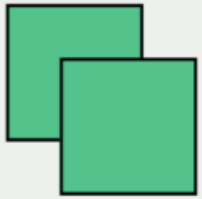
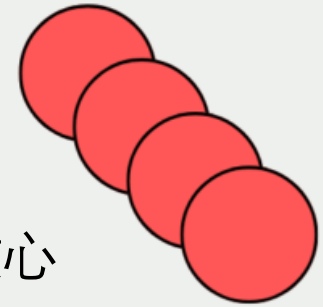
朴素贝叶斯

朴素贝叶斯分类器

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$





- **贝叶斯定理是**朴素贝叶斯分类器的核心

应用:

- 文本分类
- 垃圾邮件检测
- 情感分析

贝叶斯定理（Bayes' theorem）

- 贝叶斯定理是概率论中的重要的定理，用来计算在给定新证据或数据之后某个事件的概率（**后验概率**）。

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

- 例：我们可以通过癌症的总体平均风险来预测一个人患病的概率（**先验概率**）。使用贝叶斯定理之后，我们就可以进一步结合个体的具体信息（如年龄、性别、遗传因素等）来提供更精确的风险估计（**后验概率**），而不是仅仅依赖于总体平均风险。

先验概率/后验概率

- **先验概率（Prior Probability）**：这是在考虑任何相关证据或数据之前，根据以往经验或主观判断预估某事件发生的概率。

  例：如果某地区过去十年中每年都发生地震，那么我们可能会基于这一经验判断来年也会发生地震的概率较高，这就是先验概率。

- **后验概率（Posterior Probability）**：在观察或获取新的数据后，对事件发生概率的重新评估。

  例：如果在考虑了最近的地质活动，天气情况后，我们发现该地区的地震活动有所下降，那么我们可能会降低来年发生地震的概率评估，这就是后验概率。

分类任务下的先验概率/后验概率

- 先验概率 $P(C)$ 表示在任何特定观测数据X被考虑之前，事件 C（通常是一个类别标签或某种结果）的概率。（比如：给定如下数据集，现在有一个水果，问它属于每个类别的概率（Banana，Orange，Other））

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- 后验概率$P(C|X)$表示在给定观测数据X（即特征）的条件下，目标变量 C（即类别）的概率。（比如：我进一步给出水果特征为Long，Not sweet，Yellow，问它属于每个类别的概率（Banana，Orange，Other））

- 分类任务中核心的问题：在给定输入数据特征的情况下，预测该数据属于各个类别的概率 ⇨ **后验概率**。

朴素贝叶斯分类

假如我们的分类模型中的某一个样本是:

$$\mathbf{x} = (x_1, \ldots, x_n)$$

分类问题则可以转化为求如下后验概率的问题

$$p(C_k \mid x_1, \ldots, x_n)$$

$$p(C_k \mid \mathbf{x})$$

对于给定的特征X，可以通过计算后验概率$p(c|x)$来预测c (类别)

朴素贝叶斯分类

贝叶斯定理可以用来估计后验概率

根据贝叶斯定理，我们有

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

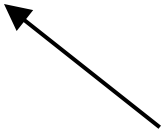
（后验概率）（先验概率）（似然）

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

（“证据”因子）

朴素贝叶斯分类

先验概率$p(C_k)$，由数据集D中的类别分布确定



$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

- $p(C_k)$可以直接通过给定的数据集来计算

朴素贝叶斯分类

似然代表了类别C_k背景下观测到样本X的概率，

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

- 如果数据量足够大，则可以准确地估计似然，但实际上样本数不足以准确估计出似然。
- 引入特征独立性假设来简化模型：朴素贝叶斯认为在**给定类别的条件下，所有特征之间是相互独立的。**那么根据条件独立性定义，似然可表示为：$p(x_1, x_2, \ldots, x_n \mid C_k) = p(x_1 \mid C_k) \times p(x_2 \mid C_k) \times \ldots \times p(x_n \mid C_k)$。

朴素贝叶斯分类

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

证据因子$p(x)$代表了整个数据集D背景下观测到样本x的概率，可以理解为$p(x|D)$

- 如果数据量足够大，则可以准确地估计证据因子，但实际上样本数不足以准确估计出证据因子。
- 因为我们只是想知道$p(C_0|x), p(C_1|x), \ldots, p(C_k|x)$这些后验概率中，哪个大哪个小的问题。不需要计算其具体的值。而这些后验概率拥有同一个分母（证据因子），所以证据因子 $p(\mathbf{x})$ **无需计算**。

朴素贝叶斯分类

先验概率$p(C_k)$，由数据集D中的类别分布确定

似然代表了类别C_k背景下观测到样本X的概率

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

证据因子$p(x)$代表了整个数据集D背景下观测到样本x的概率，可以理解为$p(x|D)$

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} = \frac{p(C_k)}{p(x)} \prod_{i=1}^{d} p(x_i, C_k)$$

朴素贝叶斯模型

$p(C_k, x_1, \ldots, x_n)$

模型估计的是类别和特征的联合概率分布

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

总结：

- 由给定数据集得到输入输出的联合概率分布模型 $p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$
- 将给定的某个样本x，输入到该联合概率分布模型，得到使后验概率最大的k

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

朴素贝叶斯分类

- 朴素贝叶斯分类器（NBC）是一种**基于贝叶斯定理**的分类方法，其核心假设是**所有特征在给定类别的条件下是相互独立的。** 在分类或预测阶段，给定一个新的输入向量X，NBC通过应用贝叶斯定理，计算并比较在每个可能的输出类别Y下的**后验概率**。最终，**模型将选择具有最大后验概率的类别作为预测结果。**

- **特征在给定类别的条件下是相互独立的：**然而在现实世界的数据中，**特征之间往往是相关的**。例如，在图像识别中，相邻像素之间就有很高的相关性；在文本分类中，某些词的出现也会影响其他词的出现。朴素贝叶斯算法忽略了这些可能的相关性，这种做法在某种意义上是“过于简化”或“朴素”的。

- 如果特征之间不独立，是不是朴素贝叶斯分类器就无法使用呢？

朴素贝叶斯分类

- 朴素贝叶斯分类器（NBC）是一种**基于贝叶斯定理**的分类方法，其核心假设是**所有特征在给定类别的条件下是相互独立的。** 在分类或预测阶段，给定一个新的输入向量X，NBC通过应用贝叶斯定理，计算并比较在每个可能的输出类别Y下的**后验概率**。最终，**模型将选择具有最大后验概率的类别作为预测结果。**

- **特征在给定类别的条件下是相互独立的：**然而在现实世界的数据中，**特征之间往往是相关的**。例如，在图像识别中，相邻像素之间就有很高的相关性；在文本分类中，某些词的出现可能会影响其他词的出现。朴素贝叶斯算法忽略了这些可能的相关性，这种做法在某种意义上是“过于简化”或“朴素”的。

- 如果特征之间不独立，是不是朴素贝叶斯分类器就无法使用呢？（否，数据集的特征之间相关性越高，朴素贝叶斯的分类效果就越差。但不代表朴素贝叶斯无法使用。）

朴素贝叶斯分类器的例子

计算例：

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

- 3个类别： banana, orange, other
- 3个特征: Long/Not long, Sweet/Not sweet, Yellow/Not yellow
- 1000个样本（500个banana, 300个orange, 200个other）

问题：现有一个长形状的、有甜味的、黄颜色的水果，你能猜测出这是什么水果吗?

朴素贝叶斯分类器的例子

根据贝叶斯公式:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

$$p(Banana|\, Long, Sweet, Yellow) = \frac{p(Banana) * p(Long, Sweet, Yellow|Banana)}{p(Long, Sweet, Yellow)}$$

$$p(Orange|\, Long, Sweet, Yellow) = \frac{p(Orange) * p(Long, Sweet, Yellow|Orange)}{p(Long, Sweet, Yellow)}$$

$$p(Other|\, Long, Sweet, Yellow) = \frac{p(Other) * p(Long, Sweet, Yellow|Other)}{p(Long, Sweet, Yellow)}$$

朴素贝叶斯分类器的例子

以Banana为例：

$$p(Banana|\ Long, Sweet, Yellow)$$

$$= \frac{p(Banana) * p(Long, Sweet, Yellow|Banana)}{p(Long, Sweet, Yellow)}$$

$$= \frac{p(Banana) * p(Long|Banana) * p(Sweet|Banana) * p(Yellow|Banana)}{p(Long, Sweet, Yellow)}$$

$$= \frac{0.5 * 0.8 * 0.7 * 0.9}{p(Long, Sweet, Yellow)} = \frac{1}{Z} * 0.252$$

朴素贝叶斯分类器的例子

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

$p(Orange|\ Long, Sweet, Yellow) = ?$

$p(Other|\ Long, Sweet, Yellow) = ?$

朴素贝叶斯分类器的例子

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

$$p(Orange|\ Long, Sweet, Yellow) = 0$$

$$p(Other|\ Long, Sweet, Yellow) = \frac{1}{z} * 0.01875$$

所以，当给定一个水果的特征：long, sweet, yellow，那么这个水果更有可能是Banana。

$$p(Banana|\ Long, Sweet, Yellow) > p(Other|\ Long, Sweet, Yellow) > p(Orange|\ \ldots)$$

朴素贝叶斯模型的优点

1. **数据需求不高**：即使在数据量较少的情况下，朴素贝叶斯分类器也能估计出必要的参数。

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

必要的参数包括类别的先验概率和给定类别下特征的条件概率，这些在数据量较少的情况下，也同样可以获得。

2. **对缺失数据不敏感**：在朴素贝叶斯中，各个特征在给定类别的条件下是独立的，也就是说每个特征对最终的后验的贡献是单独计算的。如果某个或几个特征缺失，只需在计算联合概率时忽略这些特征，而不必影响其他特征的条件概率计算。

$$p(C_k \mid x_1, \ldots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$$

计算似然时，可以忽略缺失特征，只乘以那些已知的特征的条件概率 $p(x_i \mid C_k)$

朴素贝叶斯模型的缺点

1. **特征独立性假设**：朴素贝叶斯模型假设在给定类别条件下所有特征都是相互独立的，但实际上数据集的特征之间存在某种关联，只是这种关联有时候多，有时候少。在特征关联性强的情况下朴素贝叶斯的性能会下降。

2. **概率估计问题**：当某个类别下的某个特征未在训练集中出现过时，会导致该类别的概率估计为零，即零概率问题，这会影响到后验概率的计算和分类结果。

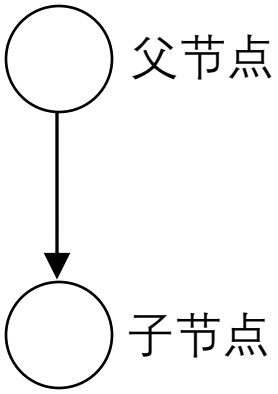
$$p(Orange|\ Long, Sweet, Yellow) = 0$$

这是因为Long这个特征并没有在Orange这个类别里出现过，也就是P(Long|Orange)=0，并最终导致后验概率为0。意味着即使其他证据强烈支持样本属于Orange类别，但特征值Long独自就能使该类别的后验概率归零，从而导致分类决策错误。

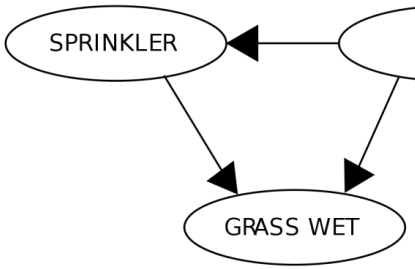
贝叶斯信念网络

贝叶斯信念网络

- 贝叶斯网络（贝叶斯信念网络），是一种表示变量之间概率依赖关系的图模型（graph model）。它由以下两个主要组成部分构成：
- **网络结构 G**：这是一个**有向无环图**（DAG），其中每个**节点（V）**代表一个随机变量。如果两个变量之间存在直接的概率依赖性，则它们之间会有一条有向边**（E）**相连。
- **参数Θ**：参数是指每个节点（V）都具有的一个**条件概率表（CPT）**，它量化了在给定父节点的特定值的情况下，该节点取特定值的概率。



RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

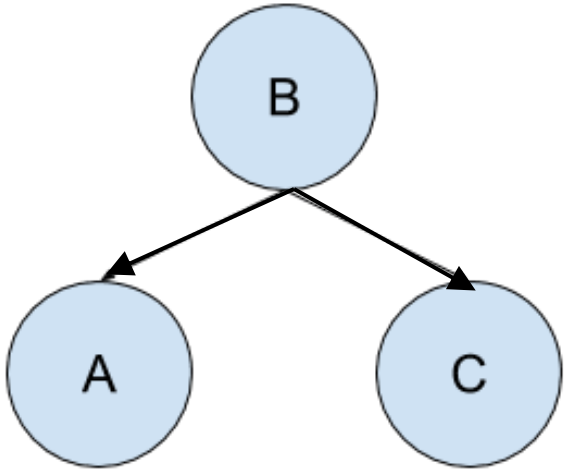


RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

条件独立属性

- 贝叶斯网络的核心是揭示了变量间的**条件独立属性。**条件独立指的是在给定某些条件的情况下，多个随机变量相互独立。（贝叶斯网络通过图结构明确地表示哪些变量在给定其他变量（通常是父节点）时是条件独立的）。



这个图所代表的含义是：A, C在B发生的前提下，相互独立，即条件独立。

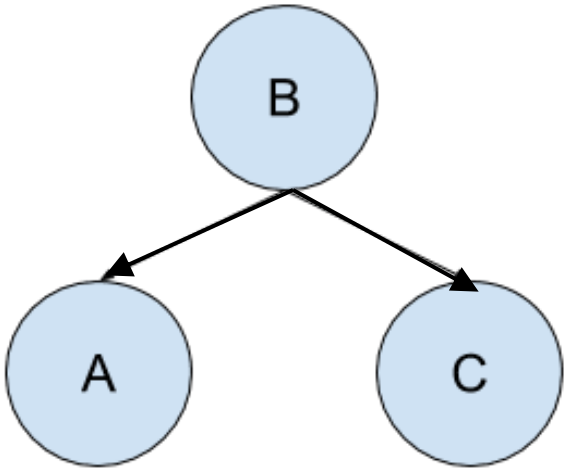
贝叶斯信念网络

- 贝叶斯网络核心应用是它能够**表示和处理多变量的联合概率分布**。它的应用具体包括以下几个方面:

  - **表示联合概率分布**
  - **条件概率查询和预测**
  - **推理和决策**

表示联合概率分布

A，B，C三个事件的联合概率分布是?



根据条件概率的定义:

$$P(A,B,C) = P(A,C \mid B)\,P(B)$$

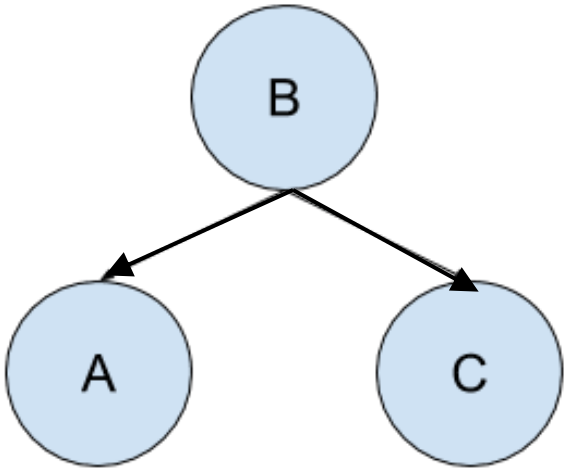
A, C在B发生的前提下，相互独立

$$P(A,C \mid B) = P(A \mid B)\,P(C \mid B)$$

$$P(A,B,C) = \underbrace{P(A \mid B)}_{\text{给定 } B \text{ 下 } A \text{ 的条件概率}} \times \underbrace{P(C \mid B)}_{\text{给定 } B \text{ 下 } C \text{ 的条件概率}} \times \underbrace{P(B)}_{\text{先验概率}}$$

表示联合概率分布

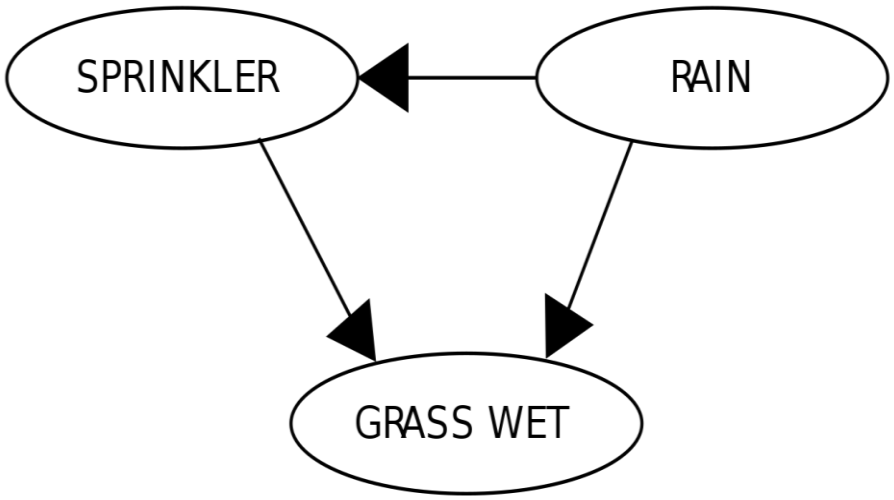
A，B，C三个事件的联合概率分布是?



- **B**没有父节点，所以它是一个先验概率P(B)
- **A**的父节点是B，因此A的条件概率是P(A|B)
- **C**的父节点是B：因此C的条件概率是P(C|B)

$$P(A,B,C) = \underbrace{P(A \mid B)}_{\text{给定 } B \text{ 下 } A \text{ 的条件概率}} \times \underbrace{P(C \mid B)}_{\text{给定 } B \text{ 下 } C \text{ 的条件概率}} \times \underbrace{P(B)}_{\text{先验概率}}$$

表示联合概率分布

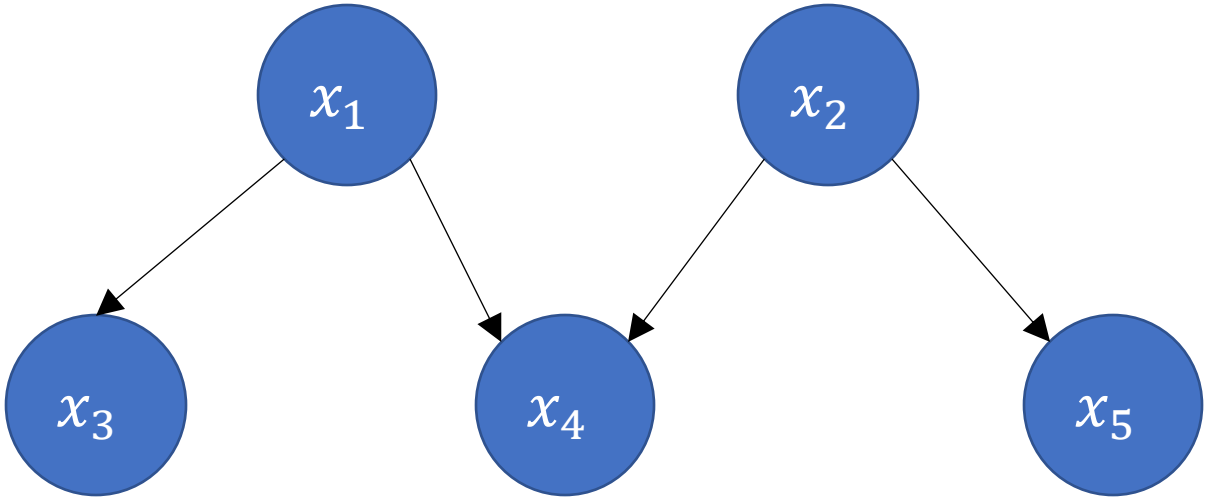


- **Rain** 没有父节点，所以它是一个先验概率P(Rain)。
- **Sprinkler** 的父节点是 Rain，因此Sprinkler的条件概率是P(Sprinkler|Rain)
- **GrassWet** 有两个父节点：Rain 和 Sprinkler。所以它的条件概率是P(GrassWet|Rain,Sprinkler)

将三个部分相乘，就得到了变量间的联合概率分布是:

$$\Pr(G, S, R) = \Pr(G \mid S, R) \Pr(S \mid R) \Pr(R)$$

表示联合概率分布

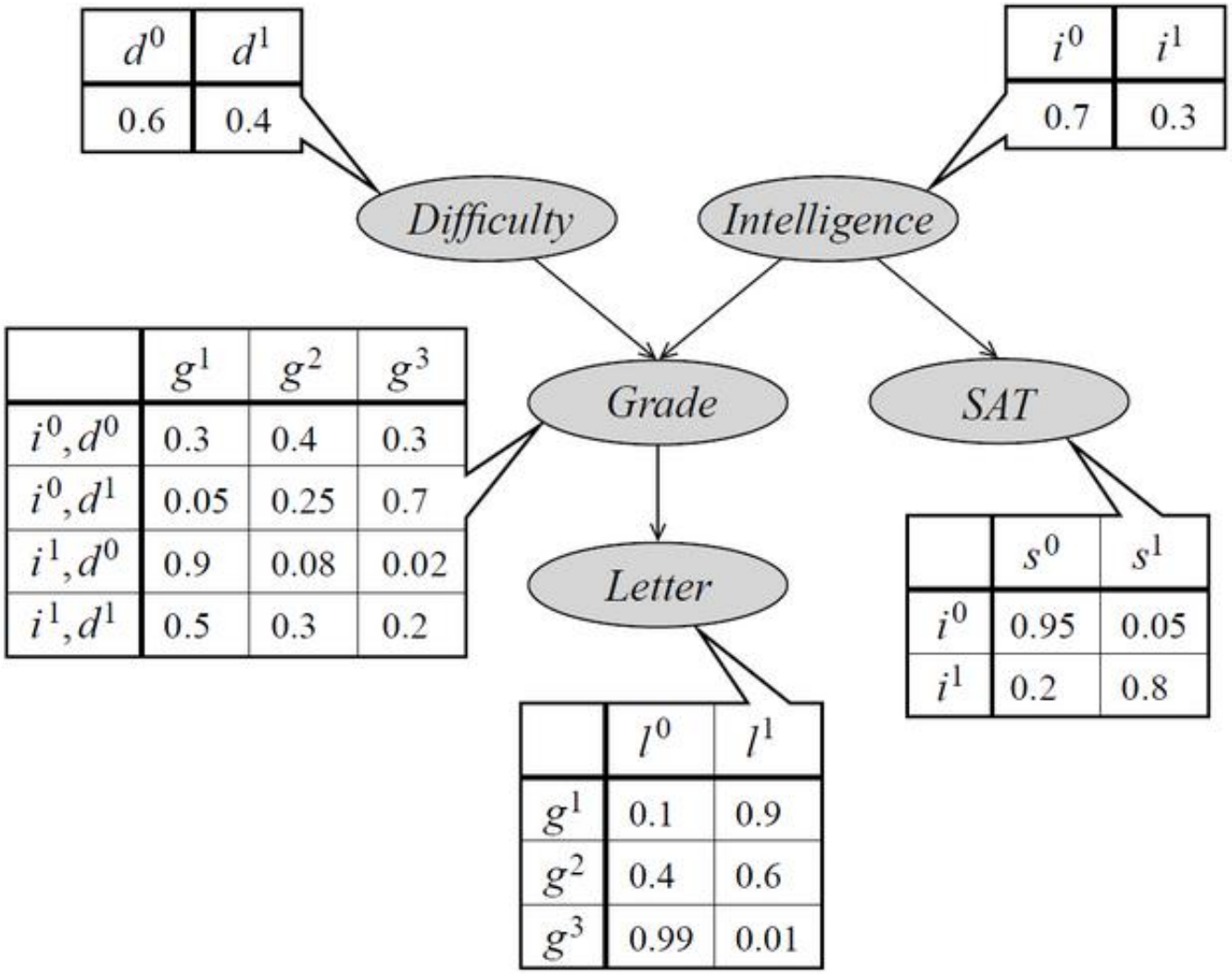


某个贝叶斯网络

它所表示的变量间的联合概率分布是:

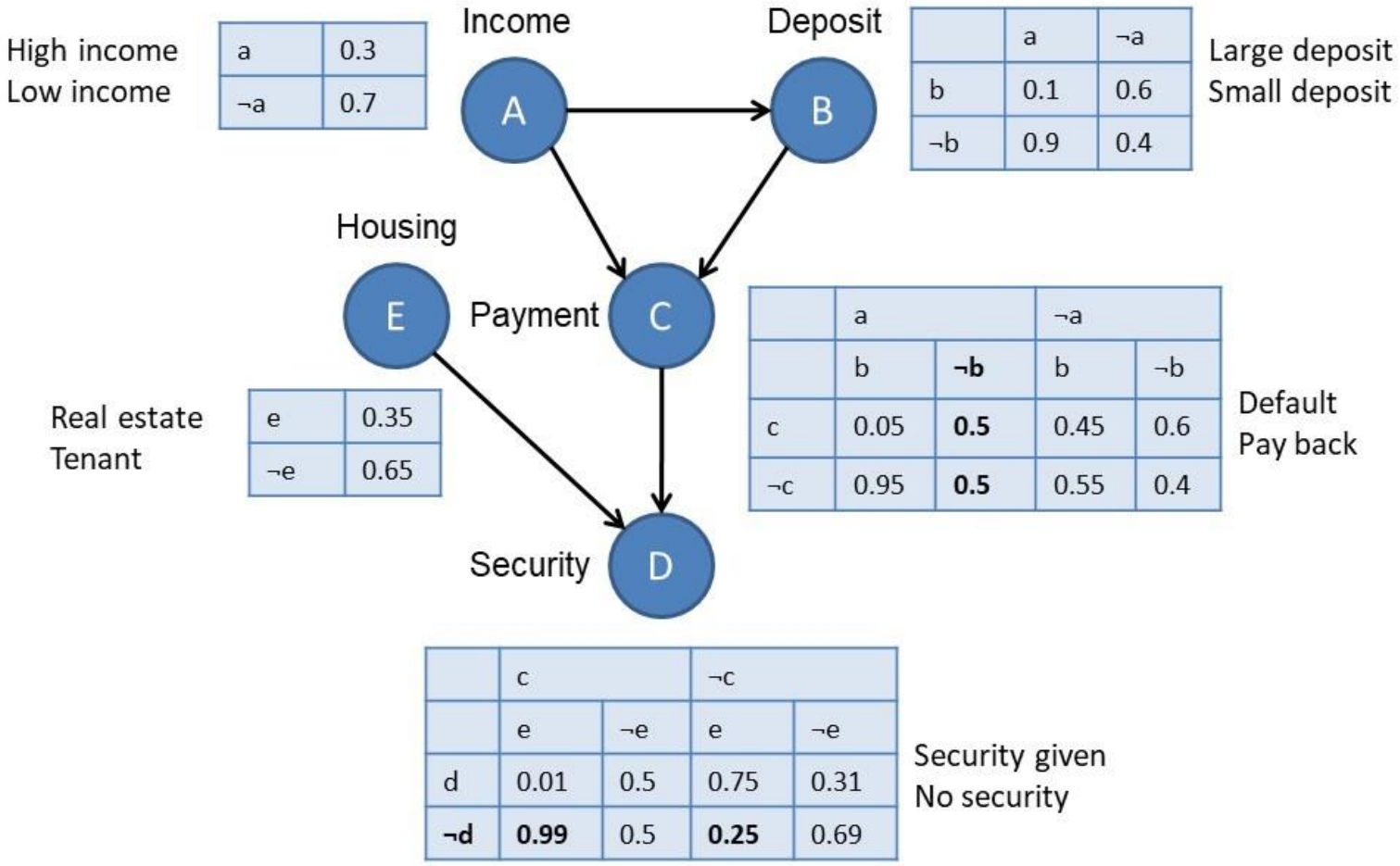
$$P(x_1, x_2, x_3, x_4, x_5) = P(x_1)P(x_2)P(x_3 \mid x_1)P(x_4 \mid x_1, x_2)P(x_5 \mid x_2)$$

贝叶斯信念网络



$P(I, D, G, S, L) = P(I)P(D)P(G \mid I, D)P(S \mid I)P(L \mid G)$.

贝叶斯信念网络



a	0.3
¬a	0.7

	a	¬a
b	0.1	0.6
¬b	0.9	0.4

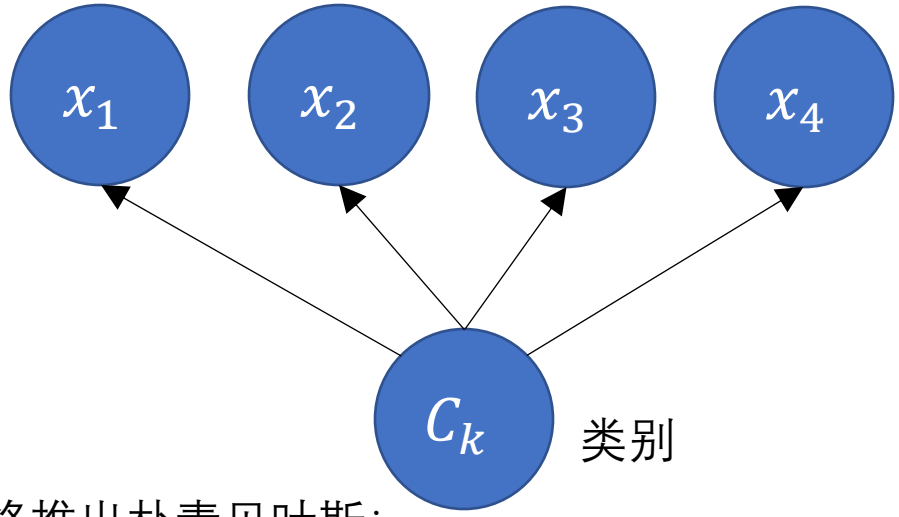
	a		¬a	
	b	**¬b**	b	¬b
c	0.05	**0.5**	0.45	0.6
¬c	0.95	**0.5**	0.55	0.4

	e	¬e
e	0.35	
¬e	0.65	

	c		¬c	
	e	¬e	e	¬e
d	0.01	0.5	0.75	0.31
¬d	**0.99**	0.5	**0.25**	0.69

贝叶斯信念网络/朴素贝叶斯

- 朴素贝叶斯是贝叶斯网络的一个特例
- 朴素贝叶斯如果用贝叶斯网络来表示



特征

（给定类别下，特征相互独立）

类别

通过贝叶斯网络推出朴素贝叶斯：

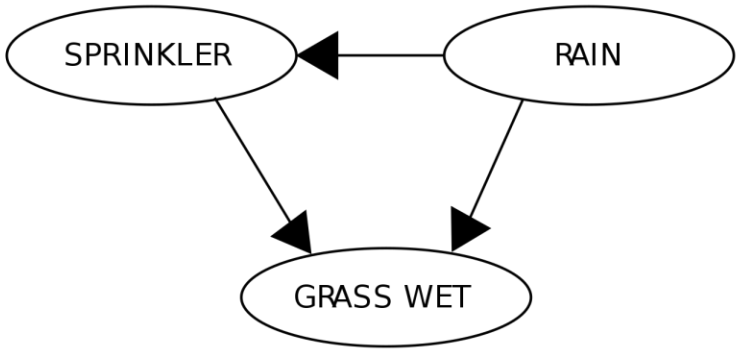
$$P(C_k|x_1,x_2,x_3,x_4)=\frac{P(C_k,x_1,x_2,x_3,x_4)}{P(x_1,x_2,x_3,x_4)}$$

$$=\frac{p(C_k)\times p(x_1|C_k)\times p(x_2|C_k)\times p(x_3|C_k)\times p(x_4|C_k)}{p(x_1,x_2,x_3,x_4)}$$

条件概率查询和预测

- 贝叶斯网络可以用来查询条件概率。例如，可以查询在给定某些证据变量的情况下其他变量的条件概率。

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

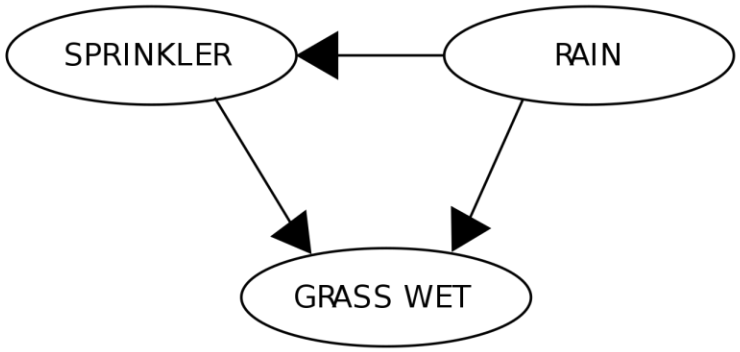
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

问：浇水器开启，下雨同时发生的情况下，草湿的概率是多少？

条件概率查询和预测

- 贝叶斯网络可以用来查询条件概率。例如，可以查询在给定某些证据变量的情况下其他变量的条件概率。

RAIN	SPRINKLER: T	SPRINKLER: F
F	0.4	0.6
T	0.01	0.99



RAIN: T	RAIN: F
0.2	0.8

P(G=T|S=T, R=T) = 0.99

SPRINKLER	RAIN	GRASS WET: T	GRASS WET: F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

问：浇水器开启，下雨同时发生的情况下，草湿的概率是多少？

条件概率查询和预测

- 草湿，浇水器开启，下雨这三者同时发生的概率是？

$$\begin{aligned}\Pr(G=T,S=T,R=T) &= \Pr(G=T \mid S=T,R=T)\Pr(S=T \mid R=T)\Pr(R=T)\\ &= 0.99 \times 0.01 \times 0.2\\ &= 0.00198.\end{aligned}$$

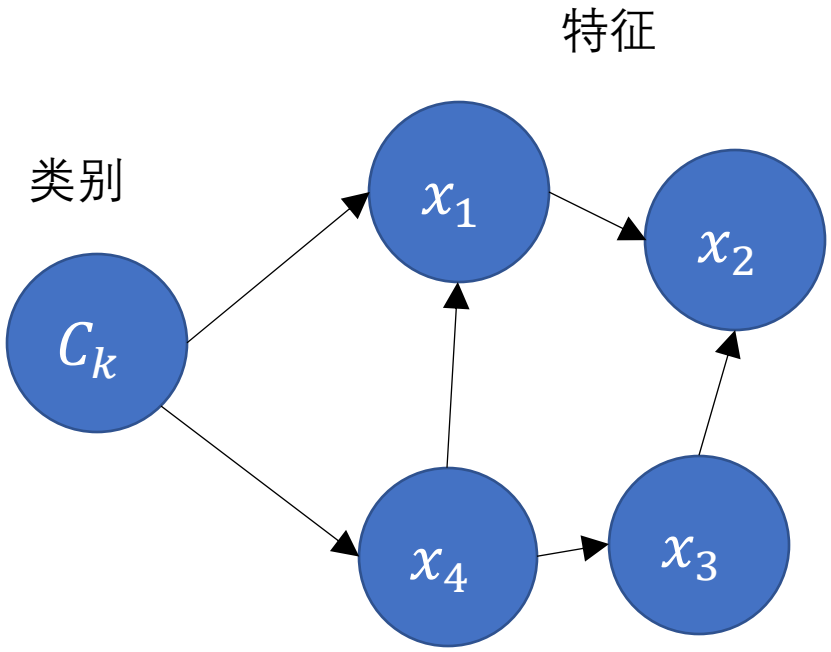
- 如果草湿了，那么下雨的概率是多少？

$$\Pr(R=T \mid G=T) = \frac{\Pr(G=T,R=T)}{\Pr(G=T)} = \frac{\sum_{x\in\{T,F\}}\Pr(G=T,S=x,R=T)}{\sum_{x,y\in\{T,F\}}\Pr(G=T,S=x,R=y)}$$

推理和决策

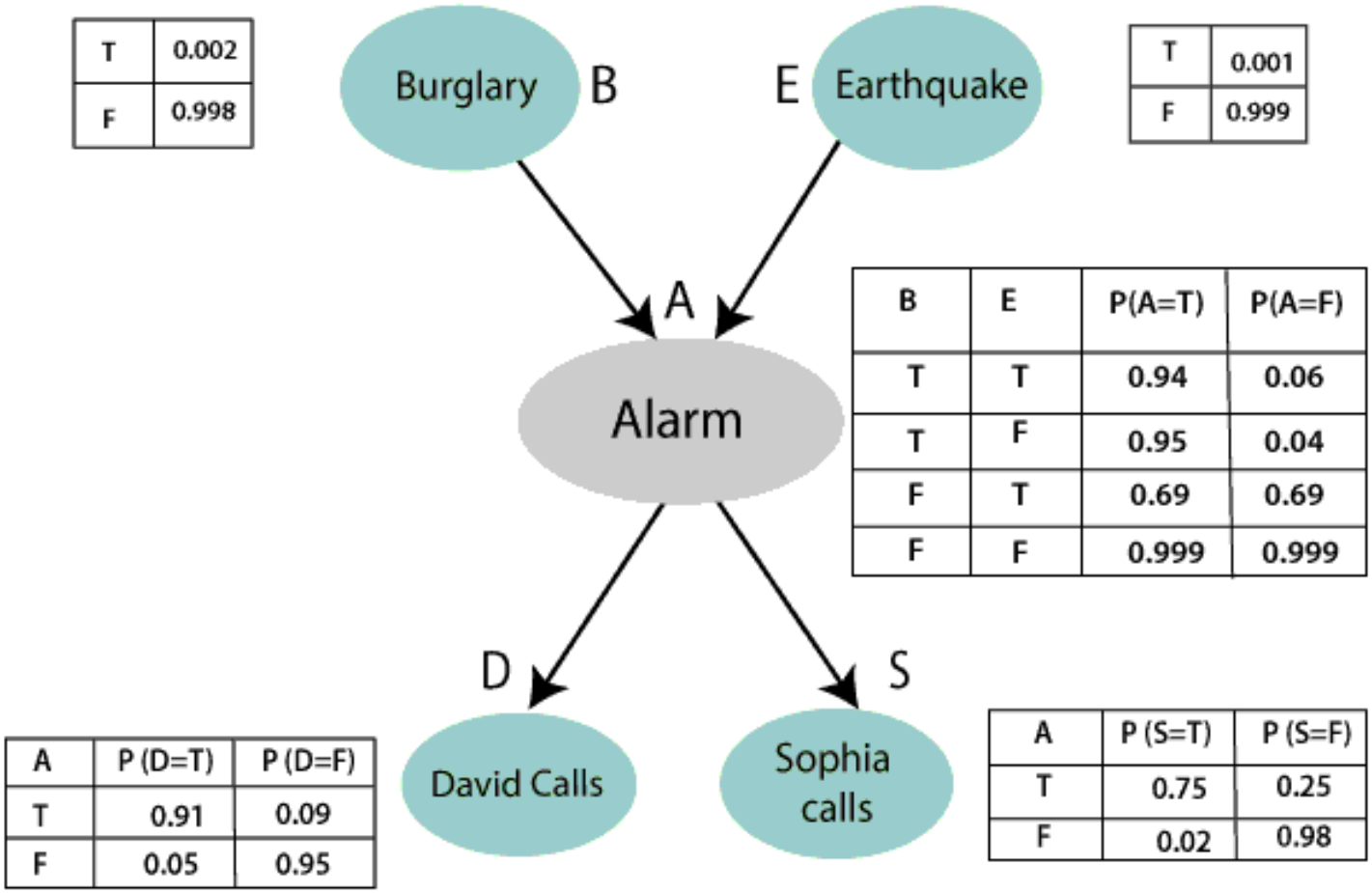
- 分类问题

将输入特征X和目标变量C_k 作为网络节点，建模节点间的依赖关系构建条件概率表，计算后验概率。



计算后验概率 $P(C_k|x_1,x_2,x_3,x_4)=\frac{P(C_k,x_1,x_2,x_3,x_4)}{P(x_1,x_2,x_3,x_4)}$

贝叶斯信念网络



T	0.002
F	0.998

T	0.001
F	0.999

B	E	P(A=T)	P(A=F)
T	T	0.94	0.06
T	F	0.95	0.04
F	T	0.69	0.69
F	F	0.999	0.999

A	P (D=T)	P (D=F)
T	0.91	0.09
F	0.05	0.95

A	P (S=T)	P (S=F)
T	0.75	0.25
F	0.02	0.98

P(B = True, E = False, A = True, D=True, S=False) =?