人工智能 I

杨杰

yangjie@seu.edu.cn

2025/12/01，五四楼-303，9:50~11:25

本周四实验课安排

周次	11	
日期	2025/12/04	
星期	星期四	
时间	16:40-18:15	19:00-20:35
机房地点	中心楼实验室 D区	五五楼435 软实七
实验任务	贝叶斯分类器	

注意两个时间段是不同机房！请勿跑错！
注意18:15机房电脑自动关机！请及时保存！

本节课安排

- 监督学习：
  - 线性回归
  - 对率回归
  - 支持向量机
  - 决策树
  - 随机森林
  - 贝叶斯分类器
  - 感知机与神经网络

贝叶斯决策论(Bayesian decision theory)

- 贝叶斯决策论（Bayesian decision theory）是**在概率框架下实施决策的基本方法**。
  - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策**考虑如何基于这些概率和误判损失来选择最优的类别标记**。

- 假设有N种可能的标记：$y = \{c_1, c_2, \ldots, c_N\}$

- 令 λ_{ij} 代表将第j类样本误分类为第i类所产生的损失，则基于后验概率将样本$\mathbf{x}$分到第i类的条件风险为：

$$R(c_i|\mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\mathbf{x})$$

贝叶斯决策论(Bayesian decision theory)

- 我们的任务是寻找一个判定准则 $h: X \mapsto Y$ 以最小化总体风险

$$R(h) = \mathbf{E}_x \left[R(h(\mathbf{x}) \mid \mathbf{x}) \right] \qquad (7.2)$$

- 显然，对每个样本 $\mathbf{x}$，若 h 能最小化条件风险 $R(h(\mathbf{x}) \mid \mathbf{x})$ 则总体风险也将被最小化。

- 这就产生了**贝叶斯判定准则**（Bayes decision rule）： 为最小化总体风险，只需在每个样本上选择那个能使条件风险 $R(c \mid \mathbf{x})$ 最小的类别标记，即

$$h^*(\mathbf{x}) = \underset{c \in \mathcal{Y}}{\arg\min}\, R(c|\mathbf{x})$$

✓ 此时，被称为**贝叶斯最优分类器**(Bayes optimal classifier)，与之对应的总体风险 $R(h^*)$ 称为**贝叶斯风险** (Bayes risk)

✓ $1 - R(h^*)$ 反映了分类器所能达到的最好性能，即通过机器学习所能产生的**模型精度的理论上限**。

贝叶斯决策论(Bayesian decision theory)

概率框架下实施决策的基本理论

- 给定N个类别，令 λ_{ij} 代表将第j类样本误分类为第i类所产生的损失，则基于后验概率将样本$\mathbf{x}$分到第i类的条件风险为：

$$R(c_i|\mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|\mathbf{x})$$

若$\lambda_{ij} = 1 - \delta(i-j)$
则$R(c|\mathbf{x}) = 1 - P(c|\mathbf{x})$

- 贝叶斯判定准则 (Bayes decision rule)：

$$h^*(\mathbf{x}) = \arg\min_{c\in\mathcal{Y}} R(c|\mathbf{x})$$

若$\lambda_{ij} = 1 - \delta(i-j)$
则$h^*(\mathbf{x}) = \arg\max_{c\in\mathcal{Y}} P(c|\mathbf{x})$

极大后验概率（Maximum a Posteriori/MAP）

✓ h^*称为 贝叶斯最优分类器 (Bayes optimal classifier)，其总体风险称为贝叶斯风险 (Bayes risk)

✓ 反映了学习性能的理论上限

贝叶斯决策论(Bayesian decision theory)

- 不难看出，使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c \mid \mathbf{x})$。

- 然而，在现实中通常难以直接获得。机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率 $P(c \mid \mathbf{x})$。

- 主要有两种策略：
  - 判别式模型（discriminative models）
    - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$，来预测 c
    - 决策树，BP神经网络，支持向量机
  - 生成式模型（generative models）
    - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
    - 生成式模型考虑 $$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \qquad (7.7)$$

贝叶斯法则

$$P(c|\mathbf{x}) = \frac{\boxed{P(c)}\,\boxed{P(\mathbf{x}|c)}}{\boxed{P(\mathbf{x})}}$$

先验概率（prior ）
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定律）

证据 (evidence)
因子，与类别无关

样本相对于类标记的 类条件概率 (class-conditional probability), 亦称 似然 (likelihood)

主要困难在于估计似然

朴素贝叶斯分类器 (Naive Bayes Classifier)

$$P(c|\mathbf{x}) = \frac{P(c)\boxed{P(\mathbf{x}|c)}}{P(\mathbf{x})}$$

表 1.1 西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

主要障碍：所有属性上的联合概率难以从有限训练样本估计获得

- 组合爆炸
- 样本稀疏

基本思路：**假定属性相互(条件)独立**

$$P(\mathbf{x}|c) = \Pi_{i=1}^{d} P(x_i|c)$$

d为属性数，x_i为$\mathbf{x}$在第i个属性上的取值

$P(x)$对所有类别相同，于是：

$$h_{nb}(\mathbf{x}) = \arg\max_{c \in \mathcal{Y}} P(c)\Pi_{i=1}^{d} P(x_i|c)$$

朴素贝叶斯分类器：算法

① 估计$P(c)$

$$P(c) = \frac{|D_c|}{|D|}$$

- D_c：类c的所有样本

② 估计$P(\mathbf{x}|c)$

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

- D_{c,x_i}：D_c中在第i个属性上取值为x_i的所有样本

朴素贝叶斯分类器：算法

表 4.3 西瓜数据集 3.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

拉普拉斯修正

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题. **比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然不合理。**

- 为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“拉普拉斯修正”（Laplacian correction）
  - 令 N 表示训练集 D 中可能的类别数，N_i 表示第 i 个属性可能的取值数，则式 (7. 16)和 (7. 17)分别修正为

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \quad (7.19) \qquad \hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i} \quad (7.20)$$

- 现实任务中，朴素贝叶斯分类器的使用：速度要求高，“查表”；任务数据更替频繁，“懒惰学习”(lazy learning)；数据不断增加，增量学习等等。

半朴素贝叶斯分类器

朴素贝叶斯分类器的“属性独立性假设”在现实中往往难以成立

半朴素贝叶斯分类器 (semi-naïve Bayes classifier)

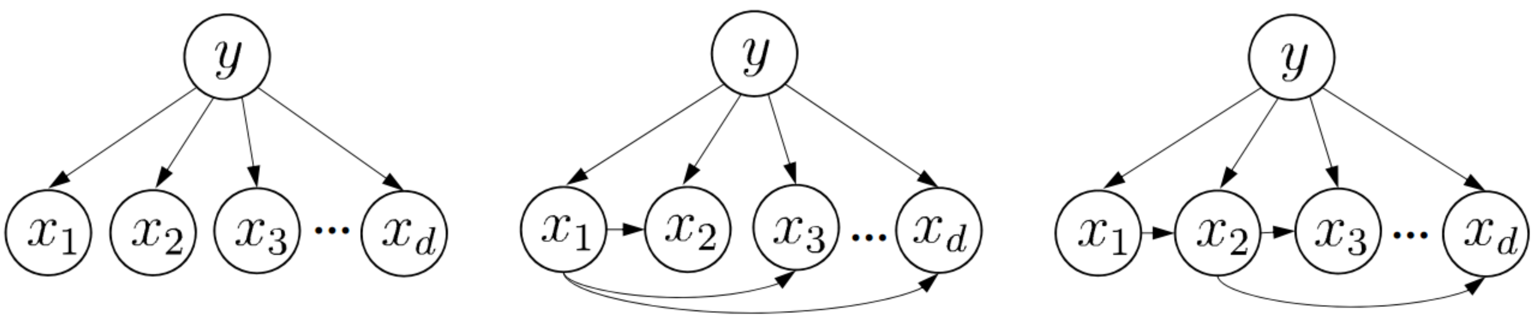
基本思路：适当考虑一部分属性间的相互依赖信息

最常用策略： **独依赖估计** (One-Dependent Estimator, ODE)

假设每个属性在类别之外最多仅依赖一个其他属性

$$P(c|x) \propto P(c)\Pi_{i=1}^{d} P(x_i|c, \boxed{pa_i})$$

x_i的“父属性”



(a) NB (b) SPODE (c) TAN

高阶依赖

能否通过考虑属性间的高阶依赖来进一步提升泛化性能?

例如最简单的做法:

将父属性pa_i替换为包含k个属性的集合$\mathbf{pa}_i$

明显障碍：随着k的增加，估计 $P(x_i|c,\mathbf{pa}_i)$ 所需的样本数将以指数级增加

- 训练样本非常充分 → 性能可能提升
- 有限训练样本 → 高阶联合概率估计困难

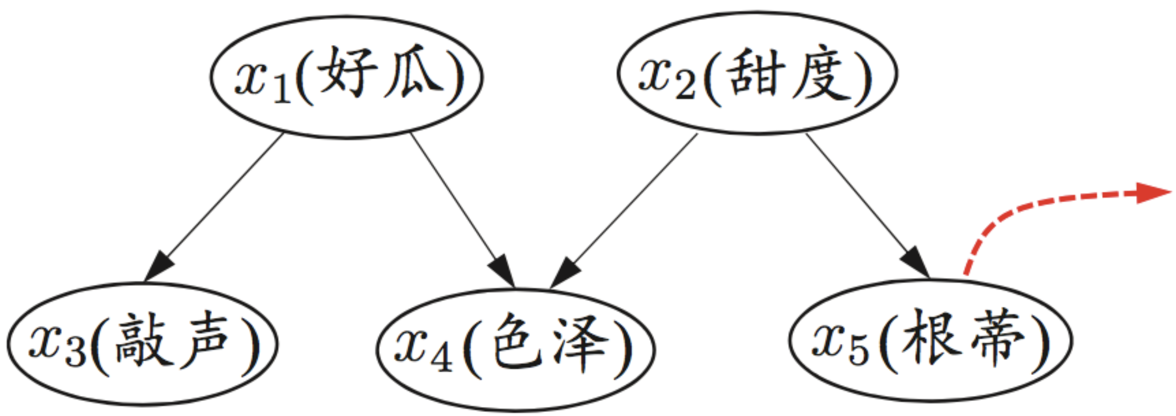
考虑属性间的高阶依赖，需要其他办法

贝叶斯网 （Bayesian network）

- **贝叶斯网 (Bayesian network)**亦称**“信念网”(brief network)**，它借助有向无环图 (Directed Acyclic Graph, DAG)来刻画属性间的依赖关系，并使用条件概率表(Conditional Probability Table, CPT)来表述属性的联合概率分布。

贝叶斯网 （Bayesian network）

- **贝叶斯网 (Bayesian network)**亦称“**信念网**”**(brief network)**，它借助**有向无环图** (Directed Acyclic Graph, DAG)来刻画属性间的依赖关系，并使用**条件概率表** (Conditional Probability Table, CPT)来表述属性的联合概率分布。



根蒂

		硬挺	蜷缩
甜度	高	0.1	0.9
	低	0.7	0.3

图7.2 西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

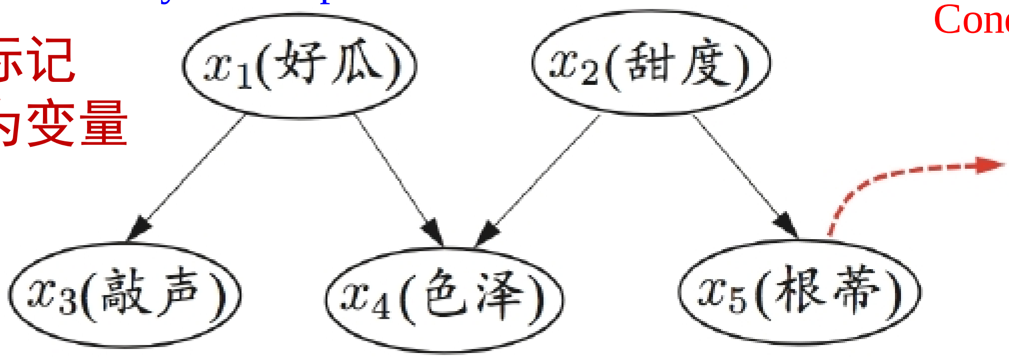
- 从网络图结构可以看出 -> “色泽”直接依赖于“好瓜”和“甜度”
- 从条件概率表可以得到 -> “根蒂”对“甜度”的量化依赖关系
  P(根蒂 = 硬挺|甜度 = 高) = 0.1

贝叶斯网 （Bayesian network）

亦称“信念网” （brief network）

有向无环图（ DAG, Directed Acyclic Graph）

类标记
作为变量



条件概率表 （ CPT, Conditional Probability Table）

		根蒂	
		硬挺	蜷缩
甜度	高	0.1	0.9
	低	0.7	0.3

贝叶斯网 $B = \langle G, \Theta \rangle$

结构 参数

1985年 J. Pearl 命名为贝叶斯网

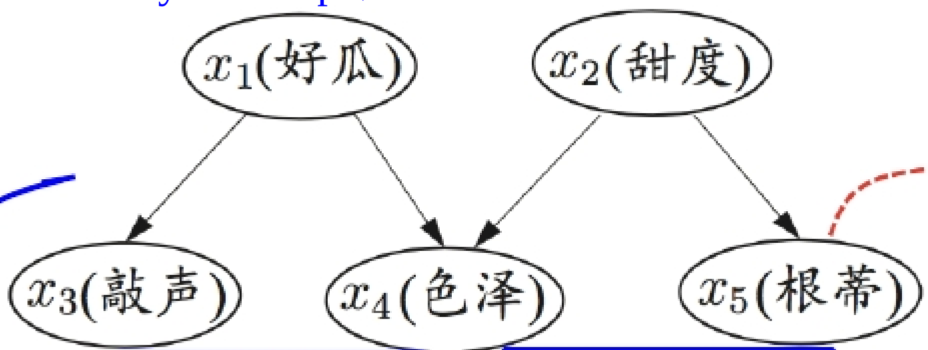
Judea Pearl
(1936 -)
2011 图灵奖

概率图模型 （Probabilistic graphical model）

- 有向图模型 → 贝叶斯网
- 无向图模型 → 马尔可夫网

贝叶斯网（Bayesian network）

有向无环图（ DAG, Directed Acyclic Graph）



$\boxed{x_3 \perp x_4 \mid x_1}$ $\boxed{x_4 \perp x_5 \mid x_2}$

条件概率表 （ CPT, Conditional Probability Table）

		根蒂	
		硬挺	蜷缩
甜度	高	0.1	0.9
	低	0.7	0.3

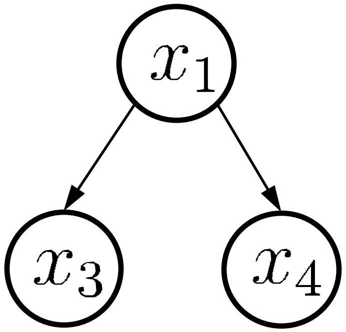
给定父节点集，贝叶斯网假设每个属性与其非后裔属性独立

$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \boxed{\pi_i}) = \prod_{i=1}^{d} \boxed{\theta_{x_i|\pi_i}}$$

父节点集

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

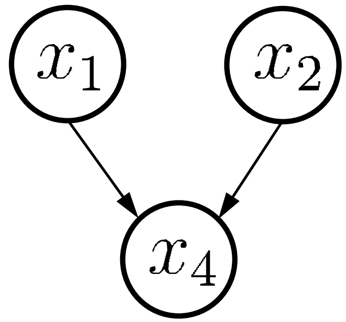
三变量间的典型依赖关系



同父结构

条件独立性

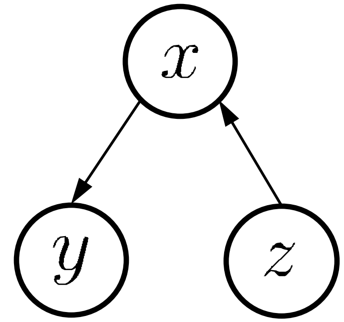
$x_3 \perp x_4 \mid x_1$



V型结构

边际独立性

$x_1 \perp\!\!\!\perp x_2$



顺序结构

条件独立性

$y \perp z \mid x$

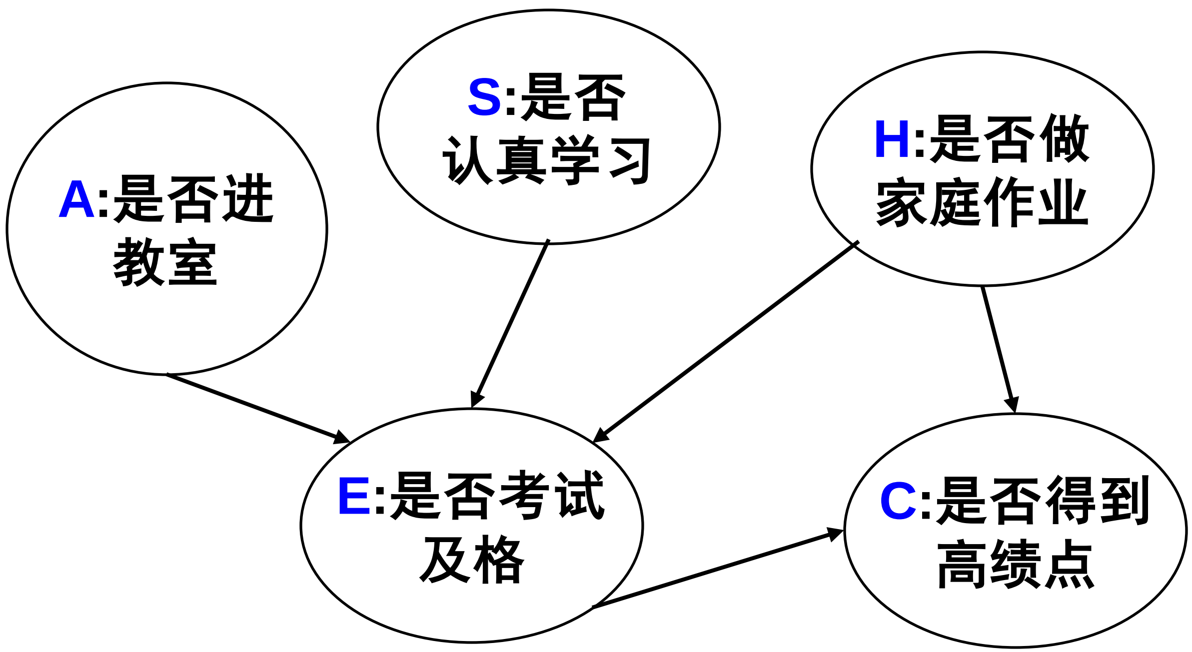
- 给定 x_4, x_1与 x_2必不独立
- 若x_4未知，则x_1与x_2独立

练习

五、（本题共 12 分）以下问题涉及贝叶斯网络相关知识，回答问题

1. （简答题，本小题共 2 分）

根据下图所示，写出联合概率表达式 $P(A,S,H,E,C)$



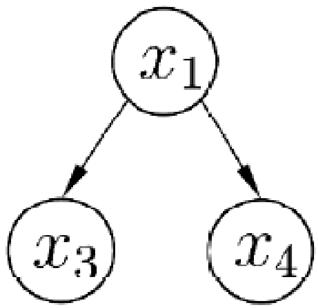
$$P(A,S,H,E,C) = P(A)P(S)P(H)P(E|A,S,H)P(C|E,H)$$

结构学习与推断

$$B = \langle G, \Theta \rangle$$

结构　　参数

变量间的依赖关系　　条件概率表



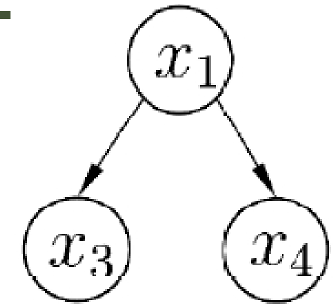
$$p(x_1)$$
$$p(x_3|x_1)$$
$$p(x_4|x_1)$$

- 结构学习：NP困难问题
- 推断(inference)：基于已知属性变量的观测值，
  推测其他属性变量的取值

已知属性变量的观测值称为“证据” (evidence)

推断

推断(inference)：基于已知属性变量的观测值，
推测其他属性变量的取值



已知属性变量的观测值称为“证据” (evidence) $P(x_1, x_3 | x_4)$

- **精确推断**：直接根据贝叶斯网定义的联合概率分布来精确计算后验概率

NP-hard

- **近似推断**：降低精度要求，在有限时间内求得近似解

注：对于不含缺失值的分类问题，精确推断的时间复杂为$O(dk)$

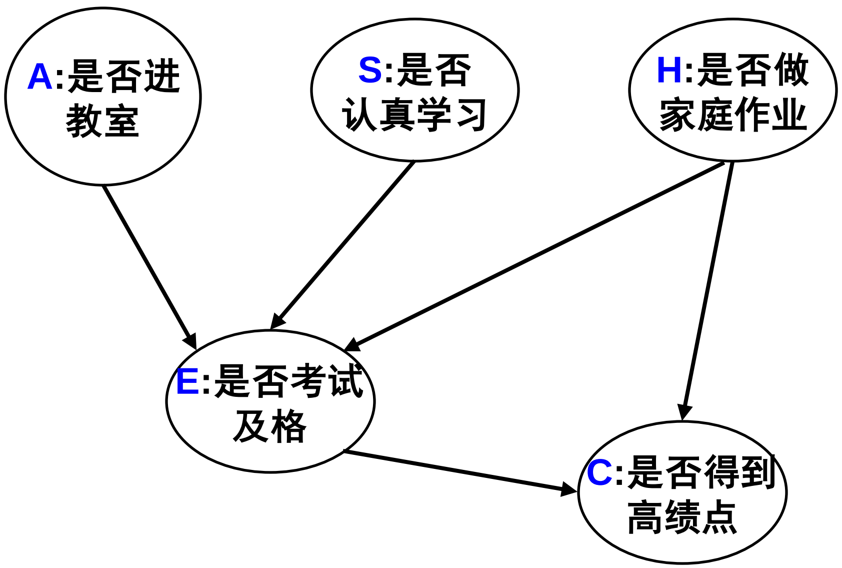
常见做法：

- 吉布斯采样 (Gibbs sampling)
- 变分推断 (variational inference)

练习

2. （问答题，本小题共 4 分）

根据上一小问，如果给定**进教室**上课、**认真学习**、但是**不做家庭作业**的情况下，请写出**得到高绩点**的概率表达式（要求写出详细推导步骤）。



$$P(C|A,S,H)=\sum_e P(E=e,C|A,S,H)=\sum_e P(C|A,E=e,S,H)P(E=e|A,S,H)=\sum_e P(C|E=e,H)P(E=e|A,S,H)$$

$$P(C=1|A=1,S=1,H=0)=\sum_e P(C=1|E=e,H=0)P(E=e|A=1,S=1,H=0)$$

练习

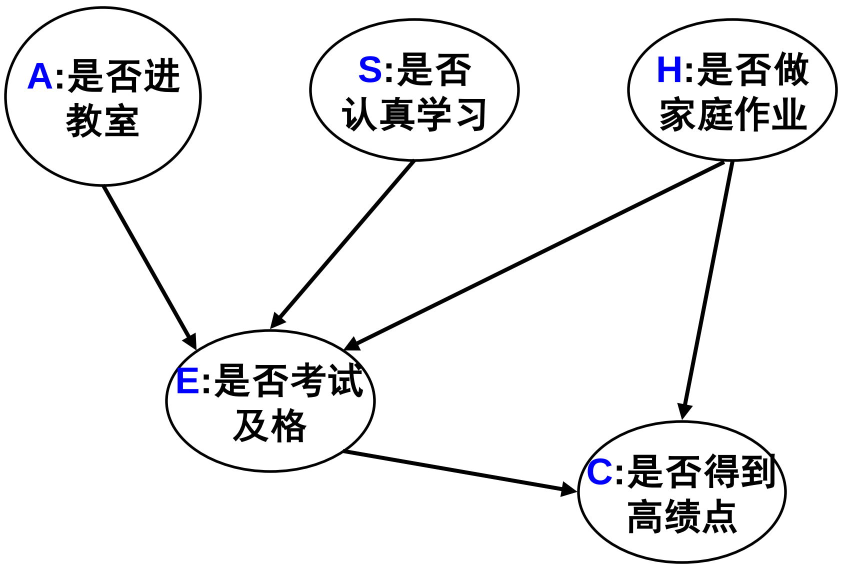
3. （简答题，本小题共 6 分）

根据第 1 小问的贝叶斯网络，已知如下条件概率表（CPT），请计算 $P(A=1|C=1,H=1)$

其中， $P(A=1)=0.5, P(S=1)=0.7, P(H=1)=0.9$

E=1

A	S	H	$P(E\|A,S,H)$
0	0	0	0.2
0	0	1	0.5
0	1	0	0.4
0	1	1	0.8
1	0	0	0.3
1	0	1	0.7
1	1	0	0.6
1	1	1	0.9



C=1

E	H	$P(C\|E,H)$
0	0	0.1
0	1	0.4
1	0	0.3
1	1	0.9

$$P(A|C,H)=\frac{P(A,C,H)}{P(C,H)}$$

$$=\frac{\sum_{s,e}P(A,S=s,H,E=e,C)}{\sum_{a,s,e}P(A=a,S=s,H,E=e,C)}$$

$$=\frac{\sum_{s,e}P(A)P(S=s)P(E=e|A,S=s,H)P(C|E=e,H)}{\sum_{a,s,e}P(A=a)P(S=s)P(E=e|A=a,S=s,H)P(C|E=e,H)}$$

练习

计算分子=0.41

P(A=1)=0.5, P(A=0)=0.5
P(S=1)=0.7, P(S=0)=0.3

P(C=1 | E=0,H=1)=0.4
P(C=1 | E=1,H=1)=0.9

P(E=1 | A=1,S=0,H=1) = 0.7
P(E=1 | A=1,S=1,H=1) = 0.9

P(E=0 | A=1,S=0,H=1) = 0.3
P(E=0 | A=1,S=1,H=1) = 0.1

$P(A=1)$ P(S=0) $P(E=0 \mid A=1,S=0,H=1)$ $P(C=1 \mid E=0,H=1)$
$= 0.5 \times 0.3 \times 0.3 \times 0.4 = 0.018$

+

$P(A=1)$ P(S=0) $P(E=1 \mid A=1,S=0,H=1)$ $P(C=1 \mid E=1,H=1)$
$= 0.5 \times 0.3 \times 0.7 \times 0.9 = 0.0945$

+

$P(A=1)$ P(S=1) $P(E=0 \mid A=1,S=1,H=1)$ $P(C=1 \mid E=0,H=1)$
$= 0.5 \times 0.7 \times 0.1 \times 0.4 = 0.014$

+

$P(A=1)$ P(S=1) $P(E=1 \mid A=1,S=1,H=1)$ $P(C=1 \mid E=1,H=1)$
$= 0.5 \times 0.7 \times 0.9 \times 0.9 = 0.2835$

练习

计算分母=0.41+0.3775

P(A=1)=0.5, P(A=0)=0.5
P(S=1)=0.7, P(S=0)=0.3

P(C=1 | E=0,H=1)=0.4
P(C=1 | E=1,H=1)=0.9

P(E=1 | A=1,S=0,H=1) = 0.7
P(E=0 | A=1,S=0,H=1) = 0.3

P(E=1 | A=1,S=1,H=1) = 0.9
P(E=0 | A=1,S=1,H=1) = 0.1

P(E=1 | A=0,S=0,H=1) = 0.5
P(E=0 | A=0,S=0,H=1) = 0.5

P(E=1 | A=0,S=1,H=1) = 0.8
P(E=0 | A=0,S=1,H=1) = 0.2

P(A=0)P(S=0)P(E=0 | A=0,S=0,H=1)P(C=1 | E=0,H=1)
= 0.5×0.3×0.5×0.4=0.03

+

P(A=0)P(S=0)P(E=1 | A=0,S=0,H=1) P(C=1 | E=1,H=1)
= 0.5×0.3×0.5×0.9=0.0675

+

P(A=0)P(S=1)P(E=0 | A=0,S=1,H=1)P(C=1 | E=0,H=1)
= 0.5×0.7×0.2×0.4=0.028

+

P(A=0)P(S=1)P(E=1 | A=0,S=1,H=1) P(C=1 | E=1,H=1)
= 0.5×0.7×0.8×0.9=0.252

$$P(A=1|C=1,H=1)=\frac{0.41}{0.7875}\approx 0.52$$

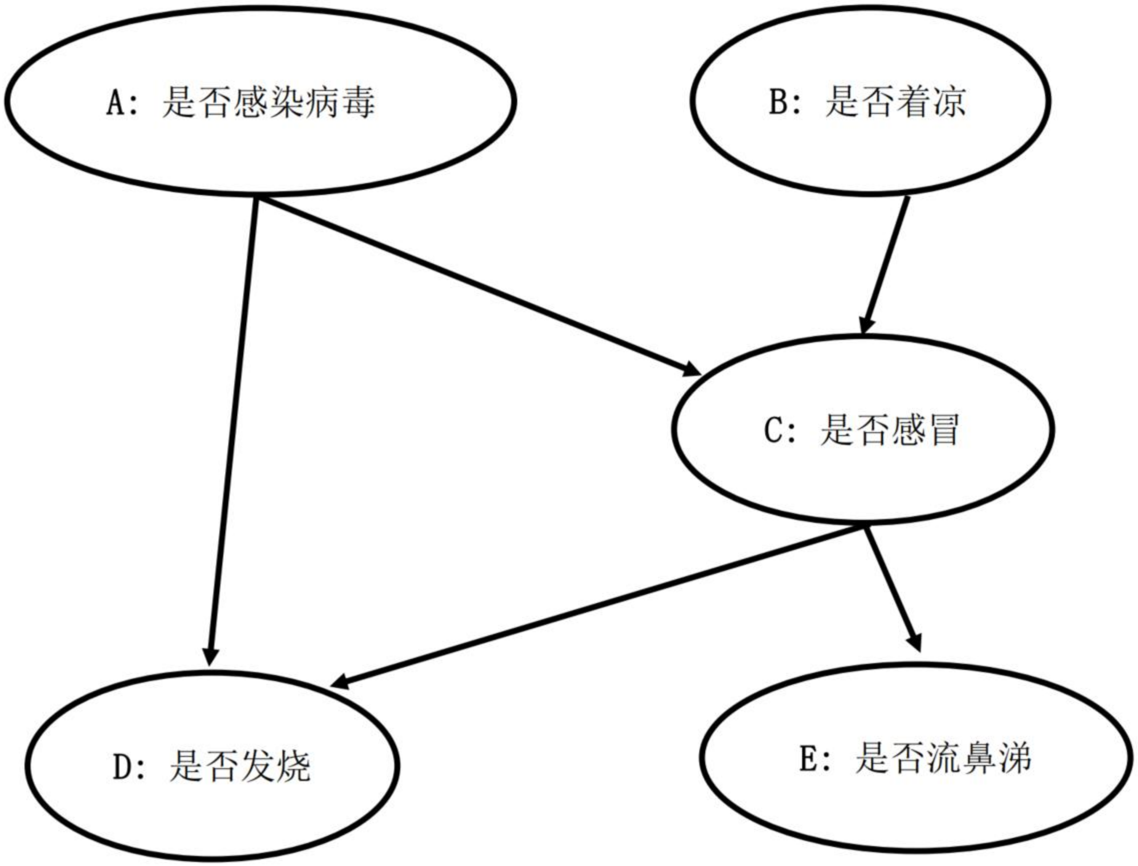
练习

四、（本题共 13 分）以下问题涉及贝叶斯网络相关知识，回答问题

1. （简答题，本小题共 3 分）

   根据下图所示，写出联合概率$P(A,B,C,D,E)$的表达式

$$P(A,B,C,D,E)=P(A)P(B)P(C|A,B)P(D|A,C)P(E|C)$$



练习

2.（问答题，本小题共 6 分）

根据上一小问，请写出在**着凉**$(B=1)$但是**没有感染病毒**$(A=0)$的情况下**发烧**$(D=1)$的概率表达式（要求写出详细推导步骤）。

$$P(D|A,B)=\sum_c P(C=c,D|A,B)$$
$$=\sum_c P(D|A,B,C=c)P(C=c|A,B)$$
$$=\sum_c P(D|A,C=c)P(C=c|A,B)$$

$$P(D=1|A=0,B=1)=\sum_c P(D=1|A=0,C=c)P(C=c|A=0,B=1)$$

练习

3. （简答题，本小题共 4 分）

已经如下条件概率表，其中$P(A=1)=0.1, P(B=1)=0.4$。请根据上一小问，计算**发烧**$(D=1)$且**流鼻涕**$(E=1)$的情况下**感染病毒**$(A=1)$的概率。（注：写成分数或精确到小数点后 2 位均可）

A	B	$P(C=1\|A,B)$
1	1	0.8
1	0	0.5
0	1	0.2
0	0	0.01

A	C	$P(D=1\|A,C)$
1	1	0.95
1	0	0.9
0	1	0.1
0	0	0.01

C	$P(E=1\|C)$
1	0.8
0	0.1

$$P(A|D,E)=\frac{P(A,D,E)}{P(D,E)}=\frac{\sum_{b,c}P(A,B=b,C=c,D,E)}{\sum_{a,b,c}P(A=a,B=b,C=c,D,E)}=\frac{\sum_{b,c}P(A)P(B=b)P(C=c|A,B=b)P(D|A,C=c)P(E|C=c)}{\sum_{a,b,c}P(A=a)P(B=b)P(C=c|A=a,B=b)P(D|A=a,C=c)P(E|C=c)}$$

$$P(A=1|D=1,E=1)=0.88$$