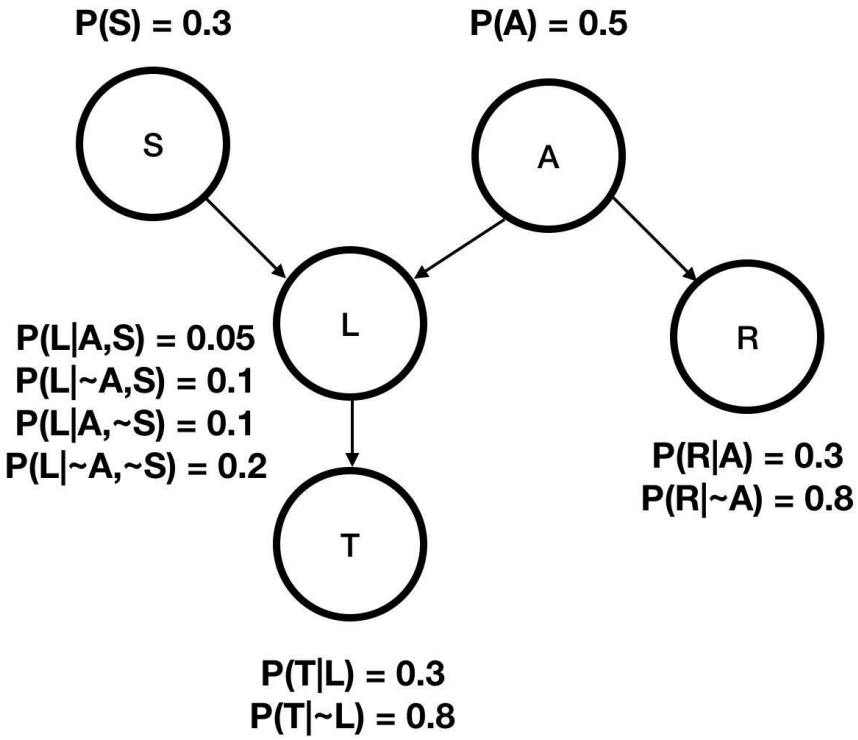


Step Two: add links.

The link structure must be acyclic.

If you assign node Y the parents $X_1, X_2, \ldots, X_n$, you are promising that, given $\{X_1, X_2, \ldots, X_n\}$, Y is conditionally independent of any other variable that's not a descendent of Y

Building a Bayes Net



A: Aameri gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

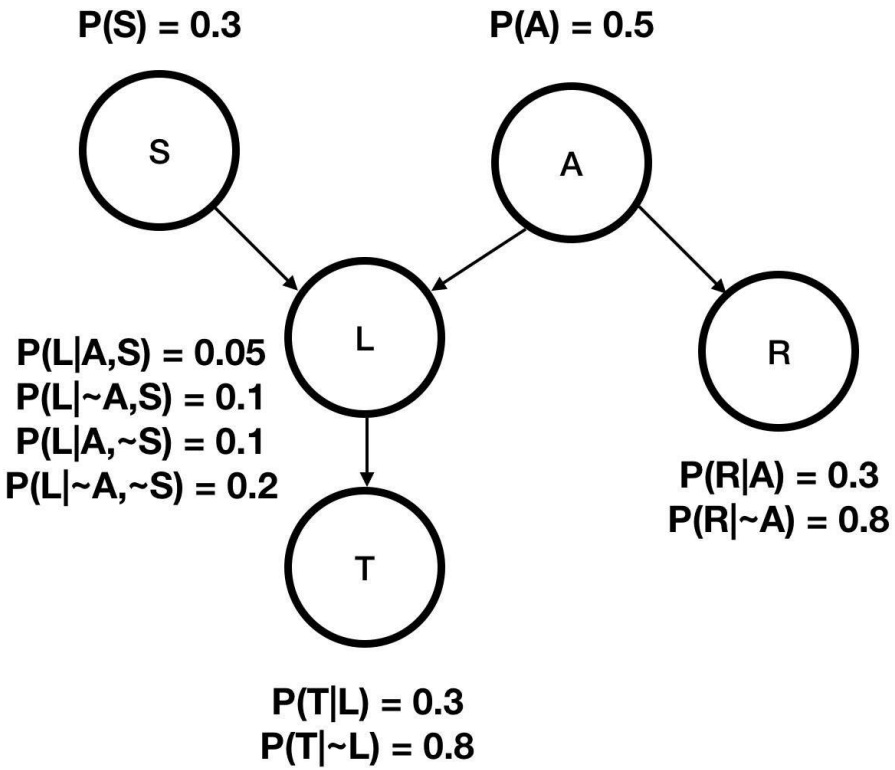
S: It is sunny out

T: Lecture starts on time.

Step Three: add a conditional probability table (CPT) for each node.

The table for X must define P(X|Parents) and for all combinations of the possible parent values.

Building a Bayes Net



A: Aameri gives the lecture

L: The lecturer arrives late

R : The lecturer concerns Reasoning with Bayes' Nets

S: It is sunny out

T: Lecture starts on time.

You can deduce many probability relations from a Bayes Net.

Note that variables that are not directly connected may still be correlated.

Building a Bayes Net

It is always possible to construct a Bayes net to represent any distribution over the variables X_1, X_2,…, X_n, using any ordering ofthe variables.

Take any ordering of the variables (say, the order given). From the chain rule we obtain.

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

Now for each X_i go through its conditioning set $X_1,\ldots,X_{i-1}$, and iteratively remove all variables X_j such that X_i is conditionally independent of X_j given the remaining variables. Do this until no more variables can be removed.

The final product specifies a Bayes net.

Constructing a Bayes Net

It is always possible to construct a Bayes net to represent any distribution over the variables $X_1, X_2, \ldots, X_n$, using any ordering of the variables.

Step 1. Apply the Chain Rule using any order of variables. (We will see later that you may wish to use causality or some other property to guide the variable ordering. 因果关系或其他准则引导排序)

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

Step 2. For each X_i go through its conditioning set $X_1, \ldots, X_{i-1}$ and iteratively remove all variables X_j such that X_i is conditionally independent of X_j given the remaining variables. Do this until no more variables can be removed.

Constructing a Bayes Net: Step 3

- Step 2 will yield a product decomposition.
  $Pr(X_n|Par(X_n))Pr(X_{n-1}|Par(X_{n-1}))\ldots Pr(X_1)$
- To create the Bayes Net, create a directed acyclic graph (DAG) such that each variable is a node and the conditioning set $Par(X_i)$ of a variable X_i are X_i's parents in the DAG.

Constructing a Bayes Net: Step 4

- Specify the conditional probability table (CPT) for each family (variable and its parents).
- Typically we represent the CPT as a table mapping each setting of $\{X_i, Par(X_i)\}$ to the numeric probability of X_i taking that particular value given that the variables in $Par(X_i)$ have their specified values.
- If each variable has d different values, we will need a table of size $d^{|\{X_i, Par(X_i)\}|}$.
- *i.e.*, exponential in the size of the parent set.

Variable Ordering Matters - Causal Intuitions

- The Bayes Net can be constructed using an arbitrary ordering of the variables. 变量序列
- However, some orderings will yield BNs with very large parent sets. This requires exponential space, and (as we will see later) exponential time to perform inference.
- Empirically, and conceptually, a good way to construct a BN is to use an ordering based on causality. This often yields a more natural and compact BN. 基于因果关系的变量序列可以获得更加自然紧致的贝叶斯网络结构

Causal Intuitions

- Malaria, the flu and a cold all cause aches. So use the ordering that causes come before effects: Malaria, Flu, Cold, Aches

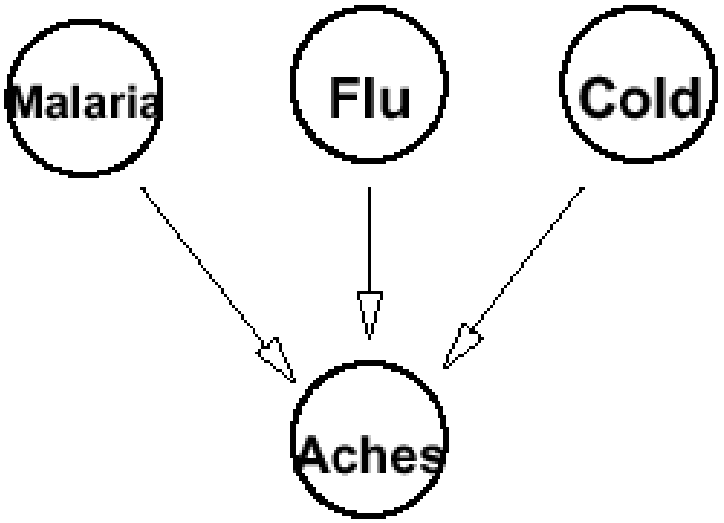
$$Pr(M,\ F,\ C,\ A) = Pr(A|M,F,C)Pr(C|M,F)Pr(F|M)Pr(M)$$

- Each of these disease affects the probability of aches, so the first conditional probability does not change.
- It is reasonable to assume that these diseases are independent of each other: having or not having one does not change the probability of having the others.
- So $Pr(C|M,F) = Pr(C),\ Pr(F|M) = Pr(F)$

Causal intuitions

This yields a fairly simple Bayes net.

We only need one big CPT, involving the family of "Aches".



Causal Intuitions

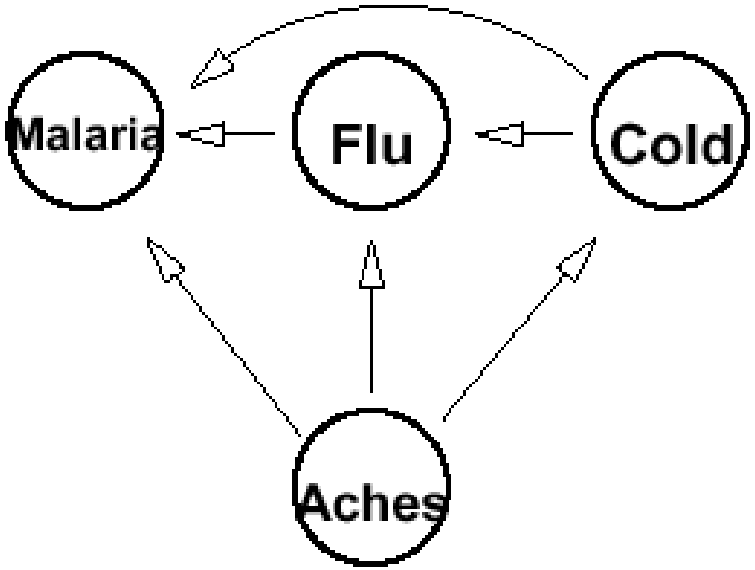
- Suppose we build the BN using the opposite ordering: Aches, Cold, Flu, Malaria

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

- Can we reduce $Pr(M|A, C, F)$? No
  - Probability of Malaria is clearly affected by knowing aches.
  - How about knowing aches and cold, or aches and cold and flu?
  - Probability of Malaria is affected by both of these additional pieces of knowledge
  - Knowing Cold and of Flu lowers the probability of Aches indicating Malaria since they "explain away" Aches!
- Similarly, we can't reduce $Pr(F|A, C)$.
- $Pr(C|A) \neq Pr(C)$

Causal intuitions

We obtain a much more complex Bayes net. In fact, we obtain no savings over explicitly representing the full joint distribution (i.e., representing the probability of every atomic event).

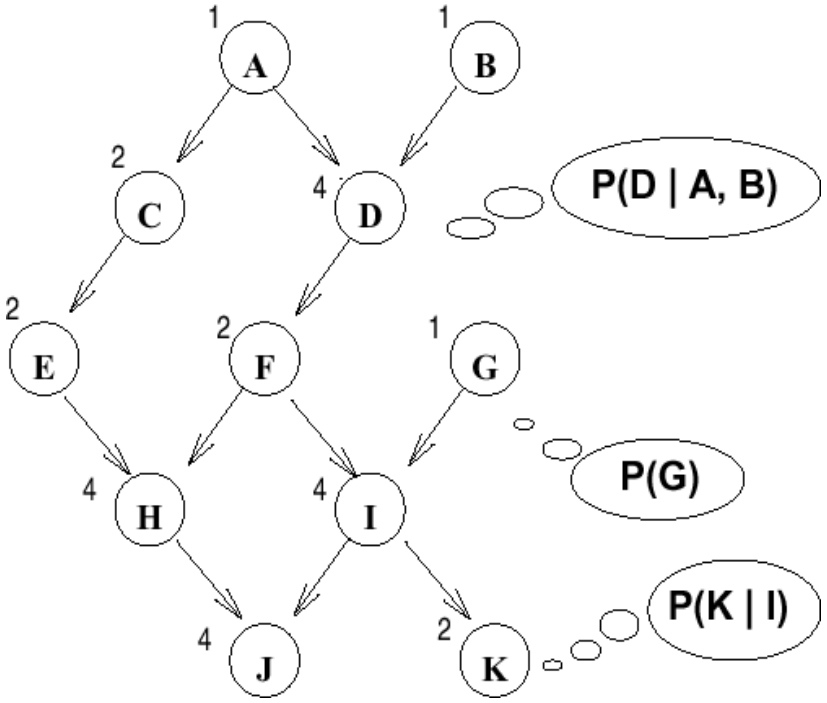


基于因果关系的变量序列可以获得更加自然紧致的贝叶斯网络结构

Example (Binary valued Variables)

Pr(A,B,C,D,E,F,G,H,I,J,K) =

Pr(A)
x Pr(B)
x Pr(C|A)
x Pr(D|A,B)
x Pr(E|C)
x Pr(F|D)
x Pr(G)
x Pr(H|E,F)
x Pr(I|F,G)
x Pr(J|H,I)
x Pr(K|I)



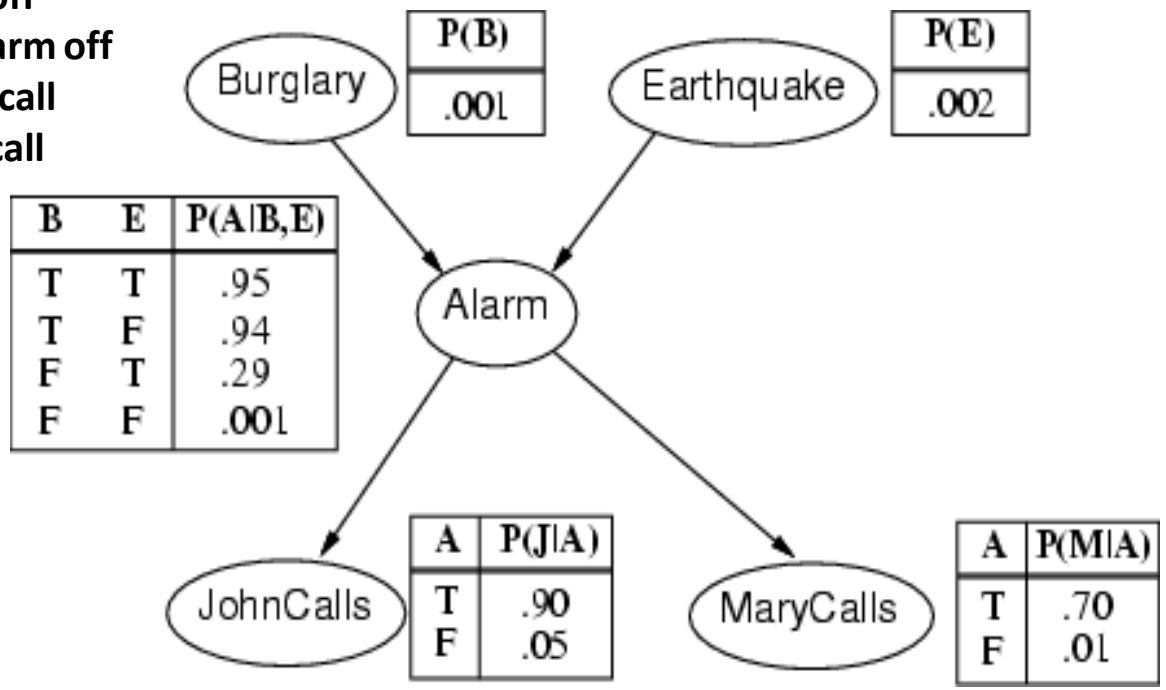
The Classic Burglary Example

- I'm at work, neighbour John calls to say my alarm is ringing, but neighbour Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?
- Variables: **Burglary入室盗窃, Earthquake, Alarm, JohnCalls, MaryCalls**
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off The
  - alarm can cause Mary to call
  - The alarm can cause John to call

Burglary example

- **A burglary can set the alarm off**
- **An earthquake can set the alarm off**
- **The alarm can cause Mary to call**
- **The alarm can cause John to call**

Note that these tables only provide the probability that Xi is true.
(E.g., Pr(A is true|B,E))
The probability that Xi is false is 1- these values



Number of Parameters: 1 + 1 + 4 + 2 + 2 = **10** (vs. 2^5-1 = **31**)

Using the Bayes Net:

$$Pr(B, E, A, J, M) = Pr(M|A)\, Pr(J|A)\, Pr(A|B, E) Pr(B) Pr(E)$$

Example of Constructing Bayes Network

- Previously we chose a causal order.
- Now suppose we choose the ordering MaryCalls (M), JohnCalls (J), Alarm (A), Burglary(B), Earthquake(E), *i.e.*, M,J,A,B,E
- These "orderings" are the ordering of the arrows in the Bayes Net DAG, which are the opposite to the ordering of variables in the chain rule, *i.e.*,
  $Pr(M,J,A,B,E) = Pr(E|B,A,J,M)*Pr(B|A,J,M)*$
  $Pr(A|J,M)*Pr(J|M)*Pr(M)$
- Now let's see if we can get rid of the conditioning sets

Burglary Example

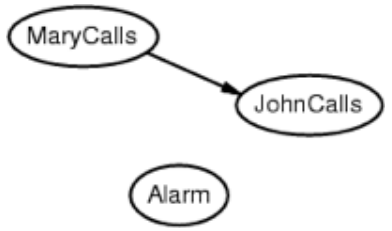
Suppose we choose the ordering M, J, A, B, E

P(J | M) = P(J)?

$$Pr(E, B, A, J, M) = Pr(E|B, A, J, M) * Pr(B|A, J, M) * Pr(A|J, M) * \mathbf{Pr(J|M)} * Pr(M)$$

Burglary Example

Suppose we choose the ordering M, J, A, B, E



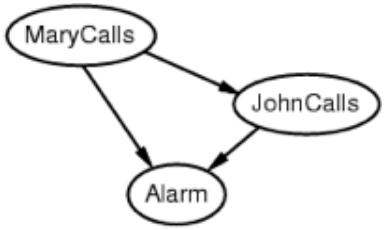
P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A| J, M) = P(A)?

$$Pr(E, B, A, J, M) = Pr(E|B, A, J, M) * Pr(B|A, J, M) * \mathbf{Pr(A|J, M)} * Pr(J|M) * Pr(M)$$

Burglary Example

Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A| J, M) = P(A)? No

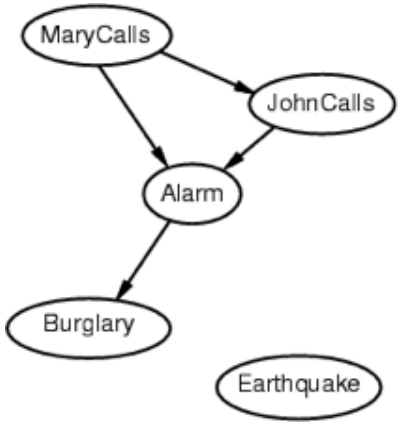
P(B | A, J, M) = P(B |A)?

P(B | A, J, M) = P(B)?

$$Pr(E, B, A, J, M) = Pr(E|B, A, J, M) * \mathbf{Pr(B|A, J, M)} * Pr(A|J, M) * Pr(J|M) * Pr(M)$$

Burglary Example

Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A| J, M) = P(A)? No

P(B | A, J, M) = P(B | A)? Yes

P(B | A, J, M) = P(B)? No

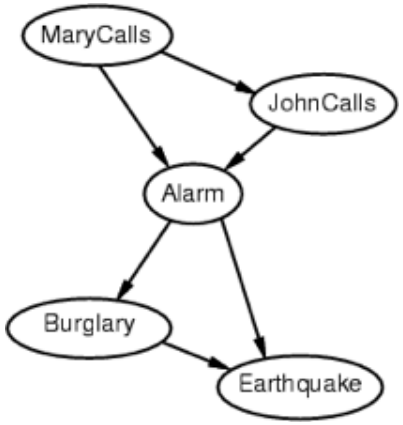
P(E | B, A ,J, M) = P(E |A)?

P(E | B, A, J, M) = P(E | A, B)?

$$Pr(E, B, A, J, M) = Pr(E|B, A, J, M) * Pr(B|A, J, M) * Pr(A|J, M) * Pr(J|M) * Pr(M)$$

Burglary Example

Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)? No

P(A | J, M) = P(A | J)? P(A| J, M) = P(A)? No

P(B | A, J, M) = P(B | A)? Yes

P(B | A, J, M) = P(B)? No

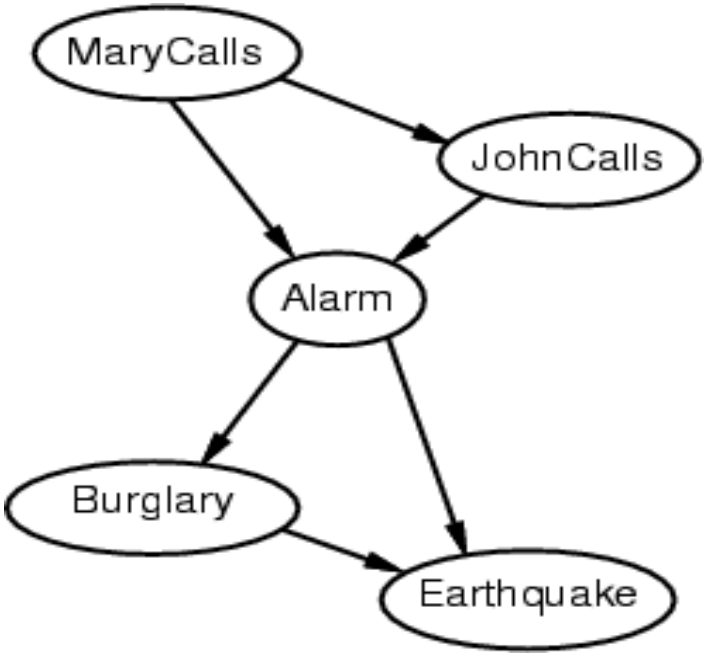
P(E | B, A ,J, M) = P(E | A)? No

P(E | B, A, J, M) = P(E | A, B)? Yes

$$Pr(E, B, A, J, M) = Pr(E|B, A, J, M) * Pr(B|A, J, M) * Pr(A|J, M) * Pr(J|M) * Pr(M)$$

Example cont'd

Deciding conditional independence is hard in the non-causal direction!
Causal models & conditional independence seem hardwired for humans.
Network is **less compact**: 1 + 2 + 4 + 2 + 4 = 13 numbers needed!



Inference in Bayes Nets 贝叶斯网络推断

Given

1) a **Bayes net**

$$\mathbf{Pr(X_1, X_2, \ldots, X_n)}$$

$$\mathbf{= Pr(X_n \mid Par(X_n)) * Pr(X_{n-1} \mid Par(X_{n-1}) * \cdots * Pr(X_1 \mid Par(X_1))}$$

2) some **Evidence**, E

$$\mathbf{E = \{a\ set\ of\ values\ for\ some\ of\ the\ variables\}}$$

We want to **compute the new probability distribution**

$$\mathbf{Pr(X_k \mid E)}$$

That is, we want to figure out

$$\mathbf{Pr(X_k = d \mid E)\ for\ all\ d \in Dom[X_k]}$$

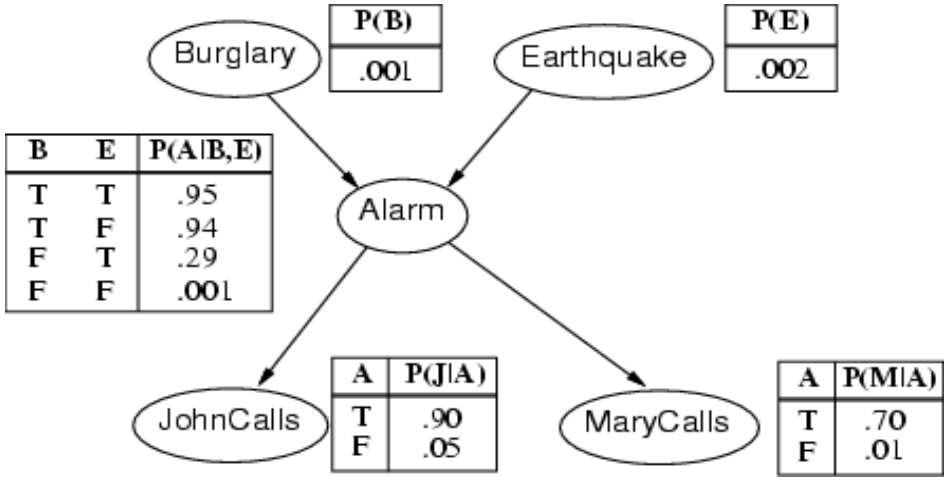
Examples

- **Computing probability of different diseases given symptoms症状**
- **Computing probability of hail storms 冰雹 given different metrological evidence**

In such cases getting a good estimate of the probability of the unknown event allows us to respond more effectively (gamble rationally)

Burglary example 回到入室盗窃例子

In the Alarm example*** we have (the compact network):



Recall Burglary(B), Earthquake(E), Alarm(A), MaryCalled (M), JohnCalled(J) And from our Bayes Net above, we determined:

Pr(B,E,A,M,J) = P(J|A) * Pr(M|A) * Pr(A|E,B) *Pr(E) * Pr(B)

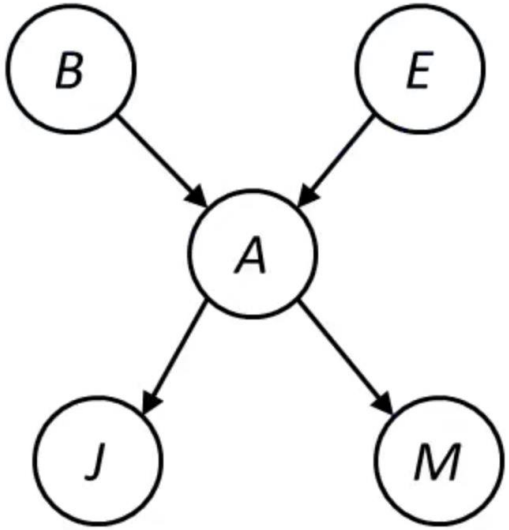
We might want to compute things like:

Pr(B=True| M=true, J=false, E=false)

The probability that there was a burglary, given that Mary called, John didn't call, and there was no earthquake

Burglary example calculation 试一下

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$

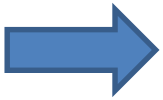
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$

$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

It tells that **independence** between variables can reduce the size of calculation!

Recall conditional independence

- Similarly
  - B is independent of E, and C, given A
  - A is independent of E, given C
- This means that:
  - $Pr(H|B,\{A,C,E\}) = Pr(H|B)$
  - $Pr(B|A,\{C,E\}) = Pr(B|A)$
  - $Pr(A|C,\{E\}) = Pr(A|C)$
  - $Pr(C|E)$ and $Pr(E)$ don't simplify



Based on the independent
$Pr(H,B,A,C,E) = Pr(H|B)*Pr(B|A)*Pr(A|C)*Pr(C|E)*Pr(E)$
Simple!

- Example: $Alarm \perp\!\!\!\perp Fire|Smoke$

火-》烟-》警报