第二部分：概率模型

南京大学 人工智能学院
School of Artificial Intelligence, Nanjing University

章宗长

2021年3月10日

内容安排

- 不确定性的表示
- 概率推理
- 参数学习
- 结构学习
- 应用案例：视频监控

不确定性的表示

- **信念度和概率**
- 概率分布
- 联合概率分布
- 条件概率分布
- 贝叶斯网络
- 时序模型

信念度

- 假设绕地遥感卫星突然与地面失去了联系，可能原因：
  - 卫星上的电力系统或通讯系统发生了故障
  - 监控卫星的地面系统发生了故障
- 假如：相比“卫星上有推进器异常”，我们更相信“卫星上有电子异常”
  - 命题E：卫星上有电子异常
  - 命题T：卫星上有推进器异常　　　那么有：$E \succ T$
- 如果认为E和T有相同的可信程度，那么$E \sim T$
- 假如：在出现了通讯故障的情况下，相比“卫星上有推进器异常”，我们更相信“卫星上有电子异常”
  - 命题C：出现了通讯故障　　　那么有：$(E \mid C) \succ (T \mid C)$

信念度

- 普适的可比性假设
  - $(A \mid C) \succ (B \mid C)$，$(A \mid C) \sim (B \mid C)$，$(A \mid C) \prec (B \mid C)$中必有一个成立
- 传递性假设
  - 如果$(A \mid D) \succcurlyeq (B \mid D)$，$(B \mid D) \succcurlyeq (C \mid D)$，那么$(A \mid D) \succcurlyeq (C \mid D)$
- 基于这两条假设，可以用实值函数P来表示**信念度**：
  - $P(A \mid C) > P(B \mid C)$当且仅当$(A \mid C) \succ (B \mid C)$
  - $P(A \mid C) = P(B \mid C)$当且仅当$(A \mid C) \sim (B \mid C)$

信念度和概率

- （**概率论公理**）假设某个试验的样本空间为Ω，对应于其中任一事件E，定义一个数P(E)。如果P(E)满足如下3个公理，则称P(E)是事件E的概率：
  - **公理1** $0 \leq \mathrm{P}(E) \leq 1$
  - **公理2** $\mathrm{P}(\Omega) = 1$
  - **公理3** 对任一列互不相容的事件$E_1, E_2, \ldots$（即如果$i \neq j$，则$E_iE_j = \emptyset$），有

$$P\left(\bigcup_{i=0}^{\infty} E_i\right) = \sum_{i=0}^{\infty} P(E_i)$$

- 假设P满足概率论公理，即$0 \leq \mathrm{P}(\mathrm{A} \mid \mathrm{B}) \leq 1$
  - 如果(A | B)是确定的，那么$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = 1$
  - 如果(A | B)是不可能的，那么$\mathrm{P}(\mathrm{A} \mid \mathrm{B}) = 0$
  - 如果(A | B)介于两者之间，那么$0 < \mathrm{P}(\mathrm{A} \mid \mathrm{B}) < 1$

条件概率

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

全概率法则

$$P(A \mid C) = \sum_{B \in \mathfrak{B}} P(A \mid B, C)\, P(B \mid C)$$

$\mathfrak{B}$：由互斥且可穷举的命题构成的集合

贝叶斯规则

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

应用贝叶斯规则：简单实例

- 将未知因素cause造成的结果effect看作证据，确定未知因素cause：

$$P(cause \mid effect) = \frac{P(effect \mid cause)\,P(cause)}{P(effect)}$$

- 例子
  - 令cause表示命题“病人患有某种传染性肺炎”
  - 令effect表示命题“病人核酸检测为阳性”
  - $P(effect \mid cause) = 1.0$
  - $P(cause) = 1/5000$
  - $P(effect) = 0.001$
  - $P(cause \mid effect) = \frac{P(effect \mid cause)\,P(cause)}{P(effect)} = \frac{1.0 \times \frac{1}{5000}}{0.001} = 0.2$

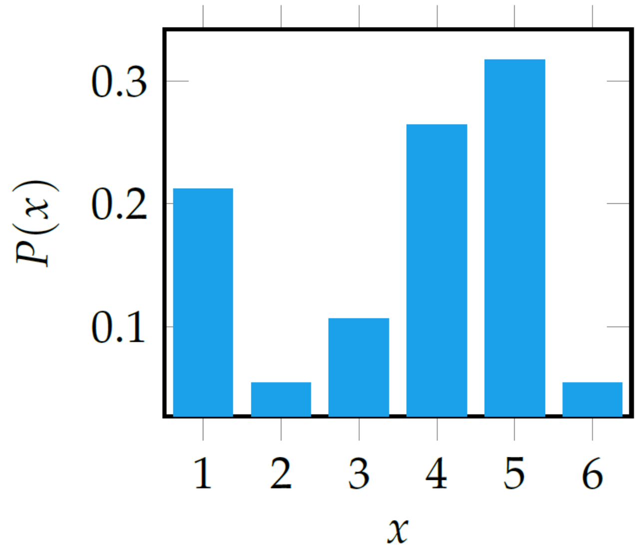
不确定性的表示

- 信念度和概率
- **概率分布**
- 联合概率分布
- 条件概率分布
- 贝叶斯网络
- 时序模型

离散概率分布

- 离散概率分布
  - 在一个离散值集合上的分布
  - 表示：概率质量函数

示例：概率质量函数

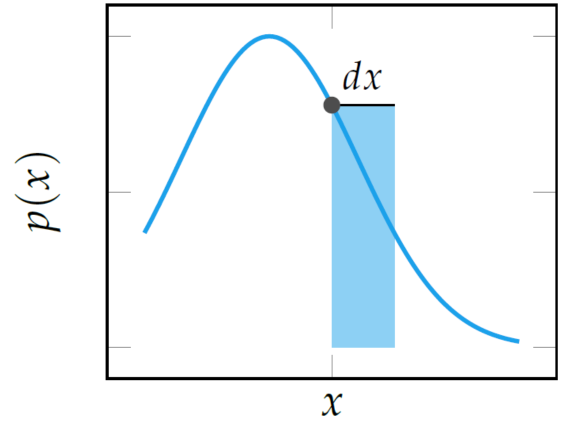


- 二值随机变量X：$\{0,1\}$
  - $P(x^0) = P(X = 0)$: X取值为0的概率
  - $P(x^1) = P(X = 1)$: X取值为1的概率

- n值随机变量X：$\{1,2,\ldots,n\}$
  - $P(x^1), P(x^2), \ldots, P(x^n)$
  - $P(x^n) = 1 - (P(x^1) + P(x^2) + \ldots + P(x^{n-1}))$

连续概率分布

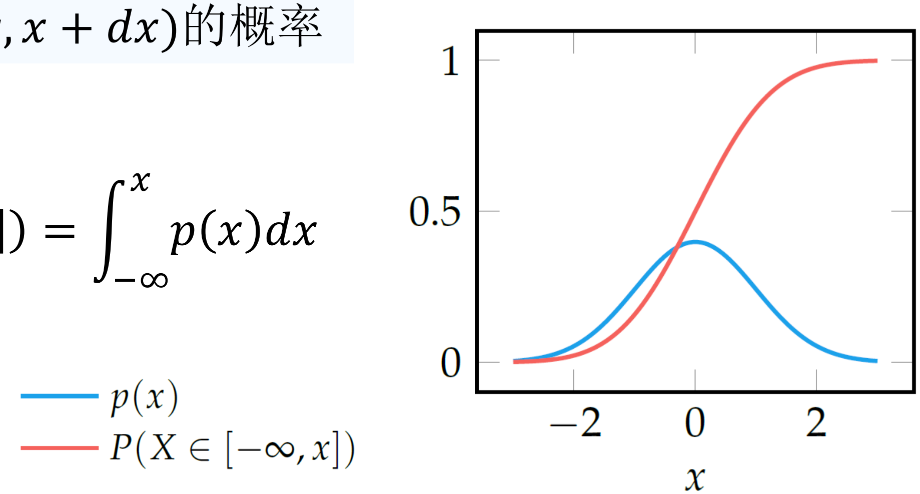
示例：概率密度函数



- 连续概率分布
  - 在一个连续值集合上的分布
  - 表示：概率密度函数
- 一个概率密度函数$p(x)$满足：$\int_{-\infty}^{+\infty} p(x)dx = 1$

随着$dx \to 0$，X的取值落入区间$(x, x+dx)$的概率

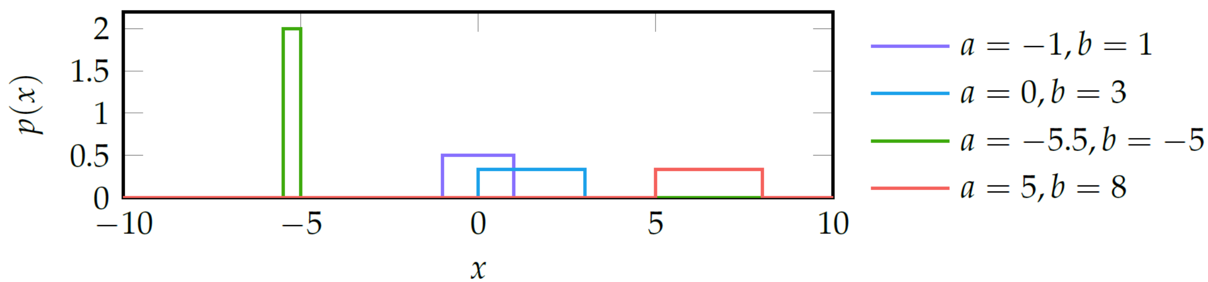
- 累积分布函数 $P(X \in [-\infty, x]) = \int_{-\infty}^{x} p(x)dx$



均匀分布

- 均匀分布$U(a,b)$的概率密度函数：

$$U(x \mid a,b)=\begin{cases}\dfrac{1}{b-a} & \text{如果} a\le x\le b\\ 0 & \text{否则}\end{cases}$$



- 例子：X服从均匀分布$U(0,10)$，那么X的取值落入区间[3, 4]内的概率：

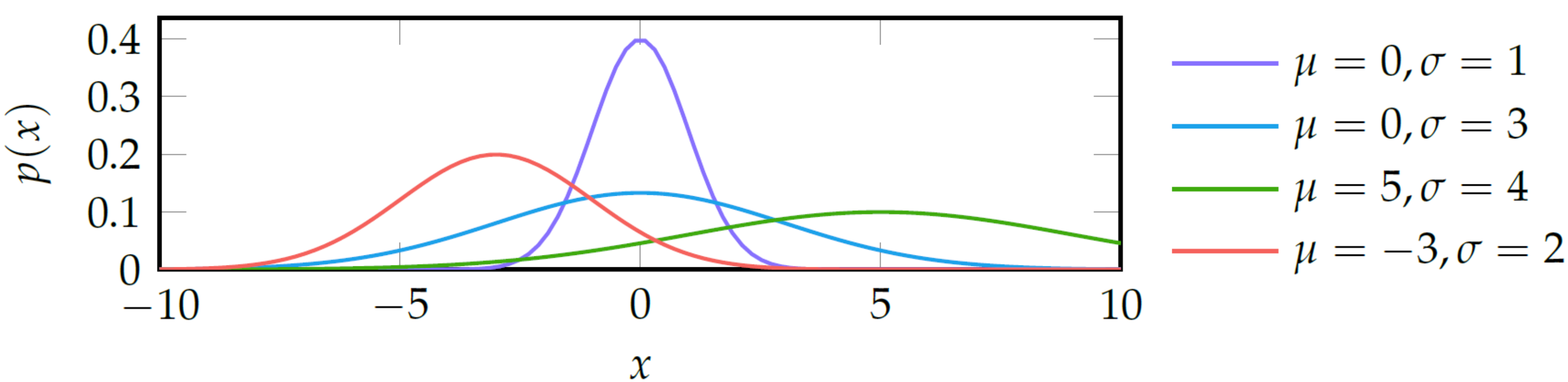
$$\int_3^4 U(x \mid 0,10)dx=\frac{1}{10}$$

高斯分布

- 高斯分布（正态分布）：连续随机变量的常见分布形式
- 随机变量服从均值为μ，方差为σ^2的高斯分布$\mathcal{N}(w \mid \mu, \sigma^2)$，其密度函数

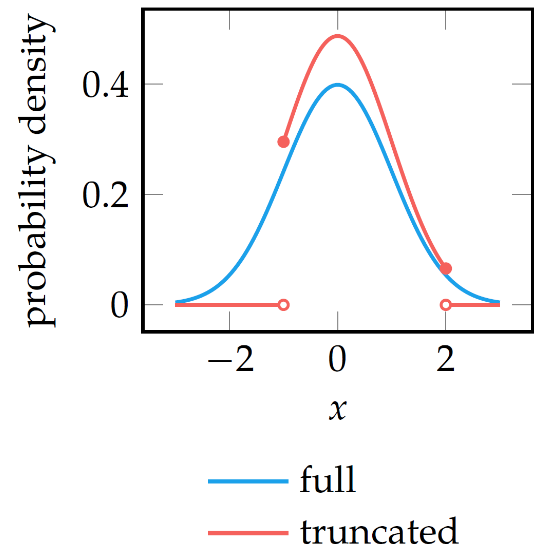
$$p(\mathrm{w}) = \mathcal{N}(w \mid \mu, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{w-\mu}{\sigma}\right)$$

$\phi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$：标准的正态密度函数



截断式高斯分布

- 假设用概率密度函数表示飞机在纽约肯尼迪机场的高度分布情况，飞机的高度范围[0, 65000]英尺(ft)



- 截断式高斯分布的密度函数

$$\mathcal{N}(w \mid \mu, \sigma^2, a, b) = \frac{\frac{1}{\sigma}\phi(\frac{w-\mu}{\sigma})}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

w：取值范围$[a, b]$

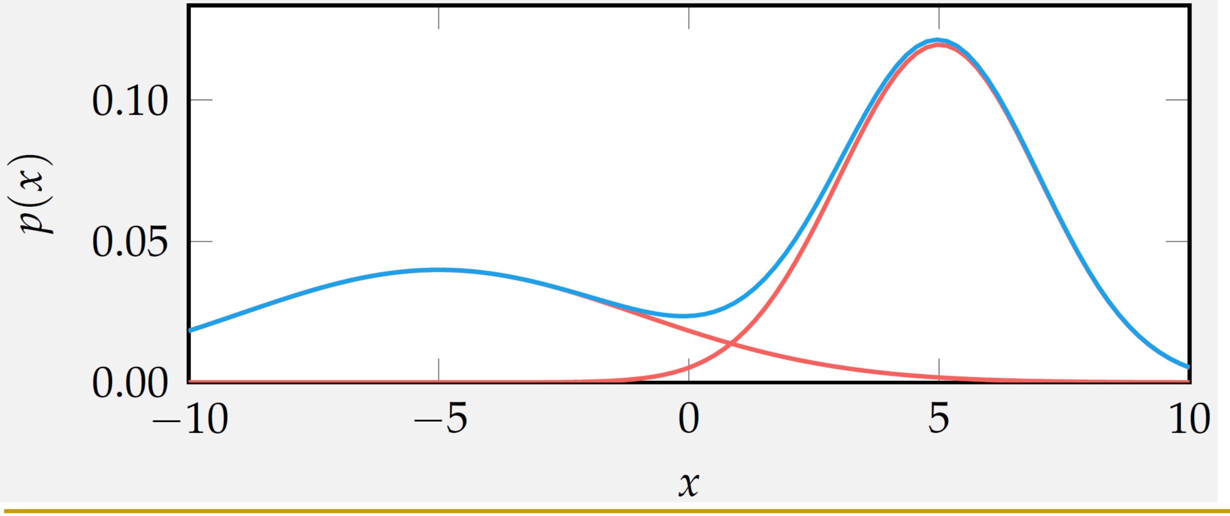
$\Phi(x) = \int_{-\infty}^{x} \phi(x)dx$：标准的正态累积分布函数

多模态的连续概率分布

- 高斯混合模型
  - 不同高斯分布的加权平均
  - 密度函数

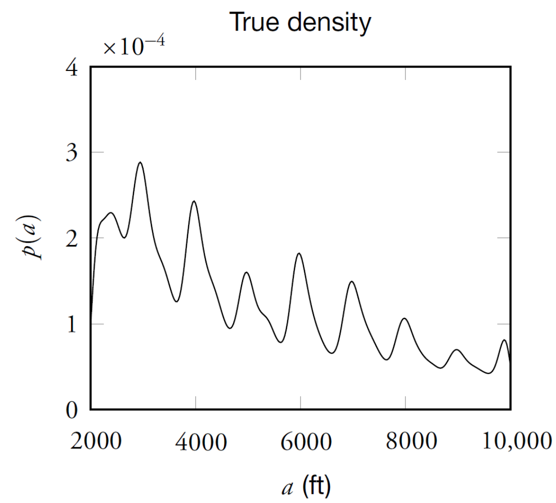
$$p(\mathrm{w} \mid \mu_1, \sigma_1^2, \ldots, \mu_n, \sigma_n^2, \rho_1, \ldots, \rho_n) = \sum_{i=1}^{n} \rho_i \ \mathcal{N}(w \mid \mu_i, \sigma_i^2)$$

其中$\sum_{i=1}^{n} \rho_i = 1$



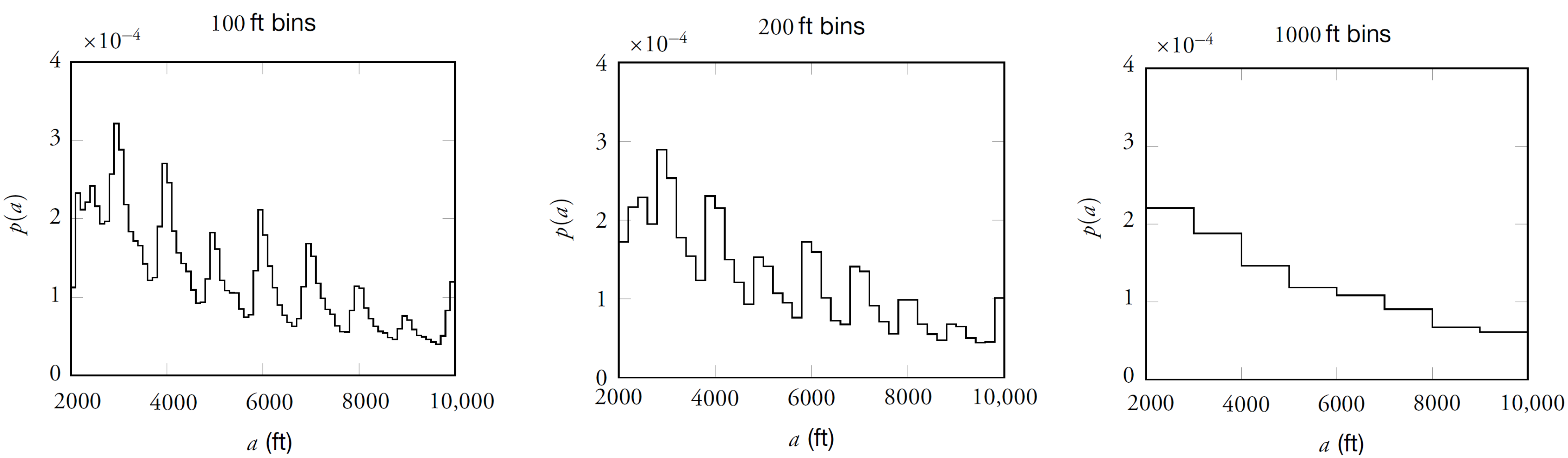
示例：高斯混合模型
$\mu_1 = 5, \sigma_1 = 2$
$\mu_1 = -5, \sigma_1 = 4$
$\rho_1 = 0.6, \rho_1 = 0.4$

多模态的连续概率分布（续）



- 右图：飞机在纽约肯尼迪机场的高度范围[2000ft, 10000ft]的概率分布

- 分段均匀密度函数



不确定性的表示

- 信念度和概率
- 概率分布
- **联合概率分布**
- 条件概率分布
- 贝叶斯网络
- 时序模型

离散联合分布

示例：联合分布的表格表示

X	Y	Z	$P(X,Y,Z)$
0	0	0	0.08
0	0	1	0.31
0	1	0	0.09
0	1	1	0.37
1	0	0	0.01
1	0	1	0.05
1	1	0	0.02
1	1	1	0.07

- 联合分布
  - 多个随机变量的概率分布
- 建模二值变量X, Y, Z的联合分布
  - 对这3个变量有8种不同可能的赋值，每种赋值对应一个概率，概率之和为1
  - θ_{i}：表中第i行出现的概率
  - $\theta_8 = 1 - (\theta_1 + \theta_2 + \cdots + \theta_7)$
- 如果有n个二值变量，则有$2^{\mathrm{n}} - 1$个独立的参数
- 独立的参数个数的指数增长，带来了表示不确定性和学习概率模型的困难

独立性假设

- 变量X和Y是独立的，当且仅当$P(X,Y) = P(X)\,P(Y)$，记为$X \perp Y$
- $X \perp Y$当且仅当$P(X) = P(X \mid Y)$
- 如果有n个独立的二值变量，那么能仅用n个独立的参数得到联合分布，而不是$2^{n} - 1$个

X	Y	Z	P(X,Y,Z)
0	0	0	0.08
0	0	1	0.31
0	1	0	0.09
0	1	1	0.37
1	0	0	0.01
1	0	1	0.05
1	1	0	0.02
1	1	1	0.07

如果知道X,Y,Z是相互独立的

则独立参数的个数减少至$n = 3$

X	P(X)
0	0.85
1	0.15

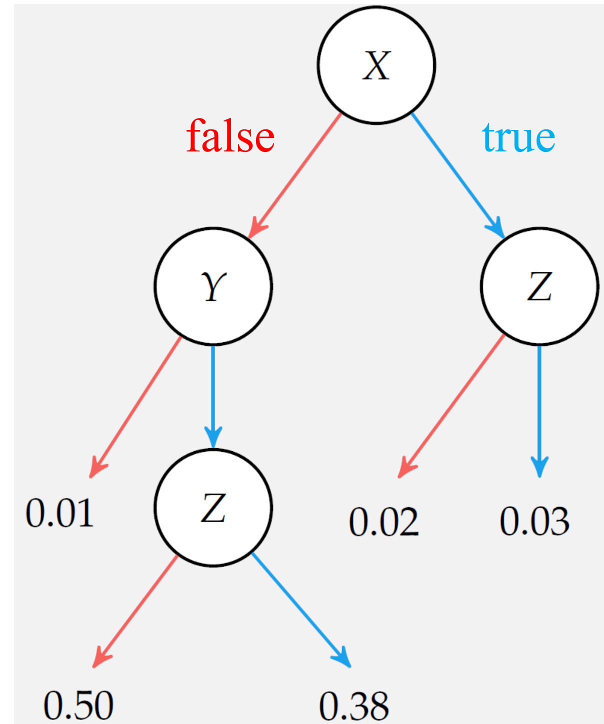
Y	P(Z)
0	0.45
1	0.55

Z	P(Z)
0	0.20
1	0.80

决策树表示

- 决策树：更紧凑地表示表格中的值

X	Y	Z	$P(X,Y,Z)$
0	0	0	0.01
0	0	1	0.01
0	1	0	0.50
0	1	1	0.38
1	0	0	0.02
1	0	1	0.03
1	1	0	0.02
1	1	1	0.03



存储8个概率值

仅存储5个概率值

多元均匀分布及其混合模型

- 多元均匀分布$U(\mathbf{a},\mathbf{b})$的概率密度函数：

2n个独立的参数

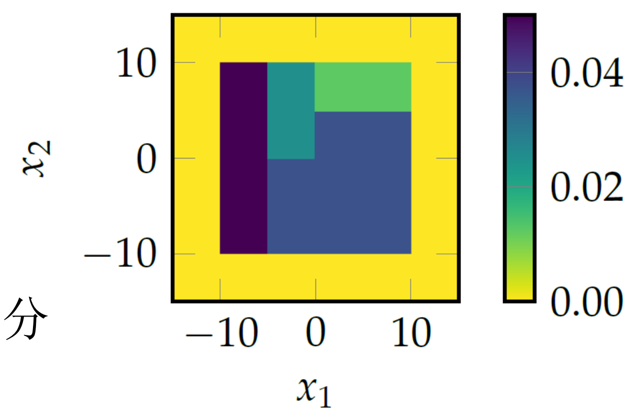
$$U(\mathbf{x}\mid\mathbf{a},\mathbf{b})=\prod_{i=1}^{n}U(x_i\mid a_i,b_i)$$

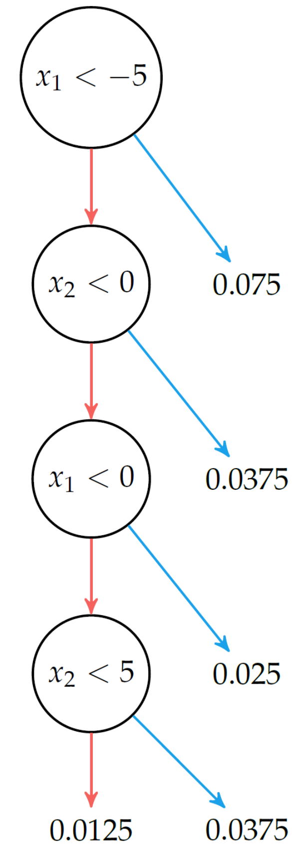
- 多元均匀分布的混合模型：

$$p(\mathbf{x}\mid\mathbf{a}_1,\mathbf{b}_1,\dots,\mathbf{a}_k,\mathbf{b}_k,\rho_1,\dots,\rho_k)=\sum_{j=1}^{k}\rho_j\,U(\mathbf{x}\mid\mathbf{a}_j,\mathbf{b}_j)$$

其中$\sum_{j=1}^{k}\rho_j=1$

$k\times 2n+(k-1)$个独立的参数





示例：$n=2,k=5$的多元均匀分布的混合模型及其决策树表示

多元高斯分布

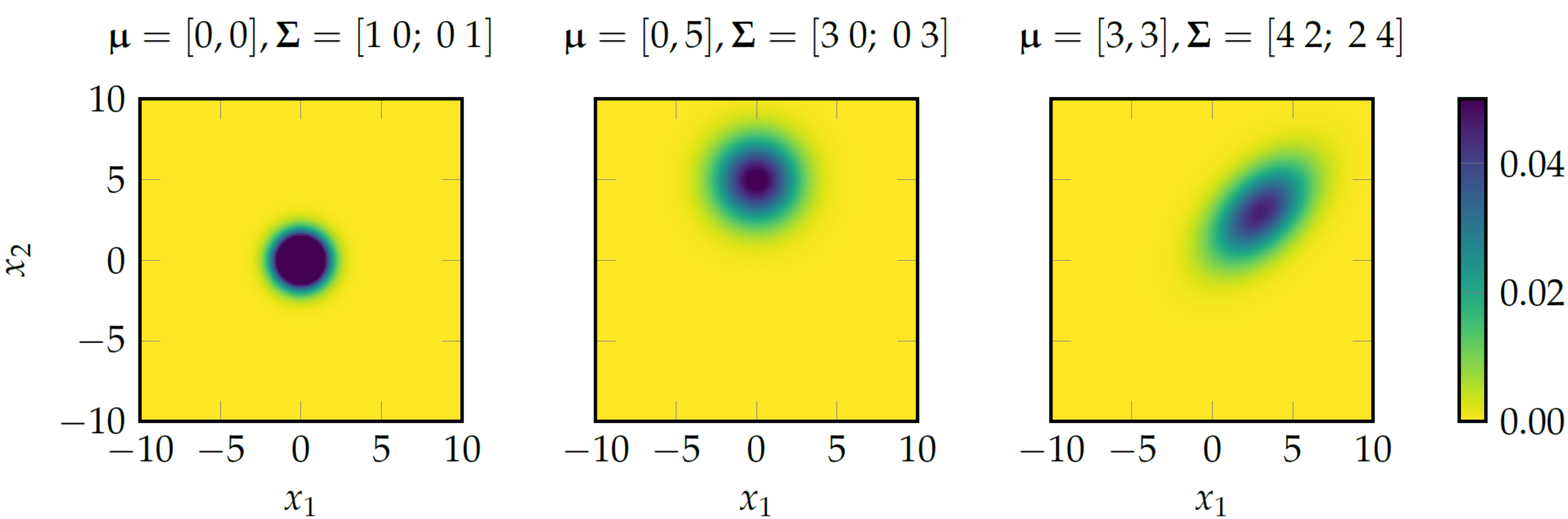
- 多元高斯分布$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$的概率密度函数：

$$\mathcal{N}(\mathbf{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

其中$\mathbf{x}\in\mathbb{R}^n$，$\boldsymbol{\mu}$是均值向量，$\boldsymbol{\Sigma}$是协方差矩阵

- 示例：

因为$\boldsymbol{\Sigma}$是对称矩阵，所以独立参数的个数为$n+(n+1)n/2$



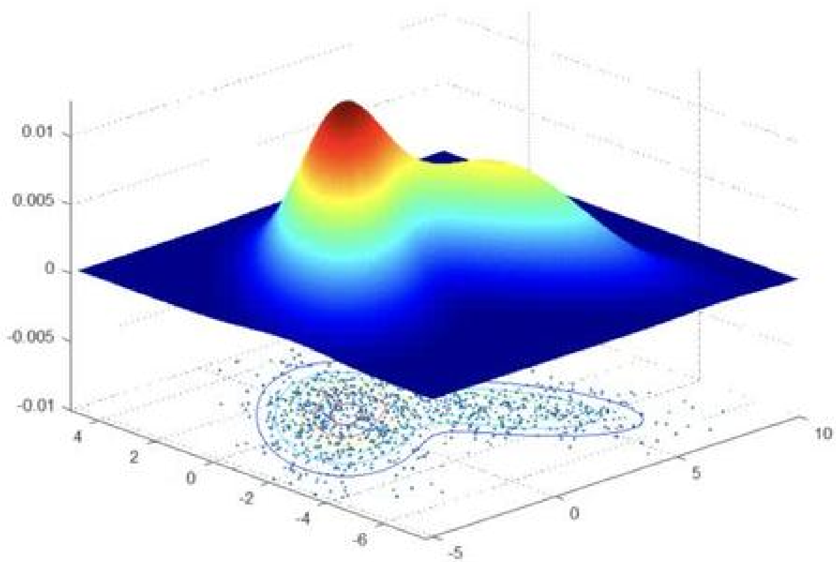
多元高斯混合模型

- 多元高斯分布的混合模型：

$$p(\mathbf{x}\,|\,\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\dots,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k,\rho_1,\dots,\rho_k)=\sum_{j=1}^{k}\rho_j\,\mathcal{N}\left(\mathbf{x}\,|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)$$

其中$\sum_{j=1}^{k}\rho_j=1$

- 示例：



不确定性的表示

- 信念度和概率
- 概率分布
- 联合概率分布
- **条件概率分布**
- 贝叶斯网络
- 时序模型

离散条件模型

X	Y	Z	$P(X \mid Y, Z)$
0	0	0	0.08
0	0	1	0.15
0	1	0	0.05
0	1	1	0.10
1	0	0	0.92
1	0	1	0.85
1	1	0	0.95
1	1	1	0.90

- 离散变量的条件概率分布
  - 表格表示
  - 给定证据y和z，所有x的概率之和为1

$$P(x^0 \mid y^0, z^0) + P(x^1 \mid y^0, z^0) = 0.08 + 0.92 = 1$$

如果每个随机变量由二值变量变为有m个离散值的变量，证据变量的个数由2个变为n个，那么表格有多少行？

m^{n+1}

独立变量的个数是多少呢？

$(m-1)m^n$

条件高斯模型

- 给定一个或多个离散变量，一个连续变量的概率分布

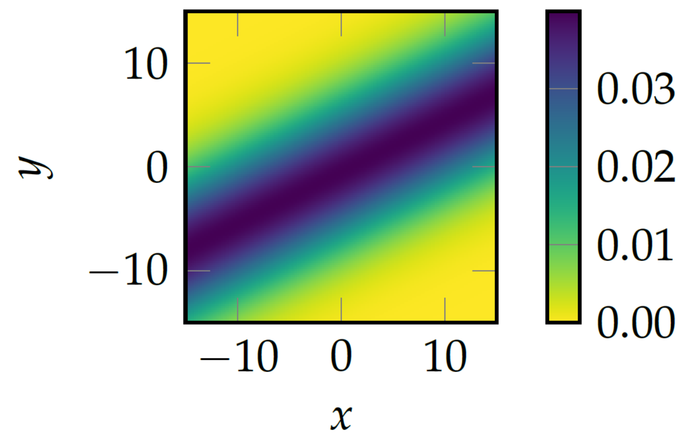
$$\mathrm{p}(x \mid y)=\begin{cases}\mathcal{N}(x \mid \mu_1,\sigma_1^2) & \text{如果} y^1 \\ \quad\quad\quad\cdots & \\ \mathcal{N}(x \mid \mu_n,\sigma_n^2) & \text{如果} y^n\end{cases}$$

$2n$个独立的参数

线性高斯模型

- 线性高斯模型：$P(X \mid Y)$
  - 连续随机变量X的高斯分布，均值为连续随机变量Y取值的线性函数
  - 条件概率密度函数：

$$p(x \mid y) = \mathcal{N}(x \mid my + b, \sigma^2)$$



$$p(x \mid y) = \mathcal{N}(x \mid 2y + 1, 10^2)$$

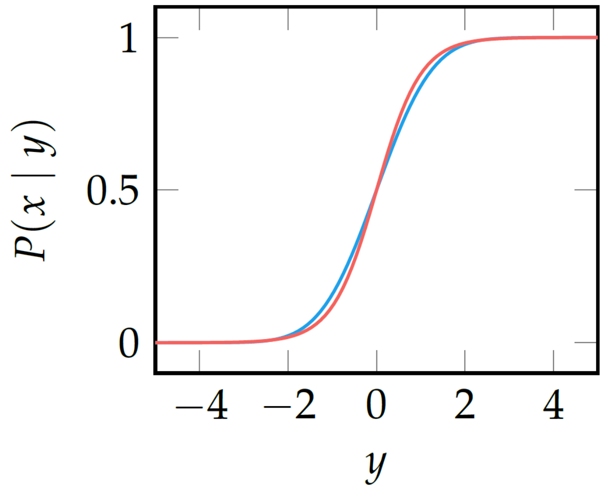
条件线性高斯模型

- 条件线性高斯模型：$P(X \mid Y, Z)$
  - 结合了条件高斯模型和线性高斯模型
  - X和Y为连续随机变量， Z为取值$1{:}n$的离散随机变量
  - 条件概率密度函数：

$$\mathrm{p}(x \mid y, z) = \begin{cases} \mathcal{N}(x \mid m_1 y + b_1, \sigma_1^2) & \text{如果} z^1 \\ \dots & \\ \mathcal{N}(x \mid m_n y + b_n, \sigma_n^2) & \text{如果} z^n \end{cases}$$

Sigmoid模型

- Sigmoid模型：给定一个连续变量，一个二值变量的概率分布
  - logit模型：$P(x^1\mid y)=\dfrac{1}{1+\exp\left(-2\frac{y-\theta_1}{\theta_2}\right)}$
  - probit模型：$P(x^1\mid y)=\Phi(y-\theta_1)/\theta_2$



红色：logit模型
蓝色：probit模型

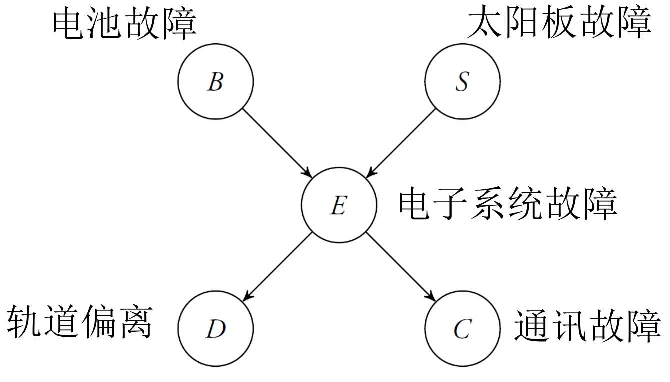
$\theta_1=0$，$\theta_2=1$

不确定性的表示

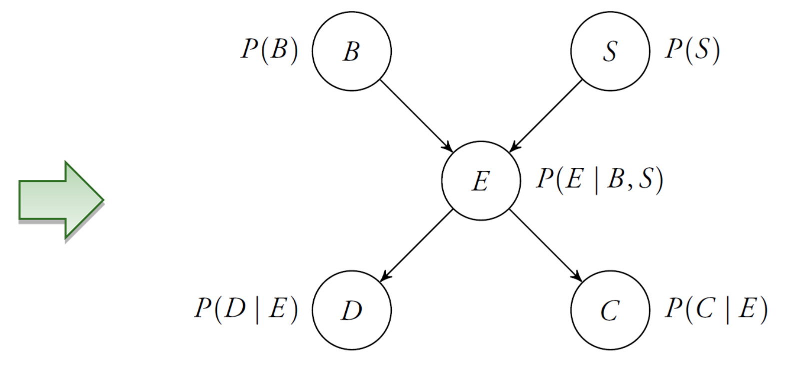
- 信念度和概率
- 概率分布
- 联合概率分布
- 条件概率分布
- **贝叶斯网络**
- 时序模型

贝叶斯网络

- 贝叶斯网络是联合分布的一种紧凑表示
- 网络结构：有向无环图
  - 结点：随机变量
  - 有向边（箭头）：联接结点对



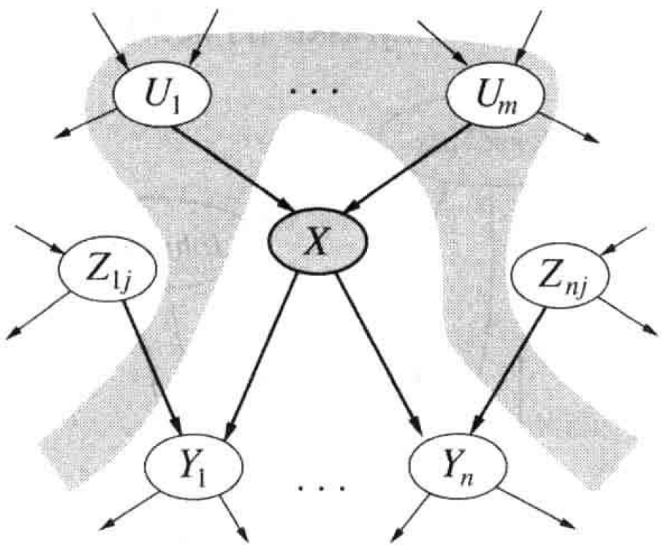
卫星监控问题的贝叶斯网络



网络的条件概率

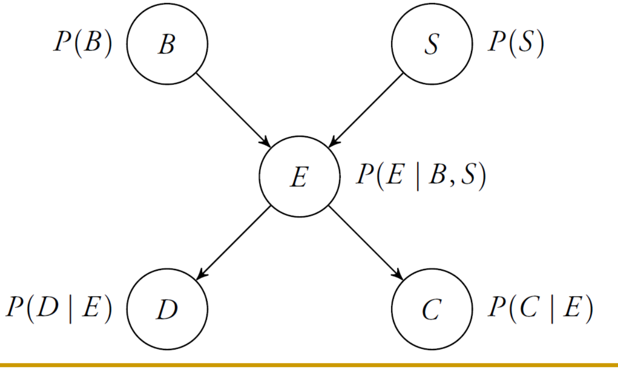
贝叶斯网络中的条件独立关系

- 给定Z，变量X和Y是独立的，当且仅当$P(X,Y \mid Z) = P(X \mid Z)\,P(Y \mid Z)$，记为$(X \perp Y \mid Z)$
  - $(X \perp Y \mid Z)$当且仅当$P(X \mid Z) = P(X \mid Y,Z)$
- 给定父结点（灰色区域中所示的各U_i），结点X条件独立于它的非后代结点（即各Z_{ij}）



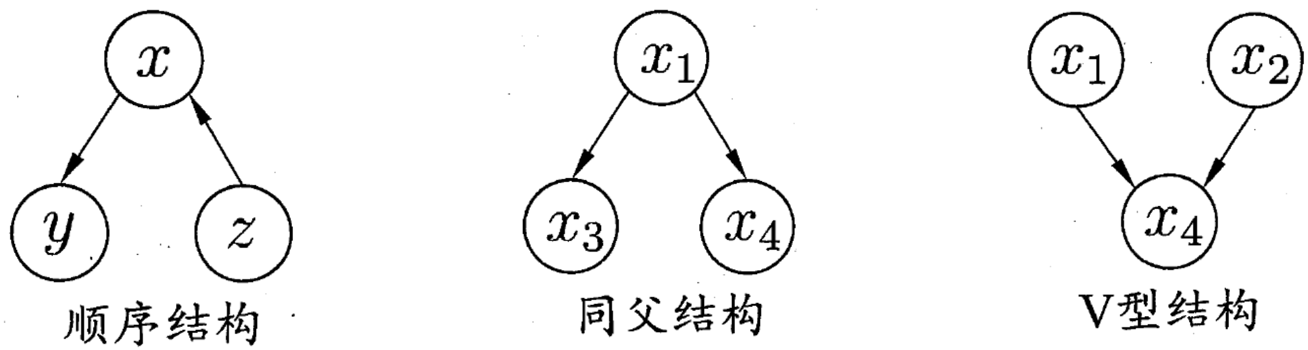
链式法则

- $P(X_i \mid Pa_{X_i})$：结点X_i的条件概率分布
  - Pa_{X_i}表示X_i的父结点
- 链式法则： $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid Pa_{x_i})$
- 例子
  - $P(b^0, s^0, e^0, d^0, c^0) = P(b^0)P(s^0)P(e^0 \mid b^0, s^0)\, P(d^0 \mid e^0)P(c^0 \mid e^0)$
  - 贝叶斯网络仅用1+1+4+2+2=10个独立的参数来计算这一联合分布



贝叶斯网络的结构

- 贝叶斯网络中三个变量的典型依赖关系：



- 在顺序结构中，给定x的值，则y和z条件独立
- 在同父结构中，给定x_1的值，则x_3和x_4条件独立
- 在V型结构中，给定x_4的值，则x_1和x_2必不独立
- 若x_4的取值完全未知，则V型结构下x_1和x_2却是相互独立的
  - 简单的验证：

$$p(x_1, x_2) = \sum_{x_4} p(x_1, x_2, x_4) = \sum_{x_4} p(x_4 \mid x_1, x_2) p(x_1) p(x_2) = p(x_1) p(x_2)$$

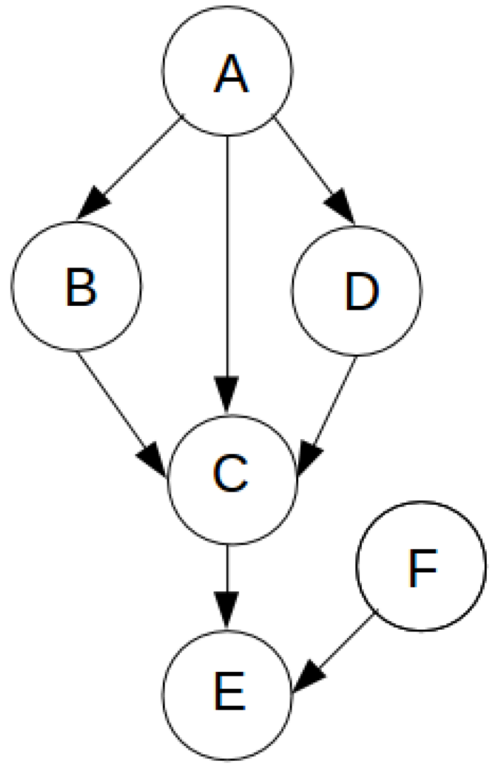
有向分离（d-分离，d-separation）

- 给定一个结点集合$\mathfrak{S}$，如果A和B的所有无向路径均由$\mathfrak{S}$有向分离， 记为$(A \perp B \mid \mathfrak{S})$，则称$\mathfrak{S}$有向分离A和B
- 如果A和B的一条路径满足如下任一情况：
  - 包含一个顺序结构：$X \longrightarrow Y \longrightarrow Z$，其中$Y \in \mathfrak{S}$
  - 包含一个同父结构：$X \longleftarrow Y \longrightarrow Z$，其中$Y \in \mathfrak{S}$
  - 包含一个V型结构：$X \longrightarrow Y \longleftarrow Z$，其中$Y$及其后代结点$\notin \mathfrak{S}$

  则该路径由$\mathfrak{S}$有向分离

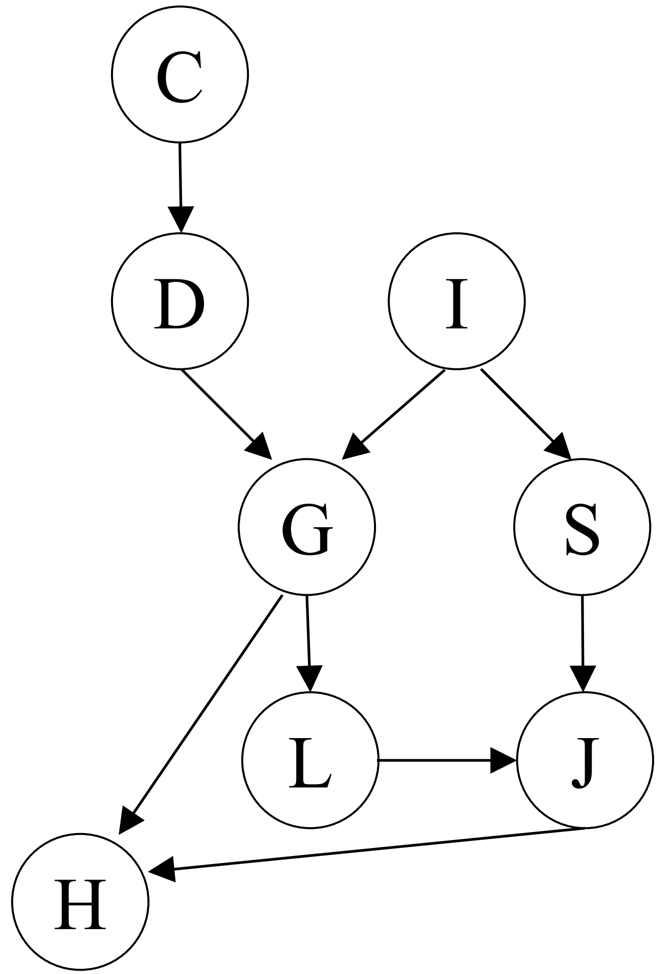
课前练习

- 给定如下贝叶斯网络，判断下列4个式子的真假。
  - $(B \perp D \mid A)$
  - $(B \perp D \mid C)$
  - $(B \perp D \mid E)$
  - $(B \perp C \mid A)$



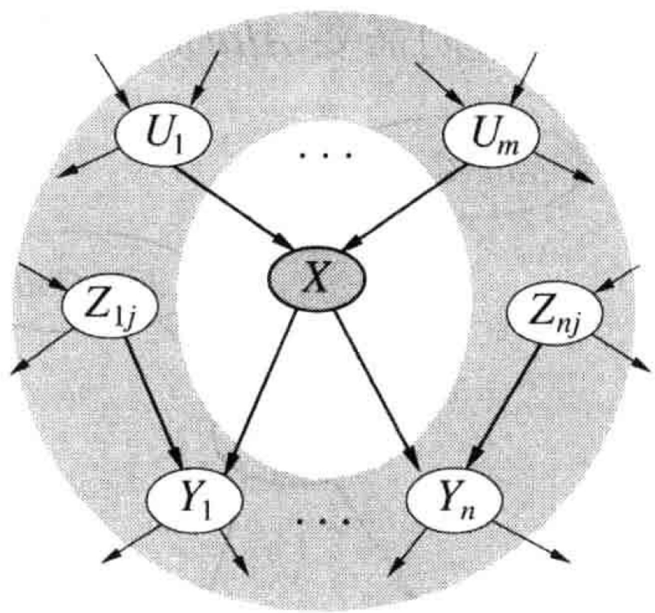
课前练习

- 给定如下贝叶斯网络，判断下列4个式子的真假。
  - $(D \perp I \mid L)$
  - $(D \perp J \mid L)$
  - $(D \perp J \mid L, I)$
  - $(D \perp J \mid L, H, I)$



马尔科夫覆盖

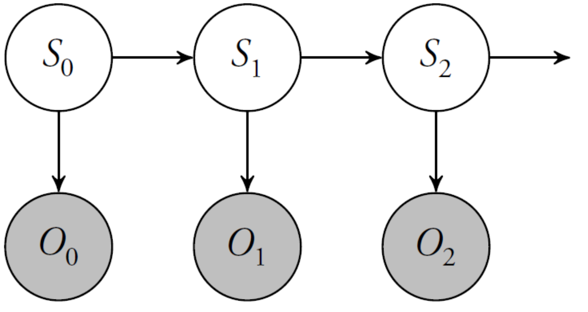
- 马尔科夫覆盖（Markov Blanket）
  - 有向分离一个结点与其他结点的最少个数的结点构成的集合
  - 这个集合由该结点的父结点、子结点以及子结点的父结点构成



灰色区域为结点X的马尔科夫覆盖

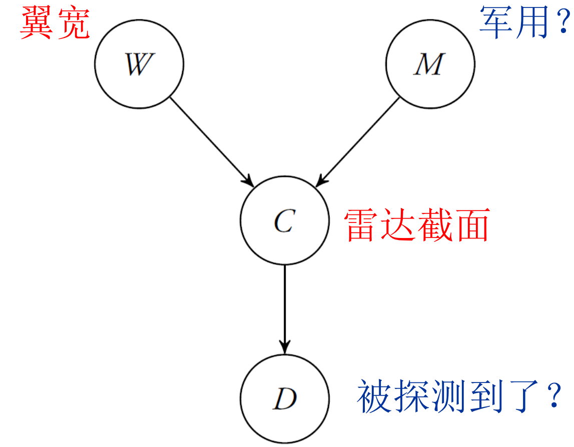
课前练习

- 以下是一个隐马尔科夫模型的贝叶斯网络，请找出结点O_1的马尔科夫覆盖。

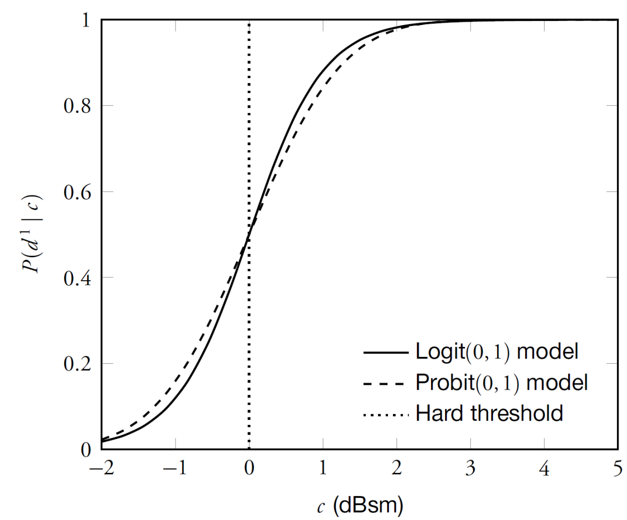


混合贝叶斯网络

- 包含离散的和连续的变量的贝叶斯网络
- 建模$P(C \mid W, M)$
  - 先忽略二值变量M，用线性高斯分布建模
    $$\mathrm{p}(c \mid w) = \mathcal{N}(c \mid \theta_1 w + \theta_2, \theta_3)$$
  - 再把二值变量M考虑进来，得到
    $$\mathrm{p}(c \mid w, m) = \begin{cases} \mathcal{N}(c \mid \theta_1 w + \theta_2, \theta_3) & 如果 m^0 \\ \mathcal{N}(c \mid \theta_4 w + \theta_5, \theta_6) & 如果 m^1 \end{cases}$$
- 建模$P(D \mid C)$
  - 硬性阈值： $P(d^1 \mid c) = \begin{cases} 0 & 如果 c < \theta \\ 1 & 否则 \end{cases}$
  - logit模型： $P(d^1 \mid c) = \dfrac{1}{1 + \exp\left(-2\dfrac{c - \theta_1}{\theta_2}\right)}$
  - probit模型：$P(d^1 \mid c) = \Phi(c - \theta_1)/\theta_2$



连续变量：翼宽、雷达截面
二值变量：军用？、被探测到了？



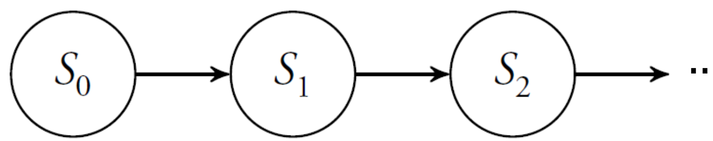
不确定性的表示

- 信念度和概率
- 概率分布
- 联合概率分布
- 条件概率分布
- 贝叶斯网络
- **时序模型**

马尔科夫链

- 时序模型表示一组变量如何随时间演进
- 马尔科夫链
  - S_t表示t时刻的状态
  - 状态转移模型： 条件分布$\boxed{P(S_t \mid S_{t-1})}$
  - 初始分布：$P(S_0)$

满足马尔科夫性质：系统下一时刻的状态仅由当前状态决定，不依赖于以往的任何状态



马尔科夫链的贝叶斯网络

- 如果条件分布不随t变化，则称模型是稳态的

马尔科夫链的实例

- 实例
  - 向量 $\boldsymbol{s}_{\mathrm{t}}$表示t时刻的状态，由飞机在t时刻的高度和垂直速率构成
  - 用线性高斯函数表示状态转移分布：

$$p(\boldsymbol{s}_{\mathrm{t}} \mid \boldsymbol{s}_{\mathrm{t-1}}) = \mathcal{N}(\boldsymbol{s}_{\mathrm{t}} \mid \boldsymbol{M}\,\boldsymbol{s}_{\mathrm{t-1}} + \boldsymbol{b}, \boldsymbol{\Sigma})$$

匀速运动：$\boldsymbol{M}\,\boldsymbol{s}_{\mathrm{t-1}} + \boldsymbol{b} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \boldsymbol{s}_{\mathrm{t-1}}$

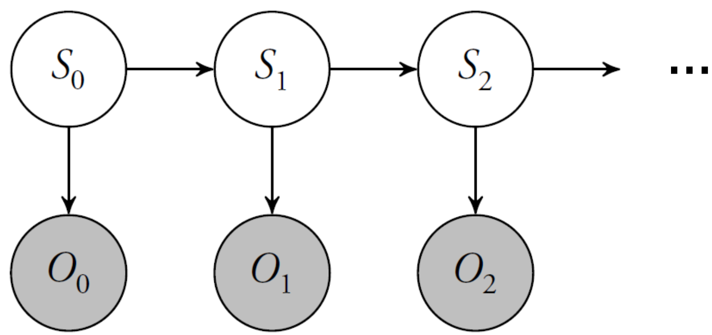
  - 初始分布： $p(\boldsymbol{s}_0) = \mathcal{N}(\boldsymbol{s}_0 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

k维高斯分布：

$$\mathcal{N}(\boldsymbol{s}_0 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})\right)$$

隐马尔科夫模型

- 在马尔科夫链上增加观察结点
  - 隐马尔科夫模型：状态变量是离散的



隐马尔科夫模型的贝叶斯网络

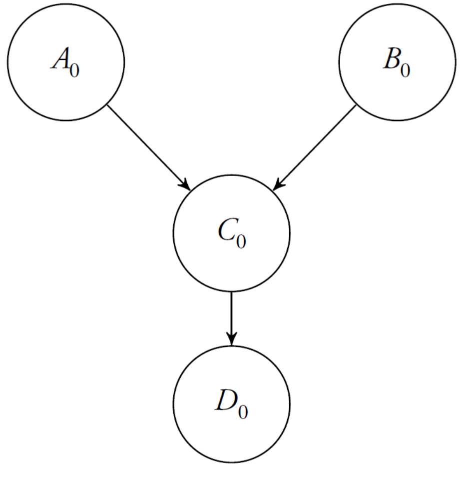
线性动态系统

- 在马尔科夫链上增加观察结点
  - 线性动态系统：状态变量是连续的且条件分布是线性高斯分布
- 实例
  - 观察是有噪声的高度信息

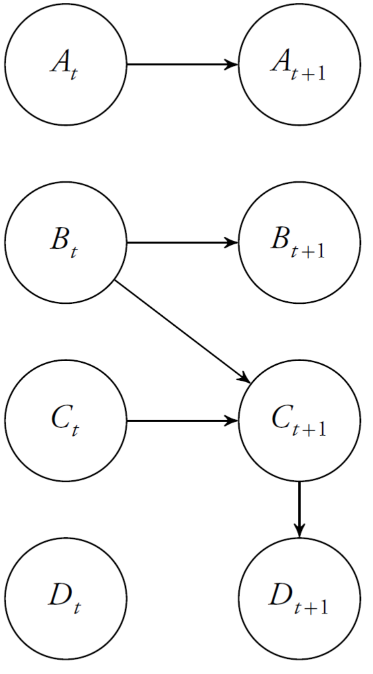
$$p(\boldsymbol{o}_{\mathrm{t}} \mid \boldsymbol{s}_{\mathrm{t}}) = \mathcal{N}(\boldsymbol{o}_{\mathrm{t}} \mid [1 \quad 0]\, \boldsymbol{s}_{\mathrm{t}}, \boldsymbol{\Sigma})$$

动态贝叶斯网络

- 可以用动态贝叶斯网络来表示稳态的时序模型
  - 一个贝叶斯网络表示初始分布
  - 一个贝叶斯网络表示转移分布



初始分布



转移分布

小结：不确定性的表示

- 不确定性的量化
  - 信念度、概率
  - 条件概率、全概率法则、贝叶斯规则
  - 概率分布、联合概率分布、条件概率分布
- 贝叶斯网络
  - 有向无环图，联合分布的一种紧凑表示
  - 条件独立关系：有向分离、马尔科夫覆盖
  - 混合贝叶斯网络
- 时序模型
  - 表示一组变量如何随时间演进
  - 马尔科夫链：状态、状态转移模型、初始分布
  - 在马尔科夫链上增加观察结点：隐马尔科夫模型、线性动态系统

课后练习2.1

- 在火星上，有50%的概率既有生命又有水，有25%的概率有生命但没有水，有25%的概率既没有生命又没有水。问：在给定有水的前提下，火星上有生命的概率是多少？

课后练习2.2

- 给定一个状态序列为$s_{0:t}$和观察序列为$o_{0:t}$的隐马尔科夫模型，试证明：

$$P(s_t \mid o_{0:t}) \propto P(o_t \mid s_t, o_{0:t-1}) P(s_t \mid o_{0:t-1})$$

请使用上式证明：

$$P\left(s_t \mid o_{0:t}\right) \propto P\left(o_t \mid s_t\right) \sum_{s_{t-1}} P\left(s_t \mid s_{t-1}\right) P\left(s_{t-1} \mid o_{0:t-1}\right)$$