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

第二部分：概率模型

章宗长

2021年3月24日

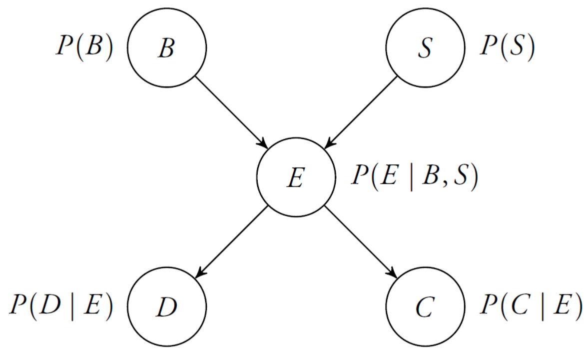
内容安排

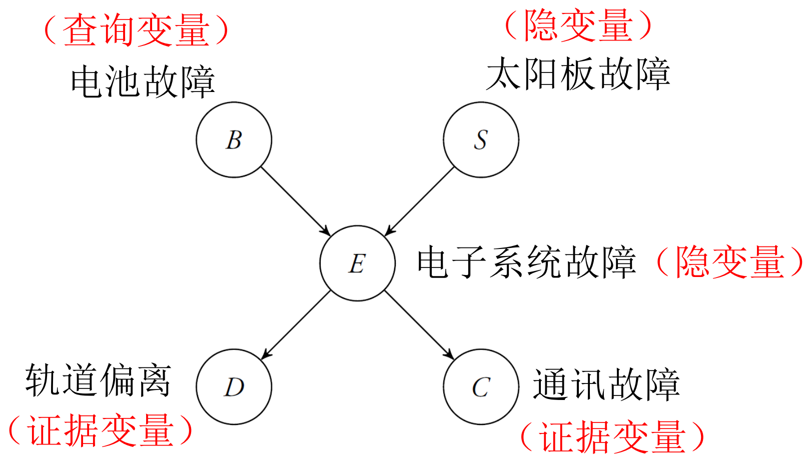
概率推理

- 贝叶斯网络中的推理
- 分类推理
- 时序模型中的推理
- 精确推理
- 精确推理的复杂度
- **近似推理**

直接采样法

- 例子：卫星监控问题



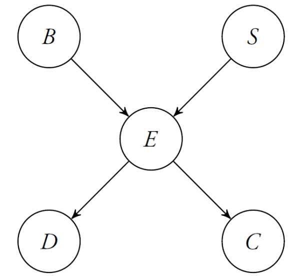


- 想要推出概率$P(b^1 \mid d^1, c^1)$
- 直接采样法：从联合分布$P(B,S,E,D,C)$中采样n个样本，然后用下式估计

$$P(b^1 \mid d^1, c^1) \approx \frac{\sum_i (b^{(i)} = 1 \wedge d^{(i)} = 1 \wedge c^{(i)} = 1)}{\sum_i (d^{(i)} = 1 \wedge c^{(i)} = 1)}$$

第i个样本：
$(b^{(i)}, s^{(i)}, e^{(i)}, d^{(i)}, c^{(i)})$

拓扑排序



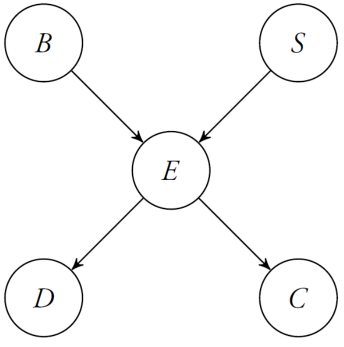
- 贝叶斯网络中结点的拓扑排序：
  - 有序列表，使得：如果图中有边$A \rightarrow B$，那么A出现在B之前
  - 拓扑排序总存在，可能不唯一
  - （右上图）的4种拓扑排序：B, S, E, D, C；B, S, E, C, D；S, B, E, D, C；S, B, E, C, D
- 寻找图G的一种拓扑排序的方法

Algorithm 2.3 Topological sort

function TopologicalSort(G)
    $n \leftarrow$ number of nodes in G
    $L \leftarrow$ empty list
    for $i \leftarrow 1$ **to** n
        $X \leftarrow$ any node not in L but all of whose parents are in L
        Add X to end of L
    return L

拓扑排序（续）

- 从条件概率分布中采样
- 链式法则： $P(x_1,\dots,x_n)=\prod_{i=1}^{n} P(x_i \mid Pa_{x_i})$
- 例子：
  - 4种拓扑排序
    - B, S, E, D, C；B, S, E, C, D；S, B, E, D, C；S, B, E, C, D
  - $P(b,s,e,d,c)=P(b)P(s)P(e \mid b,s)\,P(d \mid e)P(c \mid e)$



直接采样法（续）

Algorithm 2.4 Direct sampling from a Bayesian network

> **function** DirectSample(B)
>
> $X_{1:n} \leftarrow$ a topological sort of nodes in B
>
> **for** $i \leftarrow 1$ **to** n
>
> $x_i \leftarrow$ a random sample from $P(X_i \mid \mathrm{pa}_{x_i})$
>
> **return** $x_{1:n}$

- 右图：在一个贝叶斯网络中，通过直接采样方法得到的样本
- $P(b^1 \mid d^1, c^1) = ?$ 0.5
- 问题：浪费了很多时间来产生与观察不一致的样本

B	S	E	D	C	
0	0	1	1	0	
0	0	0	0	0	
1	0	1	0	0	
1	0	1	1	1	←
0	0	0	0	0	
0	0	0	1	0	
0	0	0	0	1	
0	1	1	1	1	←
0	0	0	0	0	
0	0	0	1	0	

似然加权法

- 特点：产生与观察一致的加权样本
- 例子：想要推出概率$P(b^1 | d^1, c^1)$
- 似然加权法：从联合分布$P(B, S, E, d^1, c^1)$中采样n个样本
  - 从联合分布$P(B, S, E)$中采样n个样本
  - 第i个样本：$P\left(b^{(i)}, s^{(i)}, e^{(i)}\right) = P\left(b^{(i)}\right)P\left(s^{(i)}\right)P\left(e^{(i)} | b^{(i)}, s^{(i)}\right)$
  - 令第i个样本的权值$w_i\left(d^{(i)} = 1 \wedge c^{(i)} = 1\right)$： $P\left(d^1 | e^{(i)}\right)P\left(c^1 | e^{(i)}\right)$
  - 则有，$P\left(b^{(i)}, s^{(i)}, e^{(i)}\right) w_i\left(d^{(i)} = 1 \wedge c^{(i)} = 1\right) = P\left(b^{(i)}, s^{(i)}, e^{(i)}, d^1, c^1\right)$

$$P(b^1 \mid d^1, c^1) \approx \frac{\sum_i w_i(b^{(i)} = 1 \wedge d^{(i)} = 1 \wedge c^{(i)} = 1)}{\sum_i w_i(d^{(i)} = 1 \wedge c^{(i)} = 1)}$$

$$= \frac{\sum_i w_i(b^{(i)} = 1 \wedge d^{(i)} = 1 \wedge c^{(i)} = 1)}{\sum_i w_i}$$

似然加权法（续）

B	S	E	D	C	Weight
1	0	1	1	1	$P(d^1 \mid e^1)P(c^1 \mid e^1)$
0	1	1	1	1	$P(d^1 \mid e^1)P(c^1 \mid e^1)$
0	0	0	1	1	$P(d^1 \mid e^0)P(c^1 \mid e^0)$
0	0	0	1	1	$P(d^1 \mid e^0)P(c^1 \mid e^0)$
0	0	1	1	1	$P(d^1 \mid e^1)P(c^1 \mid e^1)$

- 右图：在卫星监控问题的贝叶斯网络中，通过似然加权方法得到的样本
- 跟直接采样方法一样，从$P(B), P(S)$和$P(E \mid B,S)$中采样
- 当遇到D和C时，赋值$D=1$和$C=1$
- 如果样本有$E=1$，那么权值为$P(d^1 \mid e^1)P(c^1 \mid e^1)$，否则为$P(d^1 \mid e^0)\ P(c^1 \mid e^0)$
- 假设：$P(d^1 \mid e^1)P(c^1 \mid e^1)$ =0.95， $P(d^1 \mid e^0)P(c^1 \mid e^0)$=0.01
- 估算出： $P(b^1 \mid d^1, c^1) = \dfrac{0.95}{0.95+0.95+0.01+0.01+0.95} \approx 0.331$

似然加权法（续）

Algorithm 2.5 Likelihood-weighted sampling from a Bayesian network

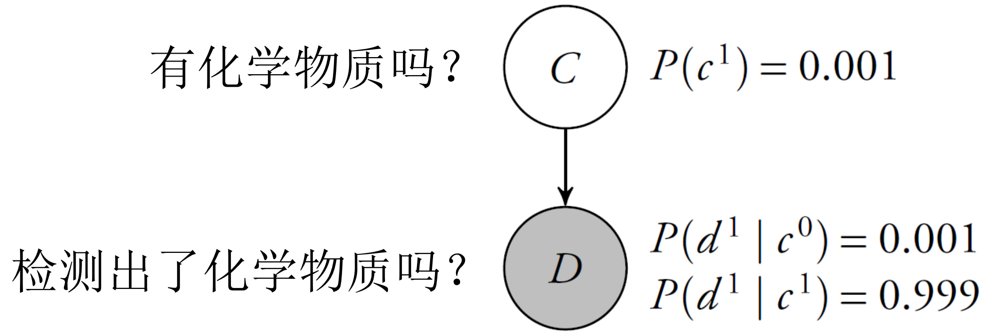
function LikelihoodWeightedSample($B, o_{1:n}$)
    $X_{1:n} \leftarrow$ a topological sort of nodes in B
    $w \leftarrow 1$
    for $i \leftarrow 1$ **to** n
        if $o_i =$ NIL
            $x_i \leftarrow$ a random sample from $P(X_i \mid \text{pa}_{x_i})$
        else
            $\boxed{x_i \leftarrow o_i}$
            $\boxed{w \leftarrow w \times P(x_i \mid \text{pa}_{x_i})}$
    return $(x_{1:n}, w)$

B：贝叶斯网络
$o_{1:n}$：观察到的值

产生与观察一致的加权样本

样本的权值：在观察到的结点处的条件概率的乘积

似然加权法（续）



- 推理$P(c^1 \mid d^1)$
- 由贝叶斯规则，有

$$P(c^1 \mid d^1)=\frac{P(d^1 \mid c^1)P(c^1)}{P(d^1 \mid c^1)P(c^1)+P(d^1 \mid c^0)P(c^0)}$$
$$=\frac{0.999\times 0.001}{0.999\times 0.001+0.001\times 0.999}$$
$$=0.5.$$

- 如果使用似然加权，那么99.9%的样本有$C=0$
- 在获得权值为0.999的样本$C=1$之前，$P(c^1 \mid d^1)$的估计将是0

吉布斯采样法

- 马尔科夫链蒙特卡洛（Markov Chain Monte Carlo, MCMC）
  - 概率模型中最常用的采样技术
  - MCMC算法家族的成员：吉布斯（Gibbs）采样算法、模拟退火算法等

- MCMC算法
  - 构造一个马尔科夫链
    - 满足：当其收敛至稳态分布时，该稳态分布恰为待估计参数的后验分布
  - 通过这个马尔科夫链来随机产生符合后验分布的样本，并基于这些样本来进行估计

马尔科夫链转移概率的构造至关重要，不同的构造方法将产生不同的MCMC算法

马尔科夫链（Markov Chain）

- 马尔科夫链可以表示为一个三元组$< S,\ \boldsymbol{\pi}(0),\ \boldsymbol{P} >$
  - S：状态集合
  - $\boldsymbol{\pi}(0)$：初始状态分布
  - $\boldsymbol{P} = [p_{ij}]$：状态转移矩阵

- 令N为状态数，$\pi_i(n)$为在第n次转移后状态i的概率，则有

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \pi_i(n) = \mathbf{1},\quad \pi_j(n+1) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \pi_i(n) p_{ij}\, , \text{for } n = 0, 1, 2, \ldots$$

- 写成向量形式，有

马尔科夫性质：时刻$n+1$的状态分布只有时刻n的状态分布有关，与更早时刻的状态分布无关

$$\boldsymbol{\pi}(n+1) = \boldsymbol{\pi}(n)\boldsymbol{P}, \text{for } n = 0, 1, 2, \ldots$$

例子：马尔科夫链及稳态分布

- 有两个状态的马尔科夫链，其状态转移矩阵$\boldsymbol{P}$为

$$\boldsymbol{P}=[p_{ij}]=\begin{bmatrix}\frac{1}{2} & \frac{1}{2}\\ \frac{2}{5} & \frac{3}{5}\end{bmatrix}$$

$$\begin{aligned}\pi(1) &= \pi(0)\boldsymbol{P}\\ \pi(2) &= \pi(1)\boldsymbol{P}=\pi(0)\boldsymbol{P}^2\\ \pi(3) &= \pi(2)\boldsymbol{P}=\pi(0)\boldsymbol{P}^3\\ \cdots & \quad \cdots\\ \pi(n) &= \pi(0)\boldsymbol{P}^n \text{ for } n=0,1,2,\ldots\end{aligned}$$

- 假设$\boldsymbol{\pi}(0)=[1,0]^{\mathrm{T}}$，有

$n=$	0	1	2	3	4	5	...
$\pi_1(n)$	1	0.5	0.45	0.445	0.4445	0.44445	....
$\pi_2(n)$	0	0.5	0.55	0.555	0.5555	0.55555	....

- 假设$\boldsymbol{\pi}(0)=[0,1]^{\mathrm{T}}$，有

$n=$	0	1	2	3	4	5	...
$\pi_1(n)$	0	0.4	0.44	0.444	0.4444	0.44444	....
$\pi_2(n)$	1	0.6	0.56	0.556	0.5556	0.55556	....

收敛到了稳态分布：与初始状态分布无关

吉布斯采样法

- 从任意样本（将证据变量固定为观察值）出发，通过对非证据变量逐个进行采样改变其取值，生成下一个样本
- 算法2.6：从一个已存在的样本$x_{1:n}$中产生一个新样本$x'_{1:n}$

Algorithm 2.6 Gibbs sampling from a Bayesian network

```
function GibbsSample(B, o_{1:n}, x_{1:n})
    X_{1:n} ← an ordering of nodes in B
    x'_{1:n} ← x_{1:n}
    for i ← 1 to n
        if o_i = NIL
            x'_i ← a random sample from P(X_i | x'_{1:n\i})
        else
            x'_i ← o_i
    return x'_{1:n}
```

不必是拓扑排序

每个样本都应满足证据变量的值等于观察到的值

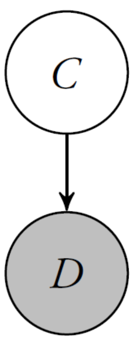
- 循环执行算法2.6得到样本序列，即马尔科夫链

吉布斯采样（续）

- 吉布斯采样方法中的第6行：非证据变量X_i的采样方法

Algorithm 2.7 Distribution at a node given observations at all other nodes

1: **function** DistributionAtNode($B, X_i, x_{1:n\setminus i}$)
2: $\mathcal{T} \leftarrow$ all conditional probability tables associated with B involving X_i
3: Remove rows that are inconsistent with $x_{1:n\setminus i}$ from all the tables in $\mathcal{T}$
4: $T \leftarrow$ product of the tables remaining in $\mathcal{T}$
5: $P(X_i \mid x_{1:n\setminus i}) \leftarrow$ normalize T
6: **return** $P(X_i \mid x_{1:n\setminus i})$



$P(c^1) = 0.001$

$P(d^1 \mid c^0) = 0.001$
$P(d^1 \mid c^1) = 0.999$

- 推理$P(c^1 \mid d^1) \Rightarrow$ 采样非证据变量C

$\mathcal{T}$:

$P(c^0) = 0.999$
$P(c^1) = 0.001$

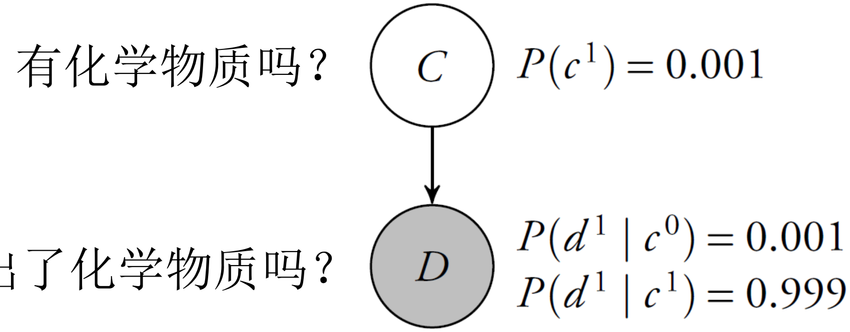
$P(d^1 \mid c^0) = 0.001$
$P(d^1 \mid c^1) = 0.999$

T:

$P(c^0)P(d^1 \mid c^0) = 0.000999$
$P(c^1)P(d^1 \mid c^1) = 0.000999$

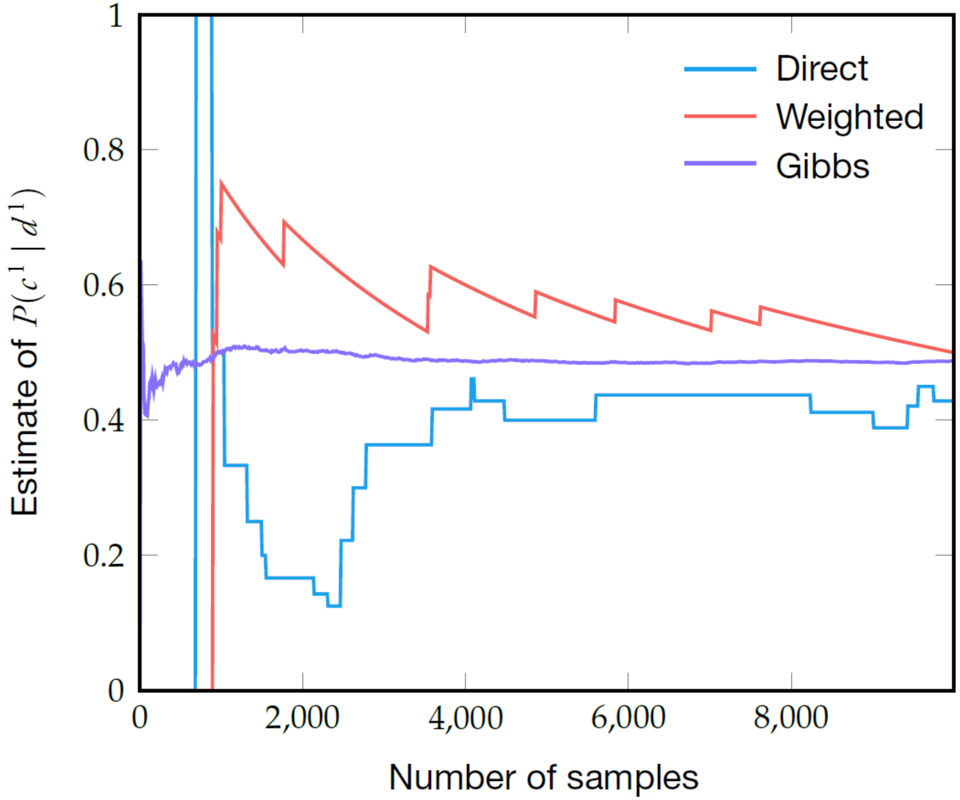
$P(c^0 \mid d^1) = 0.5$
$P(c^1 \mid d^1) = 0.5$

近似推理方法对比



- 推理$P(c^1 \mid d^1)$

- 实验对比
  - 直接采样方法（Direct）
  - 加权似然方法（Weighted）
  - 吉布斯采样方法（Gibbs）



小结：概率推理

- 概率推理
  - 已知概率模型，由一组证据变量的值确定一个或多个查询变量的分布
- 分类推理
  - 用于分类任务，从给定的一组观察或特征中推理所属类别
  - 朴素贝叶斯模型
- 时序模型中的推理
  - 滤波、预测、平滑、最可能序列
- 精确推理
  - 枚举法、变量消去法、信念传播法
  - 复杂度：NP-难
- 近似推理
  - 直接采样方法、加权似然方法、吉布斯采样方法

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

参数学习

- **极大似然参数学习：离散模型**
- 极大似然参数学习：连续模型
- 贝叶斯参数学习
- 非参数化模型的密度估算

极大似然参数学习：二项分布

- 假设随机变量C表示一架飞机是否将发生空中碰撞，想估计分布$P(C)$
- 因为C是0或1，所以估计参数$\theta = P(c^1)$就够了
- 假设D是一个跨度有十年的数据库：有n架飞机，其中m架发生了空中碰撞，想从D中推出θ
- 需要从D中学到θ的极大似然估计：

$$\hat{\theta} = \mathrm{argmax}_{\theta}\, P(D \mid \theta)$$

- 直觉地，给定D，θ的一个好的估计是$\frac{m}{n}$

极大似然参数学习：二项分布（续）

- 给定θ，n架飞机中有m架发生了空中碰撞的条件概率为：

$$P(D \mid \theta) = \frac{n!}{m!(n-m)!}\theta^m(1-\theta)^{n-m}$$
$$\propto \theta^m(1-\theta)^{n-m}$$

- 最大化$P(D \mid \theta)$ 等于最大化它的对数，即对数似然$\ell(\theta)$：

$$\ell(\theta) \propto \ln\left(\theta^m(1-\theta)^{n-m}\right)$$
$$= m\ln\theta + (n-m)\ln(1-\theta)$$

- 令$\ell(\theta)$的导数为0：

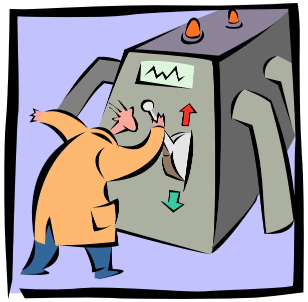
$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{m}{\hat{\theta}} - \frac{n-m}{1-\hat{\theta}} = 0$$

解得：$\hat{\theta} = \frac{\mathrm{m}}{n}$

极大似然参数学习：标准步骤

- 标准步骤
  - 为数据的似然性写下一个表达式，即参数的函数
  - 写下对数似然关于每个参数的偏导数
  - 推导出使导数为0的参数值

最需要技巧的步骤通常是最后一步



极大似然参数学习：多项分布

- 假设变量X有k个值，这k个值在D中被观察到的次数为$\mathrm{m}_{1:k}$
- 需要求解$P\left(x^i \,\middle|\, \mathrm{m}_{1:k}\right)$的极大似然估计
- 假设$\theta_i = P(x^i)$，$i = \{1, 2, \ldots, k\}$，则$\sum_{i=1}^{k} \theta_i = 1$

$$P(D \mid \theta_{1:k}) = \frac{(m_1 + \cdots \ + m_k)!}{m_1!\, m_2! \cdots m_k!} \theta_1^{m_1} \theta_2^{m_2} \cdots \theta_k^{m_k}$$

$$\propto \theta_1^{m_1}\ \theta_2^{m_2} \cdots \theta_k^{m_k}$$

$$\ln P(D \mid \theta_{1:k}) \propto \sum_{i=1}^{k} m_i \ln \theta_i$$

极大似然参数学习：多项分布（续）

- 需要从D中学到$\theta_{1:k}$的极大对数似然：

$$\hat{\theta}_{1:k} = \operatorname{argmax}_{\theta_{1:k}} \ln P(D \mid \theta_{1:k}) \quad \text{其中} \textstyle\sum_{i=1}^{k} \theta_i = 1$$

- 采用拉格朗日乘子法，得到：

$$\ell(\theta_{1:k}, \lambda) = \ln P(D \mid \theta_{1:k}) + \lambda\left(1 - \sum_{i=1}^{k} \theta_i\right) \propto \sum_{i=1}^{k} m_i \ln \theta_i + \lambda\left(1 - \sum_{i=1}^{k} \theta_i\right)$$

- 分别对θ_i和λ求偏导并令其为0，得到：

$$\frac{\partial \ell(\theta_{1:k}, \lambda)}{\partial \theta_i} = \frac{m_i}{\hat{\theta}_i} - \lambda = 0, i = \{1, 2, \dots, k\}$$

$$\frac{\partial \ell(\theta_{1:k}, \lambda)}{\partial \lambda} = 1 - \sum_{i=1}^{k} \hat{\theta}_i = 0$$

联立这$k+1$个方程，解得

$$\hat{\theta}_i = \frac{m_i}{\sum_{j=1}^{k} m_j}$$

参数学习

- 极大似然参数学习：离散模型
- **极大似然参数学习：连续模型**
- 贝叶斯参数学习
- 非参数化模型的密度估算

极大似然参数学习：高斯分布

- 学习一元高斯密度函数的参数
- 数据按下式产生：

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- 令数据为$x_{1:n}$，则对数似然：

$$\ell(\mu, \sigma^2) = \ln \mathcal{N}(x_{1:n} \mid \mu, \sigma^2) = \sum_{j=1}^{n} \ln \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_j-\mu)^2}{2\sigma^2}\right)$$

$$= n(-\ln\sqrt{2\pi} - \ln\sigma) - \sum_{j=1}^{n} \frac{(x_j-\mu)^2}{2\sigma^2}$$

极大似然参数学习：高斯分布（续）

- 令其关于参数μ和σ的导数为0：

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \mu} = -\sum_{j=1}^{n} \frac{\left(x_j - \hat{\mu}\right)}{\hat{\sigma}^2} = 0$$

$$\frac{\partial \ell(\mu,\sigma^2)}{\partial \sigma} = -\frac{n}{\hat{\sigma}} + \sum_{j=1}^{n} \frac{\left(x_j - \hat{\mu}\right)^2}{\hat{\sigma}^3} = 0$$

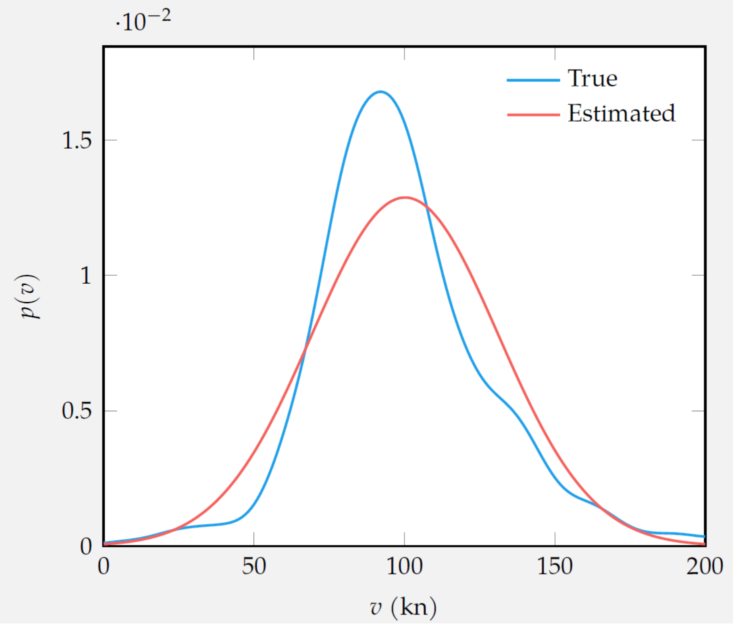
- 得到参数μ和σ的极大似然估计：

$$\hat{\mu} = \frac{1}{n}\sum\nolimits_{j=1}^{n} x_j$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum\nolimits_{j=1}^{n} \left(x_j - \hat{\mu}\right)^2$$

极大似然参数学习：高斯分布（续）

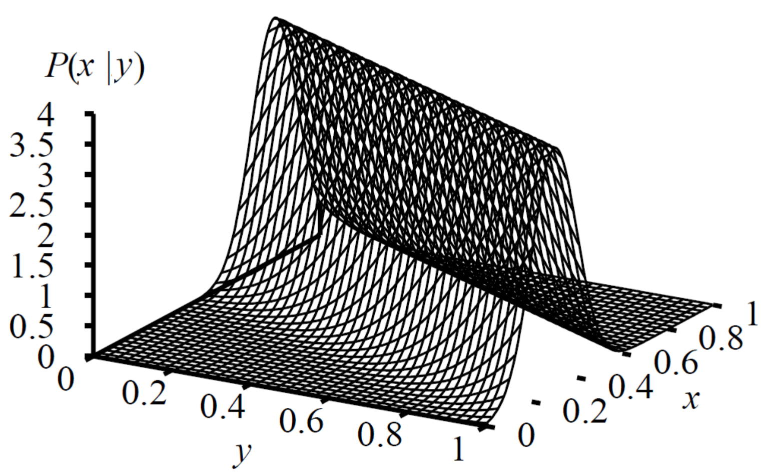
- 假设有n架飞机的飞行速度数据，用高斯模型来拟合数据，模型中的参数用极大似然方法来估计
  - $\hat{\mu}$ =100.2kt和$\hat{\sigma}$ =31kt由极大似然估计得到



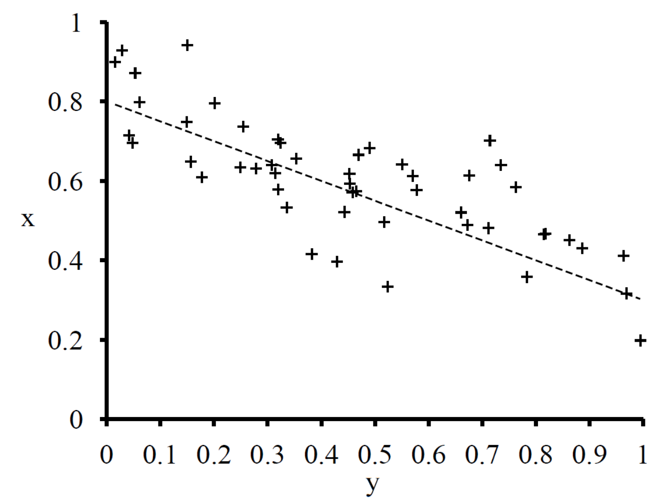
真实的和拟合的飞行速度概率密度

极大似然参数学习：线性高斯分布

- 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：$p(x \mid y) = \mathcal{N}(x \mid my + b, \sigma^2)$
- 从数据$(y_{1:n}, x_{1:n})$中学习线性高斯模型的参数m、b和σ



线性高斯模型



从该模型产生的50个数据点的集合

参数学习

- 极大似然参数学习：离散模型
- 极大似然参数学习：连续模型
- **贝叶斯参数学习**
- 非参数化模型的密度估算

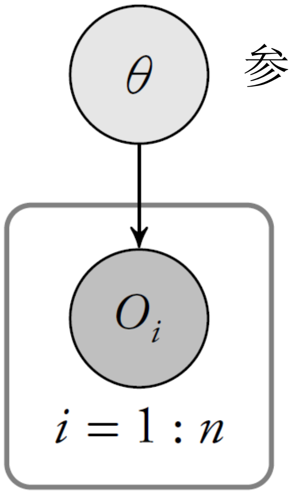
贝叶斯学习 vs. 极大后验学习

- 贝叶斯学习
  - 给定数据，计算每个假说的概率，并基于这些概率做决策
  - 用所有假说做预测，而不是使用单个“最好”的假说
  - 把学习归约于概率推理

- 参数学习中的贝叶斯方法：基于假说先验，估计θ的后验分布

- 极大后验（Maximum A Posteriori, MAP）学习
  - 基于单个最可能的假说（极大后验假说）做预测
  - 当假说先验是均匀分布时，归约为选择一个极大似然假说
  - 比贝叶斯学习更容易，要解决一个优化问题，而不是一个大规模求和（或积分）的问题

贝叶斯参数学习：二项分布



- 把参数学习过程视为贝叶斯网络中的推理过程
- （右图）用贝叶斯网络表示碰撞概率估计的问题
- 如果第i架飞机发生了碰撞，那么观察到的变量O_i为1，否则为0
- 假设观察到的变量是彼此独立的
- 具体化$P(\theta)$和$P(O_\mathrm{i} \mid \theta)$
  - 如果想使用均匀分布，则设置密度$P(\theta) = 1$
  - 设置$P\left(O_\mathrm{i}^1 \middle| \theta\right) = \theta$

贝叶斯参数学习：二项分布（续）

- 假设先验为均匀分布，则有：

$$
\begin{aligned}
p(\theta \mid o_{1:n}) &\propto p(\theta, o_{1:n}) \\
&= p(\theta)\prod_{i=1}^{n} P(o_i \mid \theta) \\
&= \prod_{i=1}^{n} P(o_i \mid \theta) \\
&= \prod_{i=1}^{n} \theta^{o_i}(1-\theta)^{1-o_i} \\
&= \theta^{m}(1-\theta)^{n-m}
\end{aligned}
$$

m: 发生了碰撞的飞机数

贝叶斯参数学习：二项分布（续）

$$\boxed{p(\theta \mid o_{1:n}) \propto \theta^m (1-\theta)^{n-m}}$$

- 归一化的常数可通过积分求得：

$$\int_0^1 \theta^m (1-\theta)^{n-m}\, d\theta = \frac{\Gamma(m+1)\Gamma(n-m+1)}{\Gamma(n+2)}$$

Γ为伽玛函数

- 把归一化放入公式里，有：

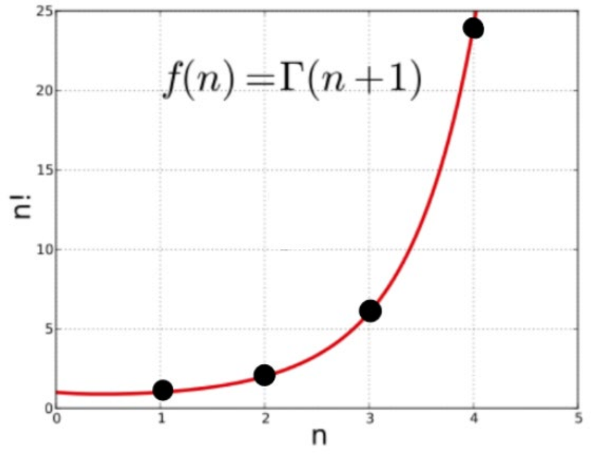
$$\begin{aligned} p(\theta \mid o_{1:n}) &= \frac{\Gamma(n+2)}{\Gamma(m+1)\Gamma(n-m+1)} \theta^m (1-\theta)^{n-m} \\ &= \mathrm{Beta}(\theta \mid m+1, n-m+1) \end{aligned}$$

伽玛函数

- 在实数域上伽玛函数的定义式：

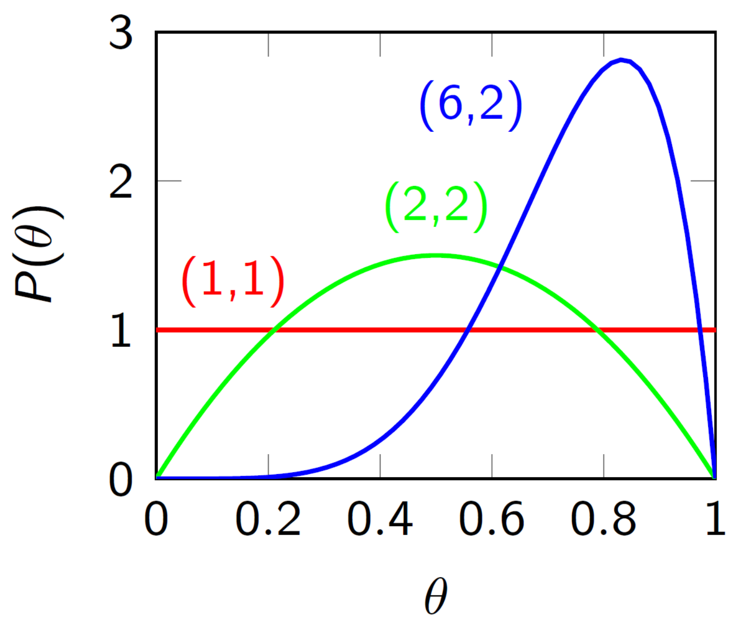
$$\Gamma(x)=\int_{0}^{+\infty}t^{x-1}e^{-t}dt \qquad (x>0)$$

- 伽玛函数是阶乘在实数域的广义形式
  - 如果n为整数，则有$\Gamma(n)=(n-1)!$



贝塔分布

- 贝塔（Beta）分布可以作为二项分布参数的先验分布
- 若选贝塔分布作为先验，后验也是贝塔分布
- Beta(1,1)是[0,1]上的均匀分布

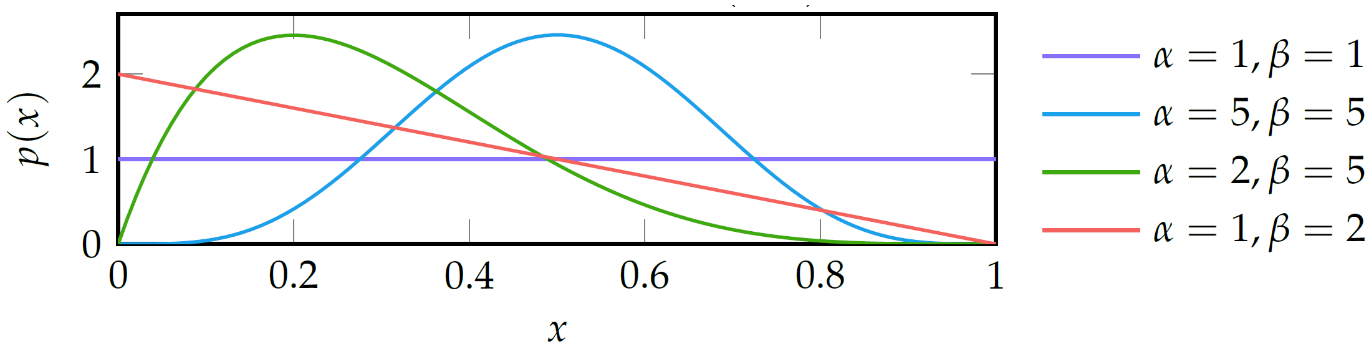


- Beta(α,β)的均值为
$$\frac{\alpha}{\alpha+\beta}$$
- Beta(α,β)的MAP估计为
$$\frac{\alpha-1}{\alpha+\beta-2}$$

条件：$\alpha \geq 1$，$\alpha+\beta>2$

贝叶斯参数学习：二项分布（续）

- 若先验为Beta(α,β)，观察为o_{i}
  - 如果o_{i}=1，则后验为$Beta(\alpha+1,\beta)$
  - 如果o_{i}=0，则后验为$Beta(\alpha,\beta+1)$



- 若先验为Beta(α,β)，数据中显示n架飞机中有m架发生了碰撞，则后验为Beta$(\alpha+m,\beta+n-m)$

- 有时称α和β为伪计数，但他们不必是整数

狄利克雷分布

- 狄利克雷（Dirichlet）分布：贝塔分布的广义形式

- 狄利克雷分布的概率密度函数：

$$\mathrm{Dir}(\theta_{1:n} \mid \alpha_{1:n}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{n}\Gamma(\alpha_i)}\prod_{i=1}^{n}\theta_i^{\alpha_i - 1}$$

其中$\alpha_0 = \sum_{i=1}^{n}\alpha_i$。如果$n = 2$，则狄利克雷分布退化为贝塔分布

- 狄利克雷分布$\mathrm{Dir}(\alpha_{1:n})$的均值向量：

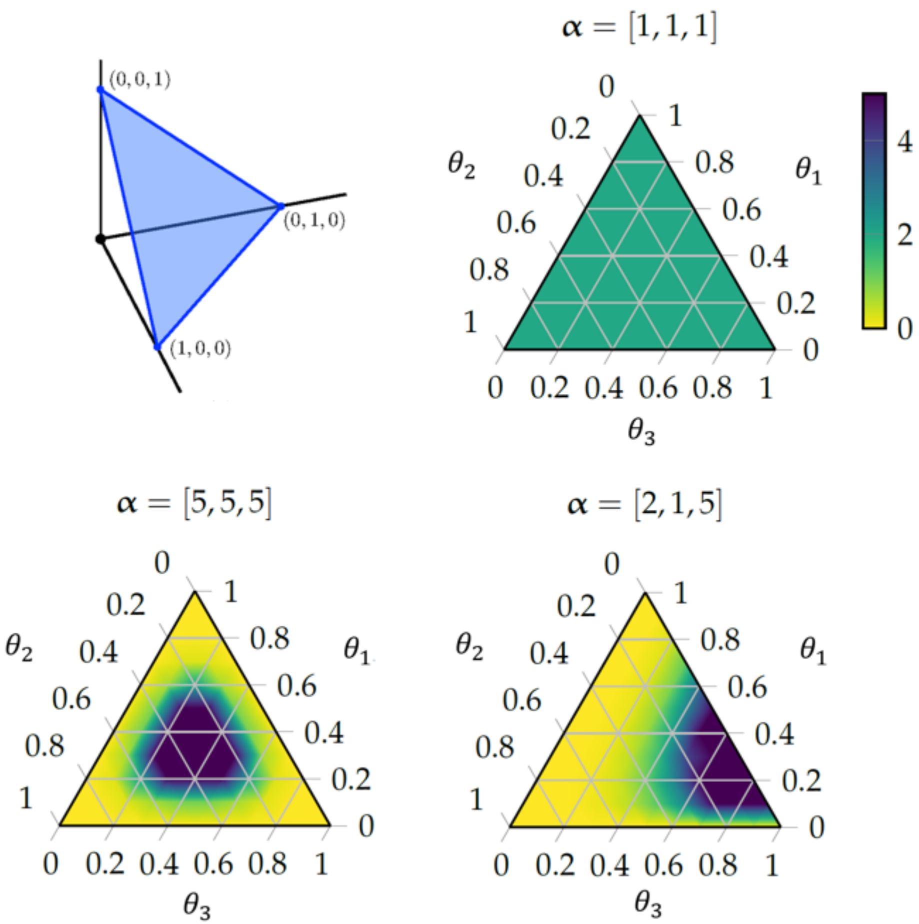
第i个元素为$\frac{\alpha_i}{\sum_{j=1}^{n}\alpha_j}$

- 狄利克雷分布$\mathrm{Dir}(\alpha_{1:n})$的MAP估计向量：

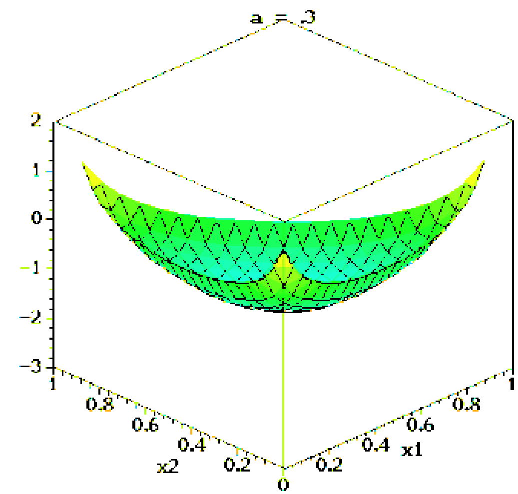
第i个元素为$\frac{\alpha_i - 1}{\sum_{j=1}^{n}\alpha_j - n}$

条件：$\alpha_i \geq 1$，$\sum_{j=1}^{n}\alpha_j > n$

狄利克雷分布的可视化



3维狄利克雷分布的示例



此图展示了当n = 3、参数从$\alpha = (0.3, 0.3, 0.3)$变化到$(2.0, 2.0, 2.0)$时，密度函数取对数后的变化

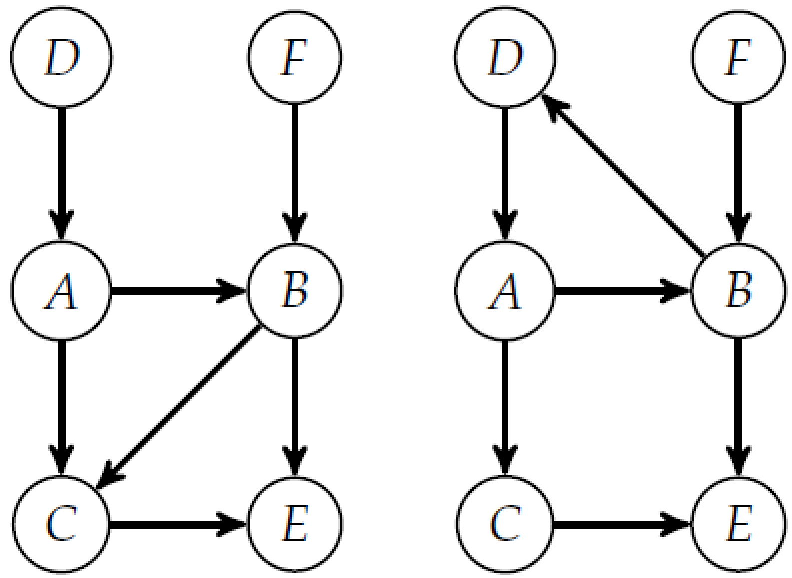
贝叶斯参数学习：多项分布

- 假设离散随机变量有n个可能值：$P(\mathrm{x}^{i}) = \theta_i$
- 狄利克雷分布可以用于表示$\theta_{1:n}$的先验和后验分布
- 狄利克雷分布由参数$\alpha_{1:n}$决定
  - 均匀先验：参数$\alpha_{1:n}$均为1
  - 参数也常用作伪计数
- 如果$\theta_{1:n}$的先验由Dir($\alpha_{1:n}$)给出，数据中观察到$X = i$有m_i次，那么$\theta_{1:n}$的后验为：

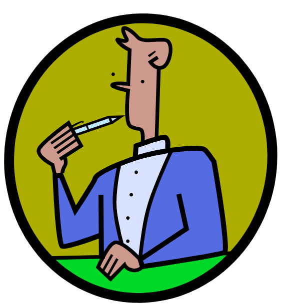
$$p(\theta_{1:n} \mid \alpha_{1:n}, m_{1:n}) = \mathrm{Dir}(\theta_{1:n} \mid \alpha_1 + m_1, \ldots, \alpha_n + m_n)$$

课后练习2.6

- 有如下两个有向图，请给出每个有向图的所有拓扑排序。如果拓扑排序不存在，试解释为什么。

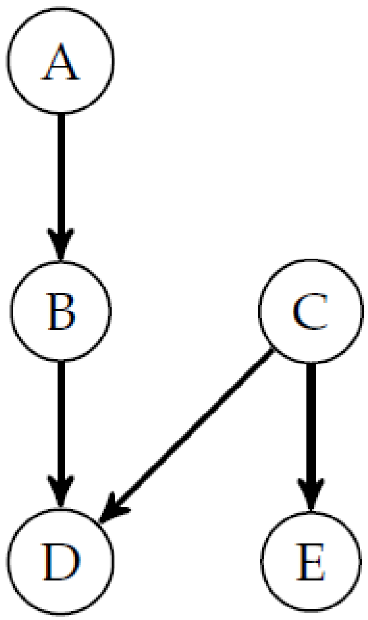


(1) (2)

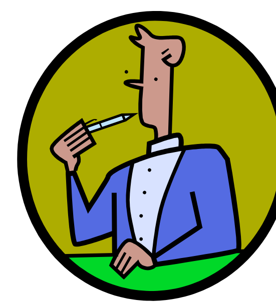


课后练习2.7

- 假设有如左下图的贝叶斯网络，想要用似然加权方法推理出$P(e^1 \mid b^0, d^1)$。右下图通过似然加权方法得到的一组样本。试写出（1）每个样本的权重表达式；（2）用样本权重w_i估计$P(e^1 \mid b^0, d^1)$的方程。



A	B	C	D	E
0	0	0	1	0
1	0	0	1	0
0	0	0	1	1
1	0	1	1	1
0	0	1	1	0
1	0	1	1	1



课后练习2.8

- 给定n个数据点(y_j, x_j)，其中x_j是按照线性高斯模型
$$p(x \mid y) = \mathcal{N}(x \mid my + b, \sigma^2)$$
从y_j产生的。试推导出使数据的条件对数似然性最大的参数m、b和σ的值。

课后练习2.9

- 假设有一个有磨损的硬币，我们想估计它正面朝上的概率，记为ϕ。如果第一次投掷的结果是正面朝上（$o_1 = 1$），则
  （1）计算ϕ的极大似然估计；
  （2）使用均匀先验假设，计算ϕ的MAP估计；
  （3）使用均匀先验假设，计算ϕ的后验分布的均值。