频率学派 vs. 贝叶斯学派
贝叶斯学派
Probability（概率）：
独立性/条件独立性：
Probability Theory（概率论）：
Graphical models （概率图模型）
什么是图模型（Graphical Models）
图是什么
计算机科学中的图模型：
为什么图模型有用?
图模型：统一框架
图模型在机器学习中的作用：
图的方向性：
贝叶斯网络
举例说明：
举例说明：
Compactness（紧致性）
全局语义
局部语义
因果链
共同原因
共同效应
构建贝叶斯网络
构建贝叶斯网络举例
因果方向
因果性?
贝叶斯网络中的推理
推理任务
枚举推理
枚举推理举例
枚举效率不高
变量消元
精确推理的复杂度
举例：朴素贝叶斯模型
举例：垃圾邮件检测
举例：数字识别器
对朴素贝叶斯模型的评价：

频率学派 vs. 贝叶斯学派

频率学派:

- 概率是事件发生的长期预期频率。
- P(A) = n/N，其中n是事件A在N次机会中发生的次数。
- "某事发生的概率是0.1"意味着0.1是在无穷多样本的极限条件下能够被观察到的比例。
- 在许多情况下，不可能进行重复实验。
- 例如问题：第三次世界大战发生的概率是多少?

贝叶斯学派

- 概率是信念的度量。
- 它是一种基于不完全知识给出事件可能性的度量。

- 贝叶斯分析从先验信念开始，根据新的数据更新这种信念。
- 贝叶斯概率的主观性可能是一个限制，因为不同的人可能有不同的先验信念，并且可能根据相同的数据以不同的方式更新他们的信念。

Probability（概率）：

- Probability（概率）是对不确定知识一种严密的形式化方法。
- 它提供了一种量化不同事件或结果的可能性的方式。
- 全联合概率分布指定了对随机变量的每种完全赋值，即每个原子事件的概率。
- 可以通过把对应于查询命题的原子事件的条目相加的方式来回答查询。
- 对于复杂的领域，联合分布可能会变得过于复杂，我们必须找到一种方法来减少它的大小。
- 独立性和条件独立性提供了分解联合分布和简化计算的工具。

独立性/条件独立性：

- 当且仅当$P(A|B)=P(A)$，或$P(B|A)=P(B)$，或$P(A,B)=P(A)P(B)$时，A和B是独立的。
- 如果$P(A|B,C)=P(A|C)$，则在给定C的条件下，A对于B是条件独立的。
- 在大多数情况下，使用条件独立性可以将全联合概率的表示从指数关系减少为线性关系。
- 条件独立性是我们关于不确定环境最基本和最强大的知识形式。
- 它可以简化复杂模型和更有效地进行推断。

Probability Theory（概率论）：

- 概率论可以用两个简单的方程式表达:
  - 加法规则（Sum Rule）：$P(X)=\sum_Y P(X,Y)$，通过边缘化或求和其他变量来获得变量的概率。
  - 乘法规则（Product Rule）：$P(X,Y)=P(X|Y)P(Y)$，联合概率可以用条件概率表达。
- 所有的概率推断和学习都可以归结为不断应用加法规则和乘法规则。

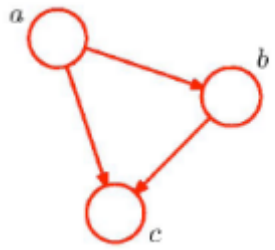
Graphical models （概率图模型）

什么是图模型（Graphical Models）

- 图模型是概率分布的图表表示。
- 它是概率论和图论的结合。
- 也被称为概率图模型（Probabilistic Graphical Models）。
- 它们增强了分析，而不是使用纯代数。

图是什么

- 由节点（也称为顶点）和链接（也称为边或弧）组成。



- 在概率图模型中，
  - 每个节点表示一个随机变量（或一组随机变量）。
  - 链接表示变量之间的概率关系。

计算机科学中的图模型：

- 处理不确定性和复杂性的自然工具，这些概念贯穿应用数学和工程。
- 图模型的基本思想是模块化，即通过组合较简单的部分来构建复杂的系统。
- 图模型为许多领域提供了一种有效的建模和推断方法，如人工智能、机器学习、计算机视觉和自然语言处理等。

为什么图模型有用？

- 概率理论提供了粘合剂，通过它，各部分得以结合，从而确保整个系统的一致性，并为模型与数据之间提供接口。
- 图论方面提供了：
  - 直观的可视化界面，使人类能够对高度交互的变量集进行建模。
  - 数据结构，自然地适合设计高效的通用算法。
- 图模型为许多领域提供了一种有效的建模和推断方法，如人工智能、机器学习、计算机视觉和自然语言处理等。

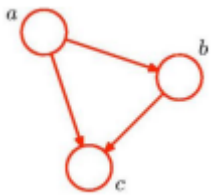
图模型：统一框架

- 将经典的多元概率系统视为共同的基础形式，如混合模型、因子分析、隐马尔可夫模型、卡尔曼滤波器等。
- 在系统工程、信息理论、模式识别和统计力学等领域中遇到。
- 观点的优点：
  - 可以在不同领域之间转移和利用特定技术。
  - 为设计新系统提供自然框架。
- 图模型为许多领域提供了一种有效的建模和推断方法，如人工智能、机器学习、计算机视觉和自然语言处理等。

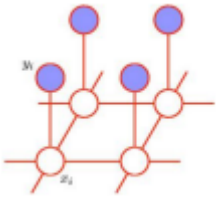
图模型在机器学习中的作用：

1. 形象化概率模型的结构，提供了一种简单的可视化方式。
2. 通过检查图，可以深入了解模型的属性，如条件独立性属性。
3. 需要进行推断和学习的复杂计算可以表示为图操作，从而简化了计算过程。

图的方向性：

- 有向图模型
  - 箭头表示方向性。
  - 贝叶斯网络 

    - 表示随机变量之间的因果关系。
  - 在人工智能和统计学中更为流行。
- 无向图模型
  - 没有箭头的链接。

- 马尔科夫随机场 
    - 更适合表示变量之间的软约束。
- 在视觉和物理学中更为流行。

贝叶斯网络

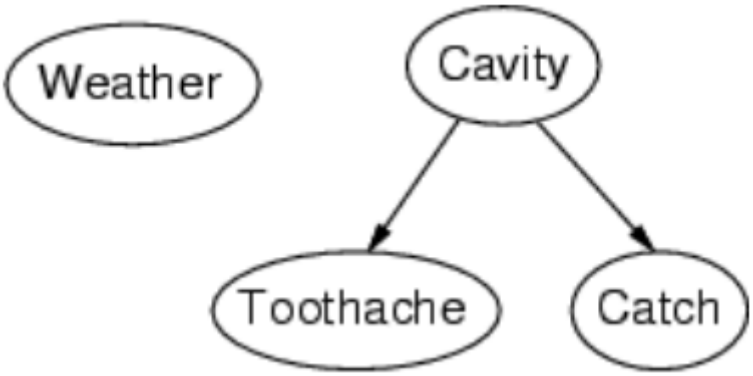
贝叶斯网络是一种简单的、图形化的数据结构，用于表示变量之间的依赖关系（条件独立性），为任何全联合概率分布提供一种简明的规范。

- 语法：
    - 一个节点对应一个变量。
    - 一个有向无环图（DAG）（链接≈“直接影响”）。
    - 每个节点都有一个条件分布，给定其父节点的条件下的概率分布：$P(X_i|Parents(X_i))$，量化其父节点对该节点的影响。
- 在最简单的情况下，条件分布可以表示为一个**条件概率表（CPT）**，给出每个父节点值组合下Xi的分布。

举例说明：

网络的拓扑结构编码了条件独立性的断言：

- 天气与其他变量无关。
- 在给定牙齿蛀牙的情况下，牙痛和牙感染（Catch）是条件独立的。



举例说明：

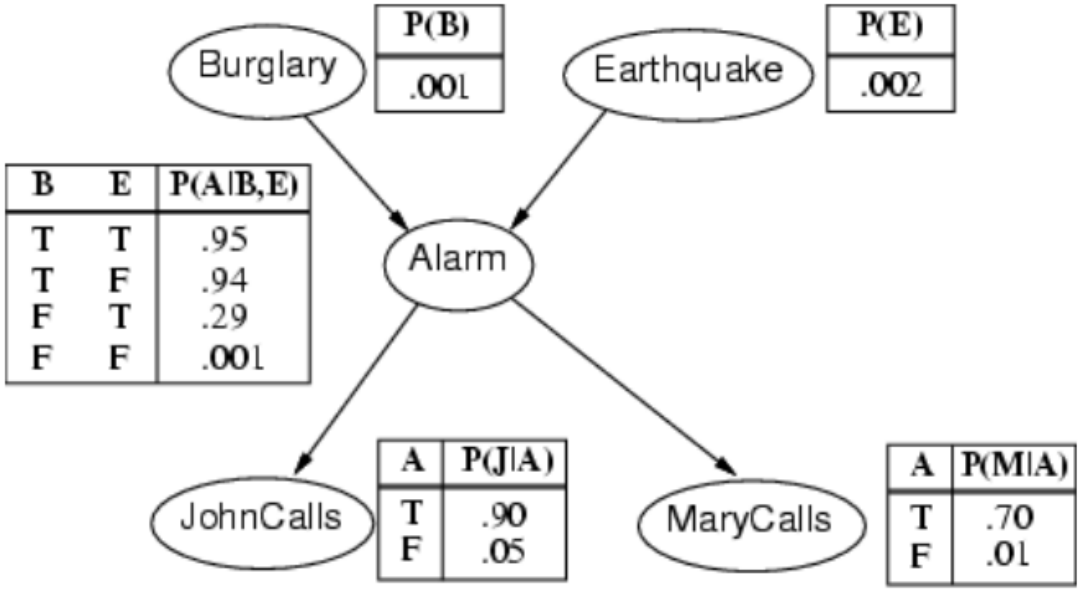
我正在工作，邻居约翰（John）打电话说我的闹钟响了，但邻居玛丽（Mary）没有打电话。有时它会因为小的地震而触发。这是不是有夜贼？

变量：入室行窃、地震、闹钟、約翰打电话、玛丽打电话。
网络拓扑反映了“因果”知识：

- 夜贼可以触发闹钟。
- 地震可以触发闹钟。
- 闹钟可以导致玛丽打电话。
- 闹钟可以导致约翰打电话。
    - P(Burglary)
    - P(Earthquake)

    - P(Alarm | Burglary, Earthquake)
    - P(JohnCalls | Alarm)
    - P(MaryCalls | Alarm)
- 给定观测到的变量，可以进行推断，例如：
    - 如果John报警，那么入室行窃的后验概率是多少？
    - 如果没有人打电话，那么地震的后验概率是多少？
    - 如果没有地震，也没有报警，那么入室行窃的后验概率是多少？

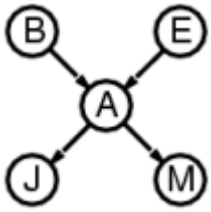
通过贝叶斯网络，可以直观地表达变量之间的依赖关系和条件独立性，从而进行推断和决策。



Compactness（紧致性）

一个具有k个布尔父节点的布尔变量的条件概率表中有2^k个独立的可指定概率，对应于父节点值的组合。

每一行需要一个数字p，表示$X_i = true$的概率（$X_i = false$的概率是$1-p$）。



如果每个变量的父节点不超过k个，则完整的网络需要$O(n \cdot 2^k)$个数字。与完整的联合分布相比，其增长率是线性的，而不是$O(2^n)$。

例如，对于入室行窃网络，有$1+1+4+2+2=10$个数字（而全联合分布有$2^5-1=31$）。

全局语义

在贝叶斯网络中，全联合概率分布可以表示为所有变量的条件概率分布的乘积，即：

$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$

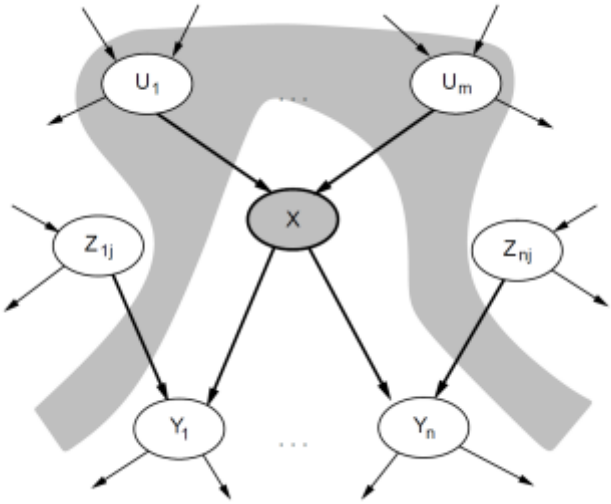
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$
\begin{aligned}
&= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\
&= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\
&\approx 0.00063
\end{aligned}
$$

这个等式表明，联合概率分布可以完全由局部条件概率分布确定。这意味着，贝叶斯网络提供了一种紧凑、可解释、易于推断的方式来表示联合概率分布。

局部语义

局部语义指的是，给定父节点，一个节点与它的非后代节点是条件独立的。换句话说，每个节点只依赖于它的父节点和自身，与其他节点无关。

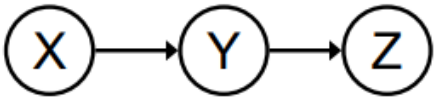


定理：局部语义 ⇔ 全局语义

即，全局语义中的联合概率分布可以由每个节点的条件概率分布表示，每个节点的条件独立性质可以保证全局条件独立性质。

因果链

- 基本结构



X: Low pressure

Y: Rain

Z: Traffic

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

  - 在给定Y的条件下，X是否与Z独立？

$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)} = \frac{P(x)P(y|x)P(z|y)}{P(x)P(y|x)}$$

$$= P(z|y) \qquad \textit{Yes!}$$

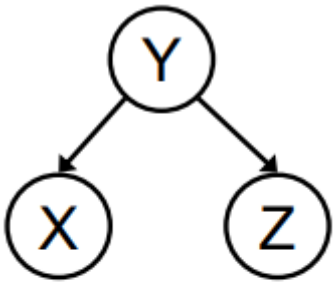
- **沿着链的证据“阻止”了影响。**

因果链是一种基本的因果结构，其中一个变量直接影响另一个变量，从而构成一个链。例如，如果X导致Y，然后Y导致Z，那么这个结构就是一个因果链。

在因果链中，如果给定Y的条件下，X和Z是独立的，那么我们可以说Y是一个“阻碍变量”，它“阻止”了X对Z的影响。这种阻碍效应是因果推断的基础，因为它们提供了关于变量间因果关系的信息。

共同原因

- 另一个基本结构：同一原因的两个影响



Y: Project due

X: Newsgroup busy

Z: Lab full

  - X和Z是否独立？
  - 在给定Y的条件下，X和Z是否独立？

$$P(z|x,y) = \frac{P(x,y,z)}{P(x,y)} = \frac{P(y)P(x|y)P(z|y)}{P(y)P(x|y)}$$

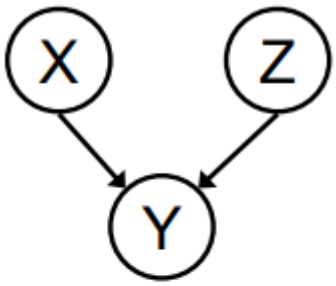
$$= P(z|y) \quad \textit{Yes!}$$

共同原因是另一个基本的因果结构，其中两个变量都受到同一原因的影响。例如，如果X和Z都是由Y导致的，那么这个结构就是一个共同原因。

在共同原因结构中，如果没有给定Y，那么X和Z可能会出现关联。但是，如果给定Y，则X和Z可能成为条件独立的，因为给定Y的情况下，它们的共同原因已经被控制。这种情况被称为“控制反应”，它提供了关于变量间因果关系的信息。

共同效应

• 最后一种结构：一个影响的两个原因（v-structures）



X: Raining

Z: Ballgame

Y: Traffic

- X和Z是否独立？
  • 是：记住球赛和下雨导致交通堵塞，没有相关性吗？
- 在给定Y的条件下，X和Z是否独立？
  • 不是：记住看到交通堵塞让雨和球赛处于竞争状态吗？

- **这与其他情况不同**
  - 观察效应可以使原因之间产生影响。

共同效应是一种因果结构，其中两个原因都可以导致同一效应。例如，如果X和Z都可以导致Y，那么这个结构就是一个共同效应。

在共同效应结构中，如果X和Z都是独立的，那么它们可能不会对Y产生影响，因为它们没有直接的关系。但是，如果给定了Y，则X和Z可能成为条件独立的，因为在给定Y的情况下，它们之间的影响路径被阻断了。这种情况被称为“掩蔽”，它提供了关于变量间因果关系的信息。

需要注意的是，观察到效应可能会导致原因之间产生影响，这与其他情况不同。例如，如果我们观察到交通堵塞，那么球赛和下雨可能会成为竞争因素，从而影响彼此的发生概率。这种情况是与其他情况相反的，因为观察到效应可以启用原因之间的影响。

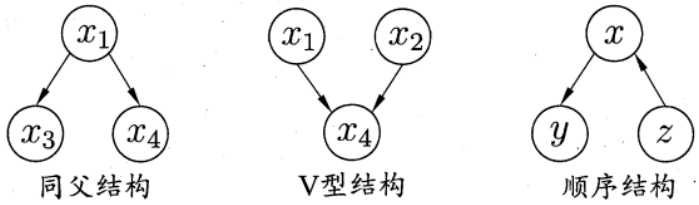


图 7.3 贝叶斯网中三个变量之间的典型依赖关系

在“同父”(common parent)结构中, 给定父结点 x_1 的取值, 则 x_3 与 x_4 条件独立. 在“顺序”结构中, 给定 x 的值, 则 y 与 z 条件独立. V 型结构(V-structure)亦称“冲撞”结构, 给定子结点 x_4 的取值, x_1 与 x_2 必不独立; 奇妙的是, 若 x_4 的取值完全未知, 则 V 型结构下 x_1 与 x_2 却是相互独立的. 我们做一个简单的验证:

$$
\begin{aligned}
P(x_1, x_2) &= \sum_{x_4} P(x_1, x_2, x_4) \\
&= \sum_{x_4} P(x_4 \mid x_1, x_2) P(x_1) P(x_2) \\
&= P(x_1) P(x_2) \ . \qquad (7.27)
\end{aligned}
$$

构建贝叶斯网络

需要一种方法使得局部的条件独立关系能够保证全局语义得以成立。

1. 选择变量X_1，X_2，...，X_n的顺序。
2. 对于$i = 1$到n，将X_i添加到网络中，并从X_1，X_2，...，X_{i-1}中选择父节点，使得条件概率符合以下条件：
   $P(X_i|Parents(X_i)) = P(X_i|X_1, \ldots, X_{i-1})$

这种父节点的选择方式可以保证全局语义成立：
$P(X_1, \ldots, X_n) = P(X_i|X_1, \ldots, X_{i-1}) = \prod_{i=1}^{n} P(X_i|Parents(X_i))$
（使用链法则、通过上述构建方法）

通过这种构建方式，可以保证贝叶斯网络中的条件独立关系与全局语义相一致。此外，这种构建方式还具有计算效率高、可解释性强等优点。

- 要求网络的拓扑结构确实反映了合适的父节点集对每个变量的那些直接影响。

- 添加节点的正确次序是首先添加“根本原因”节点，然后加入受它们直接影响的变量，以此类推。

构建贝叶斯网络举例

假设我们选择顺序为M，J，A，B，E。下面是对每个问题的解释：

1. $P(J|M) = P(J)$? No
   这个问题询问M是否影响J的概率。从网络结构来看，M是J的父节点，因此M会影响J的概率。因此，$P(J|M) \neq P(J)$。
2. $P(A|J, M) = P(A|J)$? No
   这个问题询问J和M是否一起影响A的概率，或者说是否存在直接影响A的路径上的变量被忽略了。从网络结构来看，A的父节点是M和J，因此M会影响A的概率。因此，$P(A|J, M) \neq P(A|J)$。
3. $P(B|A, J, M) = P(B|A)$? Yes
   这个问题询问是否存在一个变量，使得在给定其他相关变量的情况下，该变量与B的独立性与其他变量无关。从网络结构来看，B的父节点是M和E，而A和J不是B的父节点，因此在给定A和J的情况下，M和E对B的概率具有独立性。因此，$P(B|A, J, M) = P(B|A)$。
4. $P(B|A, J, M) = P(B)$? No
   这个问题询问B是否独立于M和J。从网络结构来看，B的父节点是M和E，而J和A不是B的父节点，因此在给定A和J的情况下，M和E对B的概率具有独立性，但是B的概率仍然会受到E的影响。因此，$P(B|A, J, M) \neq P(B)$。
5. $P(E|B, A, J, M) = P(E|A)$? No
   这个问题询问是否存在一个变量，使得在给定其他相关变量的情况下，该变量与E的独立性与其他变量无关。从网络结构来看，E的父节点是B，B对E的概率仍然有影响。因此，$P(E|B, A, J, M) \neq P(E|A)$。
6. $P(E|B, A, J, M) = P(E|A, B)$? Yes
   这个问题询问是否存在一个变量，使得在给定其他相关变量的情况下，该变量与E的独立性与其他变量无关。从网络结构来看，E的父节点是B，而J和M不是E的父节点，因此 $P(E|B, A, J, M) = P(E|A, B)$。

因果方向

决定非因果方向上的条件独立关系是困难的。因为在这种情况下，变量之间的关系可能是相互影响的，而不是单向因果关系。这使得我们不能简单地依靠因果模型和条件独立性来解决问题。

相比之下，**在因果方向上，因果模型和条件独立性是相对容易理解的**，因为它们反映了真实世界中变量之间的因果关系。这是因为我们的大脑天生被设计用来理解因果关系，而不是非因果关系。

此外，在**非因果方向上，构建网络所需的数字数量通常更多**，因为我们需要考虑所有可能的相互作用，而不仅仅是单向因果关系。例如，**在一个由5个变量组成的网络中，如果我们选择非因果方向，则需要13个数字来表示条件概率分布，而在因果方向上只需要5个数字。**

综上所述，虽然在非因果方向上决定条件独立关系是困难的，但因果模型和条件独立性仍然是人类天生熟悉的概念，可以帮助我们更好地理解复杂的关系网络。

数学公式：
在给定变量X和Y的情况下，如果变量Z与变量Y条件独立，则可以表示为：
$P(Z|X,Y) = P(Z|X)$

因果性?

- 当贝叶斯网络反映真实因果模式时：
  - 常常更简单（节点具有较少的父节点）
  - 常常更易于思考
  - 常常更容易从专家那里获取
- 贝叶斯网络不一定是因果的
  - 有时在领域中不存在因果网络（特别是如果变量缺失）
  - 最终得到的箭头反映相关性，而不是因果关系
- 箭头真正意味着什么？
  - 拓扑结构可能偶然编码了因果结构
  - **拓扑结构真正编码了条件独立性**

贝叶斯网络中的推理

推理任务

贝叶斯网络中通常需要进行以下三种推理任务：

1. 简单查询（Simple queries）：计算后验概率$P(X_i|E=e)$。例如，给定油表为空，车灯亮起且车辆未启动的情况下，计算没有汽油的概率
   $P(NoGas|Gauge_{油表}=empty, Lights=on, Starts=false)$。
2. 联合查询（Conjunctive queries）：计算$P(X_i, X_j|E=e) = P(X_i|E=e)P(X_j|X_i, E=e)$。
   这种查询涉及两个或多个变量的联合概率分布。
   这个等式成立是因为它基于条件概率的定义和贝叶斯定理。

> 根据条件概率的定义，我们有：　$P(X_i, X_j|E=e) = \frac{P(X_i,X_j,E=e)}{P(E=e)}$
> 接下来，我们可以将分子$P(X_i, X_j, E=e)$拆分为条件概率的形式，即：
> $P(X_i, X_j, E=e) = P(X_j|X_i, E=e)P(X_i, E=e)$　将上式代入分母$P(E=e)$中，得到：
> $P(X_i, X_j|E=e) = \frac{P(X_j|X_i,E=e)P(X_i,E=e)}{P(E=e)}$　接着，根据贝叶斯定理，我们可以将条件概率
> $P(X_i, E=e)$表示为：　$P(X_i, E=e) = P(X_i|E=e)P(E=e)$　代入上式中，得到：
> $P(X_i, X_j|E=e) = \frac{P(X_j|X_i,E=e)P(X_i|E=e)P(E=e)}{P(E=e)}$　化简后，即可得到：
> $P(X_i, X_j|E=e) = P(X_j|X_i, E=e)P(X_i|E=e)$

这就是等式$P(X_i, X_j|E=e) = P(X_i|E=e)P(X_j|X_i, E=e)$的推导过程。它表示在给定观测值$E=e$的条件下，变量$X_i$和$X_j$的联合概率分布可以表示为在给定$E=e$的条件下，变量$X_i$的条件概率分布和在给定$E=e$和$X_i$的条件下，变量$X_j$的条件概率分布的乘积。这个等式在贝叶斯网络中具有重要的应用价值，可以用于概率推断和条件概率分布的计算等任务。

3. 最优决策（Optimal decisions）：决策网络包括效用信息；需要概率推理以计算
   $P(outcome|action, evidence)$，其中outcome表示结果，action表示决策，evidence表示证据。
   这种推理任务通常用于在不确定性环境中做出最优决策。

这些推理任务可以帮助我们理解变量之间的概率关系，并用于许多实际应用，例如医学诊断、金融风险预测和自然语言处理等。

枚举推理

枚举推理是一种在贝叶斯网络中进行推理的方法，它通过计算条件概率的乘积并求和来回答查询，而不需要显式构建联合概率分布的完整表示。

在贝叶斯网络中，可以使用条件概率乘积的形式表示全联合概率分布。具体来说，假设我们有一个由变量$X_1, X_2, \ldots, X_n$组成的贝叶斯网络B，则可以将全联合概率分布$P(X_1, X_2, \ldots, X_n)$表示为：
$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Parent(X_i))$
其中，$Parent(X_i)$表示变量X_i的父节点集合。

> 这个公式利用了条件独立性的性质，即在给定父节点的情况下，每个变量的条件概率只依赖于其父节点，而与其他变量无关。因此，我们可以将全联合概率分布表示为每个变量的条件概率的乘积，从而简化计算的复杂度。
> 这个公式可以用于进行贝叶斯网络的推理，例如计算给定证据的情况下某个变量的后验概率。

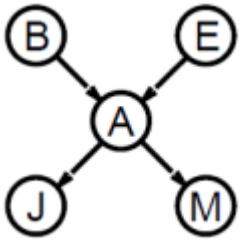
枚举推理的基本思想是枚举网络中的每个变量，对每个变量计算其条件概率，并将它们相乘。然后，将所有变量的条件概率乘积相加，得到所需的概率分布。在这个过程中，枚举推理使用了贝叶斯定理和条件独立性的性质，对计算过程进行了简化。

枚举推理是一种比较直观的方法，可以在计算条件概率时避免显式地构建完整的联合概率分布。然而，当网络结构较为复杂时，枚举推理的计算复杂度会非常高，因此需要使用其他更高效的推理方法，例如变量消元和近似推理等。

总的来说，枚举推理是一种简单而直观的推理方法，可以用于解决许多实际问题，但在处理大型网络时可能会面临计算复杂度高的问题。

枚举推理举例

从“偷窃”网络简单查询



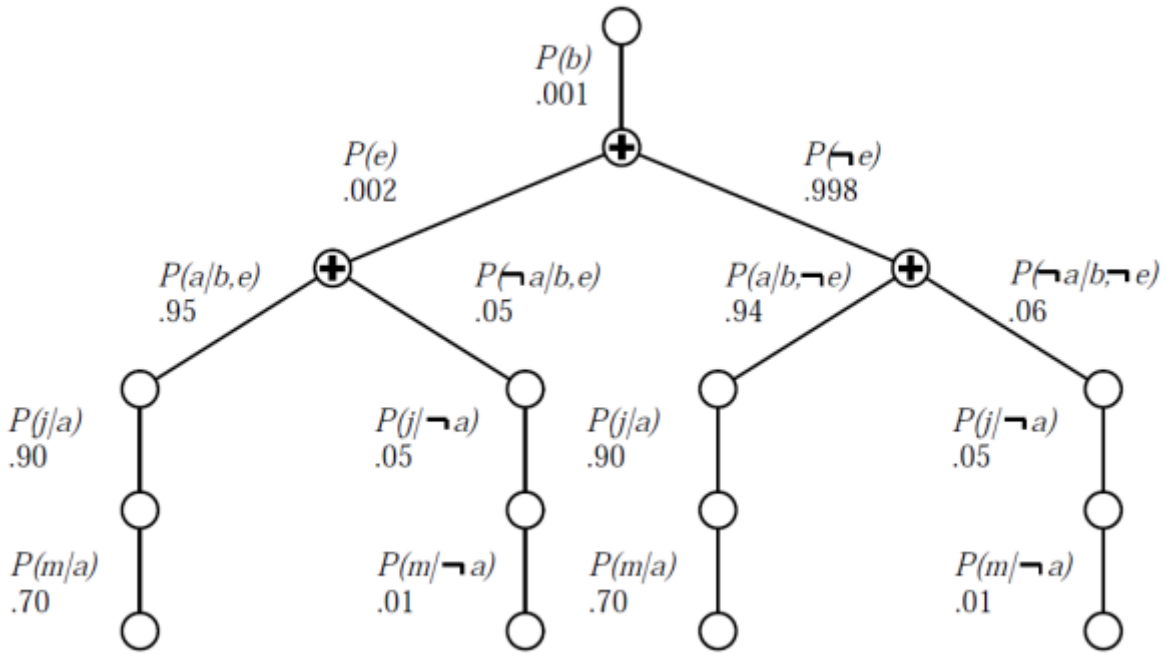
$$
\begin{aligned}
&\mathbf{P}(B|j,m) \\
&= \mathbf{P}(B,j,m)/P(j,m) \\
&= \alpha \mathbf{P}(B,j,m) \\
&= \alpha \sum_e \sum_a \mathbf{P}(B,e,a,j,m)
\end{aligned}
$$

使用条件概率表项重写全联合项

$$
\begin{aligned}
&\mathbf{P}(B|j,m) \\
&= \alpha \sum_e \sum_a \mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a) \\
&= \alpha \mathbf{P}(B) \sum_e P(e) \sum_a \mathbf{P}(a|B,e)P(j|a)P(m|a)
\end{aligned}
$$

递归深度优先搜索：$O(n)$空间复杂度，$O(d^n)$时间复杂度

枚举效率不高



重复计算$P(j|a)P(m|a)$对于e的每个取值

变量消元

变量消元是贝叶斯网络中一种常用的推理方法，它通过从右到左进行求和，并存储中间结果（因子）以避免重新计算，来简化计算复杂度。

变量消元的基本思想是将网络中的变量根据其与待求变量的关系进行分组，将每组变量的条件概率分布乘起来，得到一个新的因子。然后，对于不需要的变量，可以将它们从因子中消去，最终得到一个只包含待求变量的因子。通过对这个因子进行归一化，我们可以得到待求变量的后验概率分布。

变量消元可以通过多种方式实现，例如求和-乘积算法和置信传播算法等。其中，求和-乘积算法是最常用的变量消元算法之一。该算法使用因子表示联合概率分布，通过从右到左进行求和的方式，逐步消去不需要的变量，最终得到只包含待求变量的因子。

变量消元是一种高效且常用的推理方法，可以用于计算贝叶斯网络中任意变量的后验概率分布。

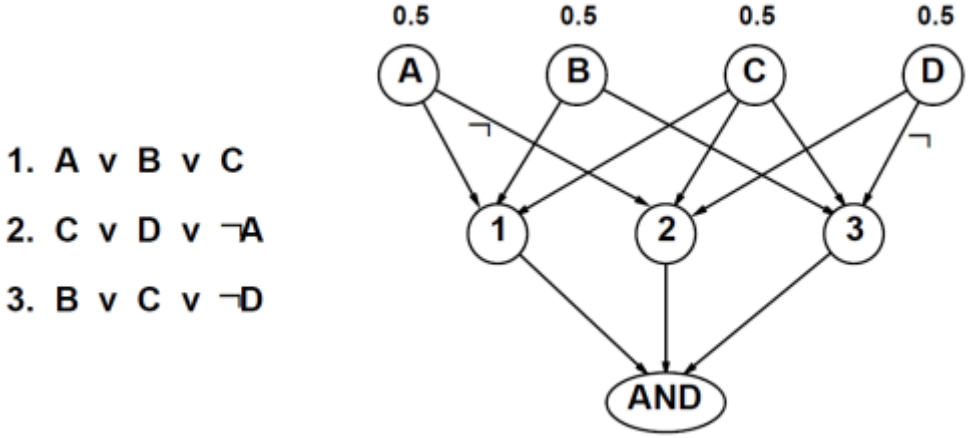
$$
\begin{aligned}
&\mathbf{P}(B|j,m) \\
&\quad = \alpha \underbrace{\mathbf{P}(B)}_{B} \Sigma_e \underbrace{P(e)}_{E} \Sigma_a \underbrace{\mathbf{P}(a|B,e)}_{A} \underbrace{P(j|a)}_{J} \underbrace{P(m|a)}_{M} \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) P(j|a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a \mathbf{P}(a|B,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) \Sigma_a f_A(a,b,e) f_J(a) f_M(a) \\
&\quad = \alpha \mathbf{P}(B) \Sigma_e P(e) f_{\bar{A}JM}(b,e) \text{ (sum out } A\text{)} \\
&\quad = \alpha \mathbf{P}(B) f_{\bar{E}\bar{A}JM}(b) \text{ (sum out } E\text{)} \\
&\quad = \alpha f_B(b) \times f_{\bar{E}\bar{A}JM}(b)
\end{aligned}
$$

精确推理的复杂度

在贝叶斯网络中，精确推理的复杂度取决于网络的结构和大小。对于单联通网络（或多树），变量消元的时间和空间复杂度都与网络规模呈线性关系，具体地说，是$O(d^k n)$，其中d是每个节点的最大父节点数，k是每个节点的最大取值数，n是网络中的节点数。

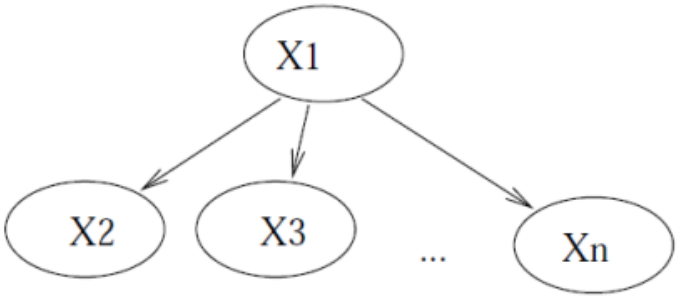
对于多联通网络，精确推理的复杂度可能非常高，实际上，它可以被归约到解决3SAT问题上，从而是NP难的。此外，计算多联通网络的模型数量等价于计数3SAT模型的数量，这被称为#P完全问题。

总的来说，精确推理在单联通网络（或多树）上是高效的，但在多联通网络上可能非常困难，需要使用近似推理或其他高级技术来解决。



举例：朴素贝叶斯模型

朴素贝叶斯模型是一种基于贝叶斯定理的分类算法，它假设每个特征与其他特征相互独立。这使得朴素贝叶斯模型的计算复杂度低，同时在许多实际应用中表现出色。



$$\begin{aligned}&\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n)\\ &\quad = \quad \mathrm{P}(X_1 = x_1)\,\mathrm{P}(X_2 = x_2 | X_1 = x_1)\ \cdots\ \mathrm{P}(X_n = x_n | X_1 = x_1)\end{aligned}$$

在朴素贝叶斯模型中，假设有一个类别变量Y和n个特征变量$X_1, X_2, \ldots, X_n$。朴素贝叶斯模型假设每个特征变量X_i与类别变量Y相互独立，但在给定类别变量Y的情况下，每个特征变量X_i可能的取值是有条件依赖的。因此，我们可以使用贝叶斯定理来计算给定类别变量Y和特征变量$X_1, X_2, \ldots, X_n$的联合概率分布：

$P(Y, X_1, X_2, \ldots, X_n) = P(Y) \prod_{i=1}^{n} P(X_i|Y)$

其中，$P(Y)$是类别变量Y的先验概率分布，$P(X_i|Y)$是在给定类别变量Y的情况下，特征变量X_i的条件概率分布。

可以使用朴素贝叶斯模型进行分类，即给定一个未知的特征向量$X = (X_1, X_2, \ldots, X_n)$，我们可以计算每个类别变量$Y$的后验概率分布$P(Y|X)$，然后选择后验概率最大的类别作为预测结果。

总的来说，朴素贝叶斯模型是一种简单而有效的分类算法，适用于许多实际应用场景，如文本分类、垃圾邮件过滤等。

举例：垃圾邮件检测

假设我们要解决自动检测垃圾邮件的问题。一个简单的起点是仅查看邮件中的“主题：”头，并通过检查一些简单的可计算特征来尝试识别垃圾邮件。我们考虑的两个简单特征是：

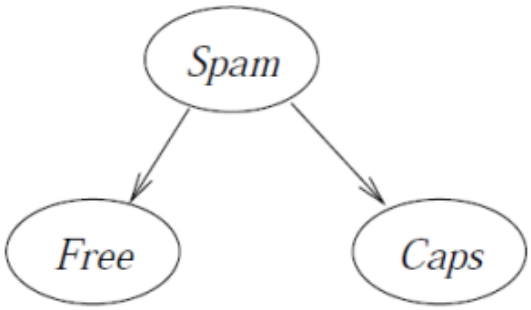
Caps：主题头是否全部大写
Free：主题头是否包含单词“free”，无论是大写还是小写

该模型基于以下三个随机变量：Caps（是否全大写）、Free（是否包含单词“free”）和Spam（是否为垃圾邮件），每个变量都取值为Y（是）或N（否）。

具体来说，当且仅当邮件主题不含小写字母时，Caps = Y；当且仅当邮件主题中包含单词“free”（不区分大小写）时，Free = Y；当且仅当邮件是垃圾邮件时，Spam = Y。

我们可以使用联合概率分布来描述这三个变量之间的关系，即：

$$P(Free, Caps, Spam) = P(Spam)P(Caps|Spam)P(Free|Spam)$$



P(*Free* , *Caps*, *Spam*)= P(*Spam*) P(*Caps*|*Spam*) P(*Free*|*Spam*)

Free	*Caps*	*Spam*	# messages
Y	Y	Y	20
Y	Y	N	1
Y	N	Y	5
Y	N	N	0
N	Y	Y	20
N	Y	N	3
N	N	Y	2
N	N	N	49
		Total:	100

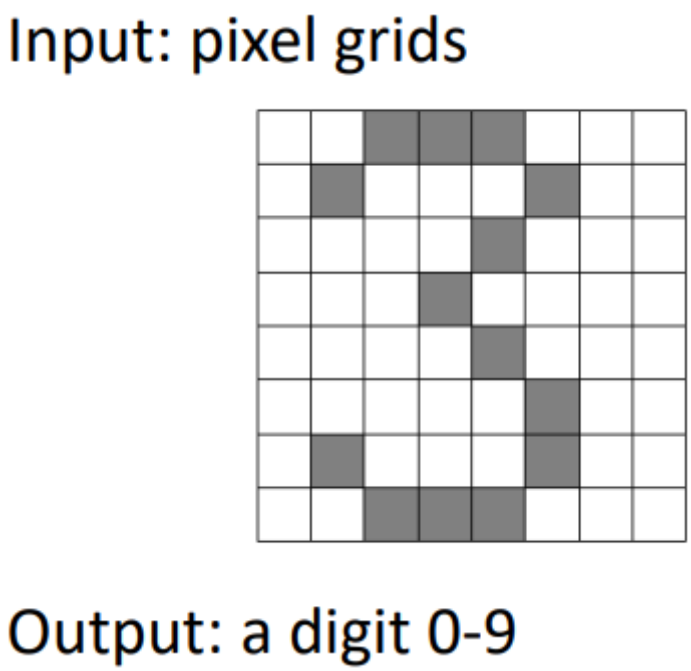
Spam	P(*Spam*)
Y	$\frac{20+5+20+2}{100} = 0.47$
N	$\frac{1+0+3+49}{100} = 0.53$

Caps	*Spam*	P(*Caps*|*Spam*)
Y	Y	$\frac{20+20}{20+5+20+2} \approx 0.8511$
Y	N	$\frac{1+3}{1+0+3+49} \approx 0.0755$
N	Y	$\frac{5+2}{20+5+20+2} \approx 0.1489$
N	N	$\frac{0+49}{1+0+3+49} \approx 0.9245$

Free	*Spam*	P(*Free*|*Spam*)
Y	Y	$\frac{20+5}{20+5+20+2} \approx 0.5319$
Y	N	$\frac{1+0}{1+0+3+49} \approx 0.0189$
N	Y	$\frac{20+2}{20+5+20+2} \approx 0.4681$
N	N	$\frac{3+49}{1+0+3+49} \approx 0.9811$

$$\begin{aligned}
&\mathrm{P}(Free = Y, Caps = N, Spam = N) \\
&\quad = \mathrm{P}(Spam = N)\ \mathrm{P}(Caps = N|Spam = N)\ \mathrm{P}(Free = Y|Spam = N) \\
&\quad \approx 0.53 \times 0.9245 \times 0.0189 \\
&\quad \approx 0.0093
\end{aligned}$$

举例：数字识别器



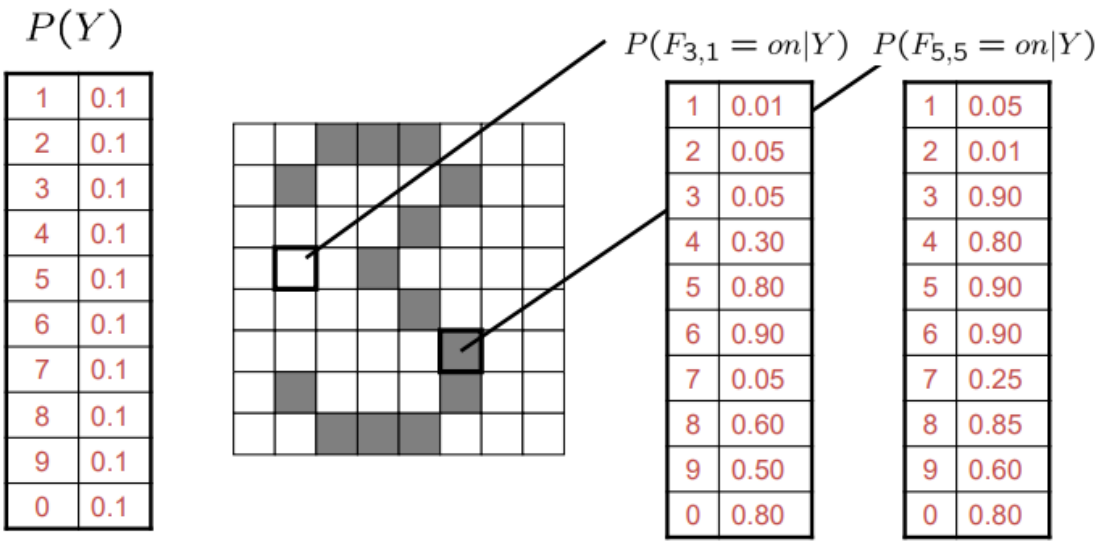
- 简单版本：
  - 每个网格位置<i,j>都有一个特征Fij
  - 可能的特征值是开/关，基于底层图像中强度是否大于0.5
  - 每个输入对应一个特征向量，例如：

$$\rightarrow \langle F_{0,0}=0 \;\; F_{0,1}=0 \;\; F_{0,2}=1 \;\; F_{0,3}=1 \;\; F_{0,4}=0 \;\; \ldots F_{15,15}=0 \rangle$$

  - 这里有很多特征，每个特征都是二进制的
  - 朴素贝叶斯模型:

$$P(Y|F_{0,0}\ldots F_{15,15}) \propto P(Y)\prod_{i,j} P(F_{i,j}|Y)$$

- 条件概率表



1	0.1
2	0.1
3	0.1
4	0.1
5	0.1
6	0.1
7	0.1
8	0.1
9	0.1
0	0.1

1	0.01
2	0.05
3	0.05
4	0.30
5	0.80
6	0.90
7	0.05
8	0.60
9	0.50
0	0.80

1	0.05
2	0.01
3	0.90
4	0.80
5	0.90
6	0.90
7	0.25
8	0.85
9	0.60
0	0.80

对朴素贝叶斯模型的评价：

1. 通过假设条件独立性，使概率推理变得可行。这个假设虽然很强，但是在许多实际应用中都表现出了较好的效果。
2. 尽管使用了这个强假设，朴素贝叶斯模型在许多应用中表现出色，尤其是在文本分类等领域。
3. 实验表明，朴素贝叶斯模型在标准数据集上与其他分类方法相比具有相当的竞争力。
4. 特别是在文本分类中，例如垃圾邮件过滤，朴素贝叶斯模型非常受欢迎。

总的来说，朴素贝叶斯模型是一种简单而有效的分类算法，尤其适用于文本分类和垃圾邮件过滤等应用。虽然它有一个强假设，但在许多实际应用中，它表现得很好，并且在竞争性实验中也表现出良好的性能。