

第13章 概率推理

《人工智能算法》

清华大学出版社

2022年7月

提纲

- ◆ 引例
- ◆ 贝叶斯网概念
- ◆ 贝叶斯网参数学习
- ◆ 贝叶斯网结构学习
- ◆ 基于贝叶斯网的概率推理
- ◆ 总结

引例

◆ 哪些事件可能导致患者呼吸困难?

遗传、长期吸烟、长期接触致癌物、
感染COVID-19、其他因素...



患者

产生呼吸困难(T)
无呼吸困难症状(F)

◆ 现实世界的推理存在不确定性

➤ 知识不完整

患者无不良生活习惯

$$P(\text{呼吸困难}=T) = 0.01$$

➤ 信息来源不准确

已知患者长期吸烟

$$P(\text{呼吸困难}=T | \text{长期吸烟}=T) = 0.6$$

➤ 测试手段的局限性

已知患者长期吸烟，且感染了COVID-19

$$P(\text{呼吸困难}=T | \text{长期吸烟}=T, \text{感染COVID-19}=T) = 0.9$$

➤



引例

◆ 概率推理 (Probabilistic Inference) 基于概率论描述随机事件带来的不确定性



基于**专家经验**
推理的局限性

数据和知识来源多 不同领域知识学习成本高昂
实际问题日益复杂 用户需求提高



◆ 概率图模型 (Probabilistic Graphical Model)

步骤:

- ① 将随机事件视为变量
- ② 构建变量间复杂依赖关系
- ③ 设计概率推理算法

常用模型:

- 贝叶斯网 (Bayesian Network)
- 马尔科夫网 (Markov Network)
- 条件随机场 (Conditional Random Field)

应用场景: 金融分析、故障检测、医疗诊断...

提纲

- ◆ 引例
- ◆ 贝叶斯网概念
- ◆ 贝叶斯网参数学习
- ◆ 贝叶斯网结构学习
- ◆ 基于贝叶斯网的概率推理
- ◆ 总结

贝叶斯网的基本概念 (1)

- ◆ **贝叶斯网 (Bayesian Network, BN)**
 - 有向无环图 (Directed Acyclic Graph, DAG)
 - 条件概率表 (Conditional Probability Table, CPT)
- ◆ **BN定义**
 - ◆ BN表示为二元组 $B = (G, \theta)$:
 - $G = (V, E)$ 是一个DAG, 其中 $V = \{v_1, v_2, \dots, v_n\}$ 为节点的集合, E 为边的集合, $\langle v_i, v_j \rangle$ ($i, j = 1, 2, \dots, n, i \neq j$) 表示节点 v_i 与节点 v_j 之间存在由 v_i 到 v_j 的依赖关系
 - θ 表示各节点参数的集合, 包括每个节点所对应的条件概率参数离散情形下构成CPT

贝叶斯网的基本概念 (2)

简单的贝叶斯网实例

变量含义：

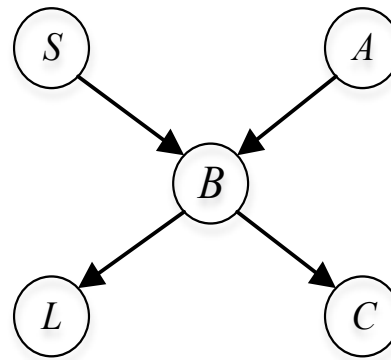
- S ：“吸烟”
- A ：“发烧”
- B ：“呼吸困难”
- L ：“肺癌”
- C ：“感染COVID-19”

所有变量为二值变量
(取值为T和F)

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

贝叶斯网的基本概念 (3)

◆ 贝叶斯网的构建方法

手工
构建

- 模型可解释性高
- 需要领域专家知识
工作量大

通过数据分
析构建BN

通过算法学习与数据尽可能吻合的模型

- 模型构建工作量小
- 模型构建速度相对快
- 得到的结构可能需要专家进一步优化

◆ 贝叶斯网的应用



推荐系统



故障诊断



机器学习



因果推断

提纲

- ◆ 引例
- ◆ 贝叶斯网概念
- ◆ 贝叶斯网参数学习
- ◆ 贝叶斯网结构学习
- ◆ 基于贝叶斯网的概率推理
- ◆ 总结

参数学习 (1)

◆ BN参数学习

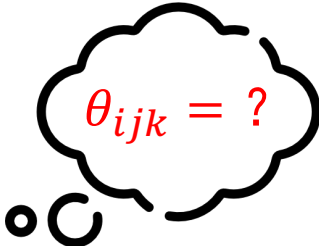
基于样本数据计算变量节点的条件概率参数

给定一个BN $\mathcal{B} = (G, \theta)$, 其中 $G = (V, E)$:

✓ 任意节点 $v_i \in V$ 取值为共有 r_i 个可能取值 $1, 2, \dots, r_i$, v_i 的父节点集记为 $\pi(v_i)$ 、共有 q_i 种可能组合 (若 v_i 无父节点, 则 $q_i = 1$)

✓ 若 v_i 取值为 k ($1 \leq k \leq r_i$), $\pi(v_i)$ 取第 j 种组合, 此时参数为

$$\theta_{ijk} = P(v_i = k | \pi(v_i) = j) (1 \leq i \leq n)$$


$$\theta_{ijk} = ?$$



如何有效从计算 θ_{ijk} , 进而计算 \mathcal{B} 中所有节点的所有参数, 得到参数集合 θ ?

参数学习 (2)

◆ BN的参数学习

- ✓ 给定一组关于 B 的独立同分布的完整样本数据集 $D = \{d_1, d_2, \dots, d_m\}$, θ 的某个取值 θ_0 与 D 的拟合程度用条件概率 $P(D|\theta = \theta_0)$ 度量, $P(D|\theta = \theta_0)$ 越大, 拟合程度越高。
- ✓ θ 的似然函数:

$$L(\theta|D) = P(D|\theta) = \prod_i^m P(d_i|\theta)$$

参数学习 (3)

◆ 最大似然估计 (Maximum Likelihood Estimation)

利用 D 中的样本数据，反推最具有可能（最大概率）导致这些样本结果出现的参数值，即求 θ 的某个取值 $\theta = \theta^*$ ，使 θ 的似然函数 $L(\theta | D)$ 值最大。

步骤：

(1) 对于任意样本 d_a ($1 \leq a \leq m$)，定义特征函数

$$\chi(i, j, k; d_a) = \begin{cases} 1, & \text{若 } d_a \text{ 中 } v_i = k \text{ 且 } \pi(v_i) = j \\ 0, & \text{其他} \end{cases}$$

参数学习 (4)

◆ 最大似然估计 (Maximum Likelihood Estimation)

(2) 对 $L(\theta|D)$ 取对数, 得到 θ 的对数似然函数:

$$\begin{aligned}l(\theta|D) &= \log P(D|\theta) = \log \prod_{a=1}^m P(d_a|\theta) = \sum_{a=1}^m \log P(d_a|\theta) \\ &= \sum_{a=1}^m \sum_{i=1}^n \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \chi(i, j, k; d_a) \log \theta_{ijk}\end{aligned}$$

其中, $P(d_a|\theta)$ 为给定 θ 时样本 d_a 出现的概率, 记为:

$$m_{ijk} = \sum_{a=1}^m \chi(i, j, k; d_a)$$

m_{ijk} 称为充分统计量, 直观上是数据集 D 中所有满足 $v_i = k$ 和 $\pi(v_i) = j$ 的样本数

参数学习 (5)

◆ 最大似然估计 (Maximum Likelihood Estimation)

(3) 将 θ 的对数似然函数化简为:

$$l(\theta|D) = \sum_{i=1}^n \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \log \theta_{ijk}$$

(4) θ_{ijk} 的最大似然估计:

$$\theta_{ijk}^* = \begin{cases} \frac{m_{ijk}}{\sum_{k=1}^{r_i} m_{ijk}}, & \text{若 } \sum_{k=1}^{r_i} m_{ijk} > 0 \\ \frac{1}{r_i}, & \text{若 } \sum_{k=1}^{r_i} m_{ijk} = 0 \end{cases}$$

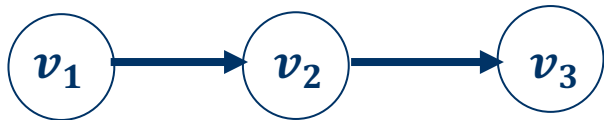
$$\text{其中, } \frac{m_{ijk}}{\sum_{k=1}^{r_i} m_{ijk}} =$$

D 中满足 $v_i=k$ 和 $\pi(v_i)=j$ 的样本实例数
—————
 D 中满足 $\pi(v_i)=j$ 的样本实例数

参数学习 (6)

◆ 基于最大似然估计的参数学习示例

给定 B 的结构 G



关于 B 的独立同分布数据 D

	v_1	v_2	v_3
d_1	F	F	F
d_2	T	T	T
d_3	T	T	F
d_4	F	F	F

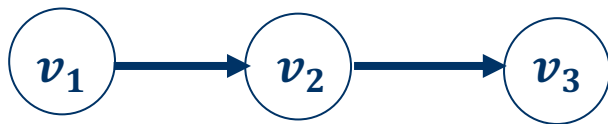
所有变量为二值变量（取值为T和F）

$$P(v_1 = T) = \frac{D \text{中满足 } v_1 = T \text{ 样本实例数}}{|D|} = \frac{2}{4}$$

$$P(v_2 = F) = \frac{D \text{中满足 } v_2 = F \text{ 样本实例数}}{|D|} = \frac{2}{4}$$

参数学习 (7)

◆ 基于最大似然估计的参数学习实例



v_2 只有一个父节点 v_1 ， $\pi(v_2)$ 共有
 $\pi(v_2) = T$ 和 $\pi(v_2) = F$ 两种组合，
分别记为**第1种**和**第2种**组合

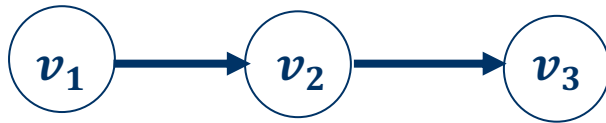
	v_1	v_2	v_3
d_1	F	F	F
d_2	T	T	T
d_3	T	T	F
d_4	F	F	F

$$\theta_{2T1}^* = \frac{D中满足v_2 = T和\pi(v_2) = T的样本实例数}{D中满足\pi(v_2) = T的样本实例数} = \frac{2}{2}$$

$$\theta_{2T2}^* = \frac{D中满足v_2 = T和\pi(v_2) = F的样本实例数}{D中满足\pi(v_2) = F的样本实例数} = 0$$

参数学习 (8)

◆ 基于最大似然估计的参数学习实例



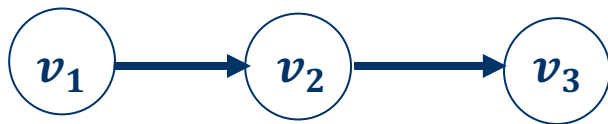
	v_1	v_2	v_3
d_1	F	F	F
d_2	T	T	T
d_3	T	T	F
d_4	F	F	F

$$\theta_{2F1}^* = \frac{D \text{中满足 } v_2 = F \text{ 和 } \pi(v_2) = T \text{ 的样本实例数}}{D \text{中满足 } \pi(v_2) = T \text{ 的样本实例数}} = 0$$

$$\theta_{2F2}^* = \frac{D \text{中满足 } v_2 = F \text{ 和 } \pi(v_2) = F \text{ 的样本实例数}}{D \text{中满足 } \pi(v_2) = F \text{ 的样本实例数}} = \frac{2}{2}$$

参数学习 (9)

◆ 基于最大似然估计的参数学习实例



计算所有节点的CPT:

	v_1	v_2	v_3
d_1	F	F	F
d_2	T	T	T
d_3	T	T	F
d_4	F	F	F

$P(v_1)$

v_1	T	F
$P(v_1)$	2/4	2/4

$P(v_2|v_1)$

$v_1 \backslash v_2$	T	F
T	2/2	0
F	0	2/2

$P(v_3|v_2)$

$v_2 \backslash v_3$	T	F
T	1/2	1/2
F	0	2/2

提纲

- ◆ 引例
- ◆ 贝叶斯网概念
- ◆ 贝叶斯网参数学习
- ◆ 贝叶斯网结构学习
- ◆ 基于贝叶斯网的概率推理
- ◆ 总结

结构学习 (1)

◆ BN的结构学习

在给定数据集的前提下寻找一个与训练样本集**匹配最好的网络结构**

结构学习主要包括以下两类方法：



结构学习 (2)

◆ 基于BIC评分和爬山法的BN结构学习

BIC评分：在大样本前提下对边缘似然函数的一种近似

$$\text{BIC}(B|D) = \sum_{i=1}^n \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \log \frac{m_{ijk}}{m_{ij*}} - \sum_{i=1}^n \frac{q_i(r_i - 1)}{2} \log m$$

- ✓ 第一项是模型结构 G 的**优参对数似然度**（Parameter Maximized Loglikelihood），度量模型**结构 G 与数据集 D 的拟合程度**。
- ✓ 若仅基于第一项选择模型，会得到一个任意两个节点之间都存在一条边的BN。

因此，增加第二项作为**惩罚项（Penalty）**，**防止模型过拟合**。

结构学习(3)

◆ 基于BIC评分和爬山法的BN结构学习

BIC评分的分解：用于减小搜索过程中的计算开销

➤ 给定BN中任意节点 v_i ， v_i 的家族（Family）为 v_i 与其父节点集 $\pi(v_i)$ 及相关边构成的局部结构

➤ v_i 的家族BIC评分：

$$\text{BIC}(\langle v_i, \pi(v_i) \rangle | D) = \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} m_{ijk} \log \frac{m_{ijk}}{m_{ij*}} - \sum_{i=1}^n \frac{q_i(r_i - 1)}{2} \log m$$

则有

$$\text{BIC}(\mathcal{B} | D) = \sum_{i=1}^n \text{BIC}(\langle v_i, \pi(v_i) \rangle | D)$$

结构学习(4)

◆ 基于BIC评分和爬山法的BN结构学习

➤ 基于爬山法找到BIC评分最高的模型

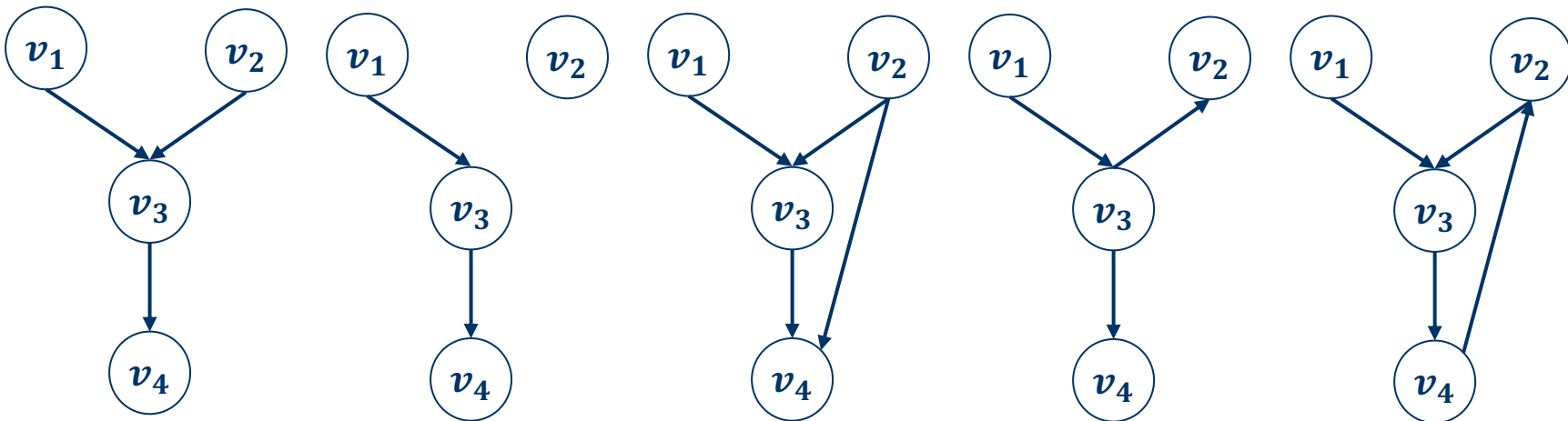
步骤:

- (1) 初始结构为无边模型，或基于领域知识设置初始结构
- (2) 通过**加边**、**减边**、**反转边**三种算子对当前结构局部进行修改，得到一系列候选模型
- (3) 计算不同候选模型参数的最大似然估计及相应的BIC评分
- (4) 迭代选出当前**BIC评分最高的候选结构**，直至收敛

结构学习(5)

◆ 基于BIC评分和爬山法的BN结构学习

三种算子:



初始结构

减边 $v_2 \rightarrow v_3$

加边 $v_2 \rightarrow v_4$

反转边 $v_3 \rightarrow v_2$

加边 $v_4 \rightarrow v_2$,
导致环, 不允许

结构学习(6)

◆ 基于爬山法的BN结构学习算法

V : 随机变量集合, D : 关于 V 的完整数据, f : BIC评分函数, G_0 : 初始BN结构
 $G \leftarrow G_0, \theta \leftarrow L(\theta | D), oldScore \leftarrow f(G, \theta | D)$

While true Do

$G^* \leftarrow \emptyset; \theta^* \leftarrow \emptyset; newScore \leftarrow -\infty$

For 每一个 G 中无边相连的节点对Do

进行加边、减边和反转边操作, 得到结构 G'

$\theta' \leftarrow L(\theta' | D); tmpScore \leftarrow f(G', \theta' | D)$

If $tmpScore > newScore$ Then

$G \leftarrow G^*; \theta \leftarrow \theta^*; oldScore \leftarrow newScore$

End for

If $newScore > oldScore$ Then

$G \leftarrow G^*; \theta \leftarrow \theta^*; oldScore \leftarrow newScore$

Else Return (G, θ)

End while

时间复杂度:

$O(|V|^2 \times |D|)$

t 为迭代次数

提纲

- ◆ 引例
- ◆ 贝叶斯网概念
- ◆ 贝叶斯网参数学习
- ◆ 贝叶斯网结构学习
- ◆ 基于贝叶斯网的概率推理
- ◆ 总结

基于贝叶斯网的概率推理 (1)

◆ 基于BN的概率推理算法

E : 证据变量集合; Q : 查询变量集合

➤ 精确推理算法

已知 E 取值为 e 的条件下, 利用联合概率和边缘概率来计算查询变量 Q 取值为 q 的后验概率, 得到精确的 $P(Q=q|E=e)$

➤ 近似推理算法

通过降低对精度的要求, 在限定时间内得到一个近似解

- 重要性采样 (Importance Sampling)
- 马尔科夫链蒙特卡罗 (Markov Chain Monte Carlo, MCMC)

基于贝叶斯网的概率推理 (2)

◆ 基于BN的精确推理算法

$L=T$ 作为证据、 $S=T$ 作为查询，计算条件概率 $P(S = T|L = T)$

➤ 一般联合概率分布推理

步骤:

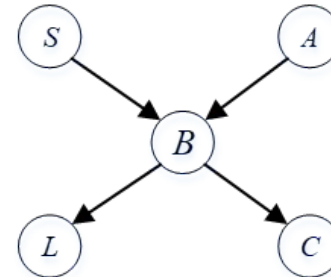
- (1) 从联合概率 $P(S, A, B, L, C)$ 出发
- (2) 计算边缘概率分布 $P(S, L) = \sum_{A, B, C} P(S, A, B, L, C)$

$$(3) \text{ 计算 } P(S = T|L = T) = \frac{P(S=T, L=T)}{P(L=T)}$$

整个联合概率分布包含 2^5-1 个独立参数，方法具有极高的复杂度

S	P(S)
T	0.6
F	0.4

A	P(A)
T	0.6
F	0.4



B	S	A	P(B S, A)
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95

L	B	P(L B)
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	P(C B)
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

基于贝叶斯网的概率推理 (3)

- 利用变量间的条件独立性，分解联合概率分布

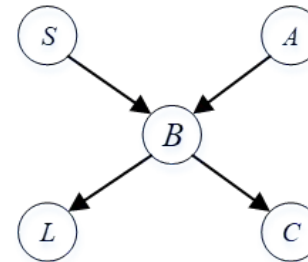
采用链式规则： $P(L) = \sum_S \sum_A \sum_B \sum_C P(S)P(A)P(B|S,A)P(L|B)P(C|B)$

计算步骤	乘法 (次)	加法 (次)
$P(S)P(A) \rightarrow P(S,A)$	4	/
$P(S,A)P(B S,A) \rightarrow P(B,S,A)$	8	/
$P(B,S,A)P(L B) \rightarrow P(L,B,S,A)$	16	/
$P(L,B,S,A)P(C B) \rightarrow P(C,L,B,S,A)$	32	/
$P(C,L,B,S,A) \rightarrow P(C,L,B,E)$	/	16
$P(C,L,B,E) \rightarrow P(L,B,E)$	/	8
$P(L,B,E) \rightarrow P(L,E)$	/	4
$P(L,E) \rightarrow P(L)$	/	2
总计	60	30

S	P(S)
T	0.6
F	0.4

A	P(A)
T	0.6
F	0.4

B	S	A	P(B S,A)
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	P(L B)
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	P(C B)
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

需要60次乘法和30次加法

基于贝叶斯网的概率推理 (4)

- 利用变量间的条件独立性，分解联合概率分布

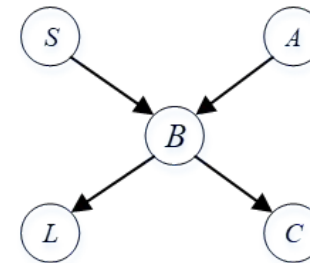
分解链式规则： $P(L) = \sum_B P(L|B) \sum_C P(C|B) \sum_S P(S) \sum_A P(A) P(B|S, A)$

计算步骤	数字乘法 (次)	数字加法 (次)
$P(A)P(B S, A) \rightarrow P(B, A S)$	8	/
$P(B, A S) \rightarrow P(B S)$	/	4
$P(S)P(B S) \rightarrow P(B, S)$	4	/
$P(B, S) \rightarrow P(B)$	/	2
$P(C B)P(B) \rightarrow P(C, B)$	4	/
$P(C, B) \rightarrow P(B)$	/	2
$P(L B)P(B) \rightarrow P(L, B)$	4	/
$P(L, B) \rightarrow P(L)$	/	2
总计	20	10

S	P(S)
T	0.6
F	0.4

A	P(A)
T	0.6
F	0.4

B	S	A	P(B S, A)
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	P(L B)
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	P(C B)
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

仅需要20次乘法和10次加法

基于贝叶斯网的概率推理 (5)

◆ VE算法

- 变量消元法 (Variable Elimination, VE) : 通过分解联合分布简化推理
步骤:

(1) 设 $N(v_1, v_2, \dots, v_n)$ 为 $\{v_1, v_2, \dots, v_n\}$ 的函数, 用 $\mathcal{F} = \{f_1, f_1, \dots, f_b\}$ 表示一组函数, 其中, 每个 f_i ($1 \leq i \leq b$) 涉及变量为 $\{v_1, v_2, \dots, v_n\}$ 的一个子集。如果

$$\mathcal{F} = \prod_{i=1}^b f_i$$

则称 \mathcal{F} 是 N 的一个分解 (Factorization), f_1, f_1, \dots, f_b 称为这个分解的因子 (Factor)。

基于贝叶斯网的概率推理 (6)

◆ VE算法

(2) 消元 (Elimination) :

$$N(v_1, v_2, \dots, v_n) \xrightarrow{K(v_2, \dots, v_y) = \sum_{v_1} \mathcal{F}(v_1, v_2, \dots, v_y)} \text{得到变量}\{v_2, \dots, v_n\} \text{的一个函数}$$

设 $\mathcal{F} = \{f_1, f_1, \dots, f_b\}$ 为函数 $N(v_1, v_2, \dots, v_n)$ 的一个分解, 从 N 中消去变量 v_1 的过程:

- 1) 从 \mathcal{F} 中删去所有 v_1 涉及的函数 (设这些函数为 $\{f_1, f_1, \dots, f_k\}$) ;
- 2) 将新函数 $\sum_{v_1} \prod_{i=1}^k f_i$ 放回 \mathcal{F} 中。

基于贝叶斯网的概率推理 (7)

◆ VE算法

B : BN; E : 证据变量; e : 证据变量取值, Q : 查询变量; ρ :
待消元变量顺序, 包括所有不在 $E \cup Q$ 中的变量

```
 $\mathcal{F} \leftarrow \mathcal{N}(v_1, v_2, \dots, v_n)$  //得到 $B$ 中所有变量条件概率分布的函数  
在 $\mathcal{F}$ 的因子中, 将证据变量 $E$ 设置为其观测值 $e$   
While  $\rho \neq \emptyset$  Do  
     $\rho \leftarrow \rho \setminus \{Z\}$  // $Z$ 为 $\rho$ 中第一个变量, 将 $Z$ 从 $\rho$ 中删除  
     $\mathcal{F} \leftarrow \text{Elim}(\mathcal{F}, Z)$  //对变量 $Z$ 进行消元  
End While  
 $h(Q) \leftarrow \prod_{i=1}^{|\mathcal{F}|} f_i$  //将 $\mathcal{F}$ 中所有因子相乘, 得到 $Q$ 的函数 $h(Q)$   
Return  $h(Q) / \sum_Q h(Q)$ 
```

```
Elim( $\mathcal{F}, Z$ )  
 $\mathcal{F} \leftarrow \mathcal{F} \setminus \{f_1, \dots, f_k\}$   
//从 $\mathcal{F}$ 中删去所有涉及 $Z$   
的函数 $\{f_1, \dots, f_k\}$   
 $g \leftarrow \prod_{i=1}^k f_i$   
 $h \leftarrow \sum_Z g$   
 $\mathcal{F} \leftarrow \mathcal{F} \cup \{h\}$  //将 $h$ 放回 $\mathcal{F}$   
Return  $\mathcal{F}$ 
```

假设 B 中每个变量有 x 种取值, ρ 为待消元变量个数, 最坏情况下进行 $x \times \prod_{i=1}^{\rho} x$ 次运算; 假设 B 中一共有 n 个变量, 则算法时间复杂度为 $O(n \times x^{\rho+1})$ 。

基于贝叶斯网的概率推理(8)

◆ VE算法示例

基于VE算法计算 $P(S|L=F)$

(1) 设置变量消元顺序 $\rho = \langle A, B, C \rangle$, 联合概率分解为 $\mathcal{F} = \{P(S), P(A), P(B|S, A), P(L|B), P(C|B)\}$;

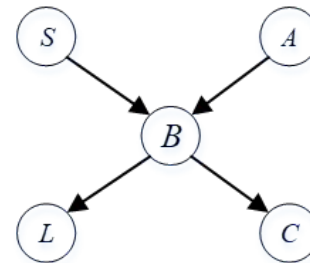
(2) 设置证据 $L=F$, 得到 $\mathcal{F} = \{P(S), P(A), P(B|S, A), P(L=F|B), P(C|B)\}$;

(3) 消去 A : A 为第一个消元变量, 与之相关的函数是 $P(A)$ 和 $P(B|S, A)$, 消去 A 得到 $\mathcal{F} = \{P(S), P(L=F|B), P(C|B), \varphi_1(B, S)\}$, 其中, $\varphi_1(B, S) = \sum_A P(A)P(B|S, A)$ 。

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

基于贝叶斯网的概率推理(9)

◆ VE算法示例

(4) 消去 B : 与之相关的函数是

$P(L=F|B)$ 、 $P(C|B)$ 和 $\varphi_1(B, S)$, 消去 B 得到

$\mathcal{F}=\{P(S), \varphi_2(S, C)\}$, 其中, $\varphi_2(S, C)=$

$\sum_B P(L = F|B)P(C|B) \varphi_1(B, S)$;

(5) 消去 C : 与之相关的函数是 $\varphi_2(S, C)$,

消去 C 得到 $\mathcal{F}=\{P(S), \varphi_3(S)$, 其中, $\varphi_3(S)$

$= \sum_C \varphi_2(S, C)$;

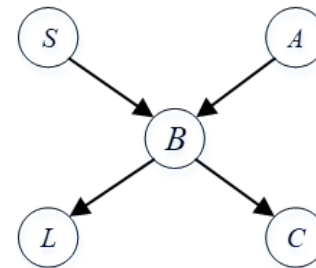
(6) 计算 $h(S)=\varphi_3(S)$;

(7) 返回 $h(S)/\sum_S h(S)$ 。

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

基于贝叶斯网的概率推理(10)

◆ 近似推理算法

Gibbs采样: 随机产生一个与证据 $E=e$ 一致的样本 D_1 作为初始样本, 每一步都**从当前样本出发**产生下一个样本。

步骤:

- (1) 对于当前第 $i - 1$ 步, 设 $D_{i-1} = D_i$;
 - (2) 按某个顺序对 D_{i-1} 中非证据变量逐个采样;
 - (3) 设 Z 是下一个待采样变量, 根据分布 $P(Z|mb(Z) = z_i)$ 对 Z 采样;
 - (4) 用采样结果替代 D_i 中 Z 的当前取值。
- $mb(Z)$ 是 Z 的**马尔科夫覆盖** (包括 Z 的直接孩子节点、直接父亲节点、以及直接孩子节点的其他父亲节点) 上的变量集合
 - z_i 是 $mb(Z)$ 在 D_i 中的当前取值

基于贝叶斯网的概率推理 (11)

B : BN, η : 采样次数, E_i : 证据变量, e : 证据变量取值, Q : 查询变量, q : 查询变量取值, ρ : 非证据变量采样顺序

随机生成一个样本 D_1 , 使 $E_i=e$

If $Q=q$ **Then** $m_q \leftarrow m_q + 1$

For $i = 2$ **To** η **Do**

$D_i \leftarrow D_{i-1}$

For ρ 中每一个变量 Z **Do**

$y \leftarrow \sum_i P(z_i | mb(Z))$ // 计算 Z 的下一个状态, z_i 为 Z 的不同状态取值

生成一个随机数 $r \in [0, y]$, Z 的取值为

$$Z = \begin{cases} z_1, & r \leq P(z_1 | mb(Z)) \\ z_2, & P(z_1 | mb(Z)) < r \leq P(z_1 | mb(Z)) + P(z_2 | mb(Z)) \\ \dots & \dots \end{cases}$$

$D_i \leftarrow \text{replace}(D_i, Z)$

If $Q=q$ **Then** $m_q \leftarrow m_q + 1$

Return m_q / η

假设每个变量的取值状态有 $|z|$ 种, 其父变量共有 c 种组合, 则算法的时间复杂度为 $O(\eta \times \rho \times |z|^c)$

基于贝叶斯网的概率推理 (12)

◆ 基于Gibbs采样的概率推理示例

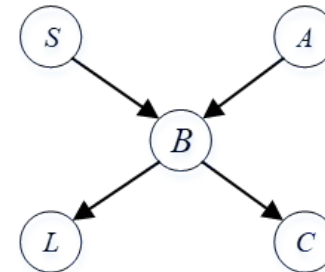
基于Gibbs采样算法近似计算 $P(S|L=F)$

- 随机生成一个与证据 $L=F$ 一致的样本，假设为 $D_1=\{S=T, A=F, B=T, L=F, C=F\}$
- 由 D_1 生成样本 D_2
- 算法从 $D_2=D_1=\{S=T, A=F, B=T, L=F\}$ 出发，对非证据变量逐个采样
- 采样顺序为 $\langle S, A, B, L, C \rangle$

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

基于贝叶斯网的概率推理 (13)

◆ 基于Gibbs采样的概率推理示例

采样过程如下：

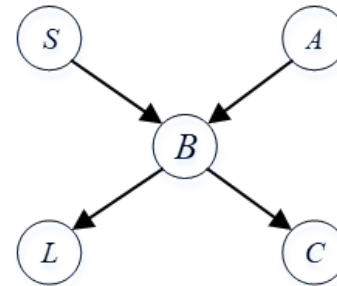
① 对 S 进行采样， $mb(S)$ 包含节点 A 和 B ，计算 S 的概率分布 $P(S|A=F, B=T)$ ，假设采样结果为 $S=F$ ，则有 $D_2=\{S=F, A=F, B=T, L=F, C=F\}$

② 对 A 进行采样，此时 $S=F$ ， $mb(A)$ 包含节点 S 和 B ，计算 A 的概率分布 $P(A|S=F, B=T)$ 。假设采样结果为 $A=T$ ，则有 $D_2=\{S=F, A=T, B=T, L=F, C=F\}$

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

基于贝叶斯网的概率推理 (14)

◆ 基于Gibbs采样的概率推理示例

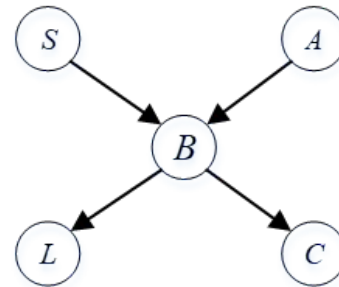
③ 对 B 进行采样, 此时 $A=T$, $mb(B)$ 包含节点 S 、 A 、 L 和 C , 因此, 计算 B 的概率分布 $P(B|S=F, A=T, L=F, C=F)$, 假设采样结果为 $B=T$, 则有 $D_2=\{S=F, A=T, B=T, L=F, C=F\}$

④ 对 L 进行采样, 此时 $B=T$, $mb(L)$ 包含节点 B 和 C , 计算 L 的概率分布 $P(L|B=T, C=F)$, 假设采样结果为 $L=T$, 则有 $D_2=\{S=F, A=T, B=T, L=T, C=F\}$

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

基于贝叶斯网的概率推理 (15)

◆ 基于Gibbs采样的概率推理示例

⑤ 对 C 进行采样, 此时 $L=T$, $mb(C)$ 包含节点 B 和 L , 计算 C 的概率分布 $P(C|B=T, L=T)$, 假设采样结果为 $C=F$, 则有 $D_2=\{S=F, A=T, B=y, L=T, C=F\}$, 即为 D_2 的最终值。

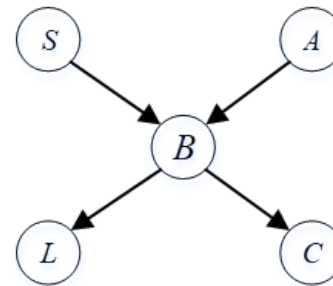
假设采样共得到 η 个样本, 其中满足 $Q=q$ 的有 m_q 个, 近似的后验概率为

$$P(Q=q) \approx \frac{m_q}{\eta}$$

S	$P(S)$
T	0.6
F	0.4

A	$P(A)$
T	0.6
F	0.4

B	S	A	$P(B S, A)$
T	T	T	0.9
F	T	T	0.1
T	T	F	0.83
F	T	F	0.17
T	F	T	0.2
F	F	T	0.8
T	F	F	0.05
F	F	F	0.95



L	B	$P(L B)$
T	T	0.8
F	T	0.2
T	F	0.1
F	F	0.9

C	B	$P(C B)$
T	T	0.6
F	T	0.4
T	F	0.01
F	F	0.99

提纲

- ◆ 引例
- ◆ 贝叶斯网概念
- ◆ 贝叶斯网参数学习
- ◆ 贝叶斯网结构学习
- ◆ 贝叶斯网概率推理
- ◆ 总结

总结

- ◆ 不确定性知识表示和处理能力，是智能系统走向实用的重要要求
 - ✓ 概率图模型：图模型的概率性质，概率论+图论
 - ✓ 概率推理基本思想和分类
- ◆ 贝叶斯网的概念、参数学习和结构学习
- ◆ 基于贝叶斯网的概率推理（精确推理，近似推理）



结语

谢谢！