Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 23, 2015

Today:

- Graphical models
- Bayes Nets:
  - Representing distributions
  - Conditional independencies
  - Simple inference
  - Simple learning

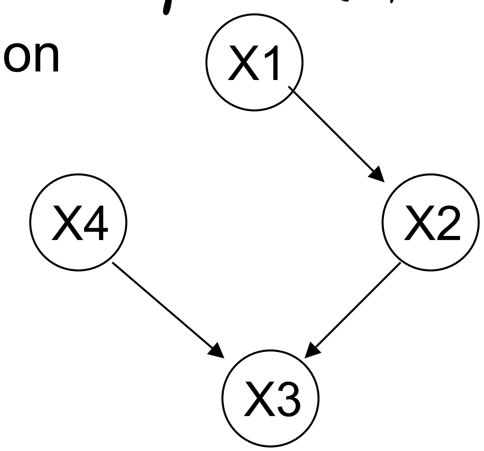
Readings:

- Bishop chapter 8, through 8.2
- Mitchell chapter 6

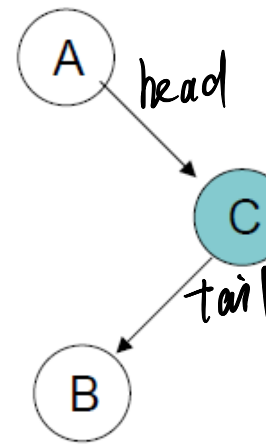
Conditional Independence, Revisited

- We said:
  与非子代节点、都独立
  - Each node is conditionally independent of its non-descendents, given its immediate parents.

- Does this rule give us all of the conditional independence relations implied by the Bayes network?
  - No!
  给定{X2, X3}，那么X1, X4条件独立
  - E.g., X1 and X4 are conditionally indep given {X2, X3}
  - But X1 and X4 not conditionally indep given X3 但只给定X3，X1，X4不条件独立
  - For this, we need to understand D-separation



Easy Network 1: Head to Tail



prove A cond indep of B given C?

ie., p(a,b|c) = p(a|c) p(b|c) ⇒ 条件独立.

$P(a|c)\cancel{P(c)}$

条件独立 $P(a,b|c) = \dfrac{P(a,b,c)}{P(c)} = \dfrac{\boxed{P(a)P(c|a)}P(b|c)}{\cancel{P(c)}} = P(a|c)P(b|c)$

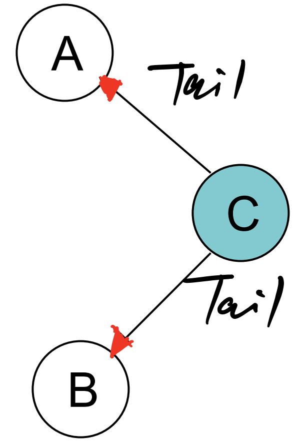
无法化简掉

边缘独立： $P(a,b) = \sum_c P(a,b,c) = \sum_c P(a)P(c|a)P(b|c) \neq P(a)P(b)$

⇒ 在给定C时条件独立 但不独立(边缘独立)

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 2: Tail to Tail



prove A cond indep of B given C?   ie., p(a,b|c) = p(a|c) p(b|c)

$$P(a,b|c) = \frac{P(c)P(a|c)P(b|c)}{P(c)}$$

$$= p(a|c)\,p(b|c)$$

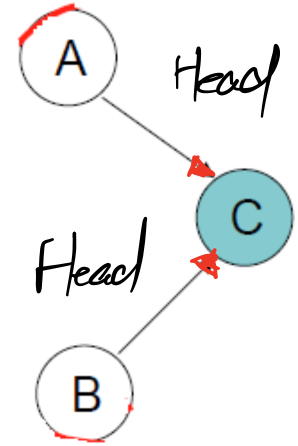
$$P(a,b) = \sum_c P(c)\underline{P(a|c)P(b|c)}$$ 无法化简

$\Rightarrow$ 也是给定C时条件独立 但本身不独立.

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 3: Head to Head

prove A cond indep of B given C?   ie., $p(a,b|c) = p(a|c)\, p(b|c)$



$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a)p(b)\underline{p(c|a,b)}}{p(c)}$$

→ 化简不掉

不独立

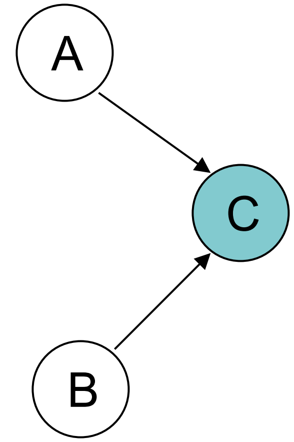
$$p(a,b) = \sum_c p(a,b,c) = \sum_c p(a)\,p(b)\,\underline{p(c|a,b)} = p(a)\,p(b)$$

→ 求和是1

⇒ 条件不独立. 边缘独立.

let's use p(a,b) as shorthand for p(A=a, B=b)

Easy Network 3: Head to Head



prove A cond indep of B given C?   NO!

Summary:

- p(a,b)=p(a)p(b)
- p(a,b|c) NotEqual p(a|c)p(b|c)

Explaining away.

e.g.,

- A=earthquake
- B=breakIn) 原本不相关
- C=motionAlarm ⟶ 一旦发生C，那么认为不是A就是B.

AB同时发生概率太小，不考虑，

故AB不再独立.

X and Y are conditionally independent given Z, **<u>if and only if</u>** X and Y are D-separated by Z.

[Bishop, 8.2.2]

Suppose we have three sets of random variables: X, Y and Z

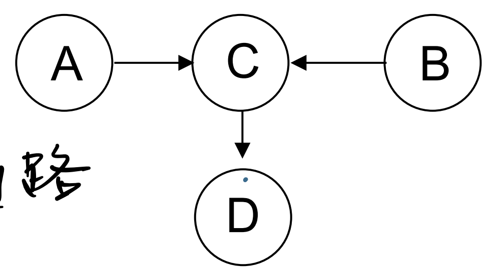
X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from every variable in X to every variable in Y is **<u>blocked</u>**

被观测的相当于被打断.

A path from variable X to variable Y is **blocked** if it includes a node in Z such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z



X与Y通过Z相连可能有多个通路

只有都满足才能说明独立.

X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from every variable in X to every variable in Y is **<u>blocked</u>**

A path from variable A to variable B is **blocked** if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

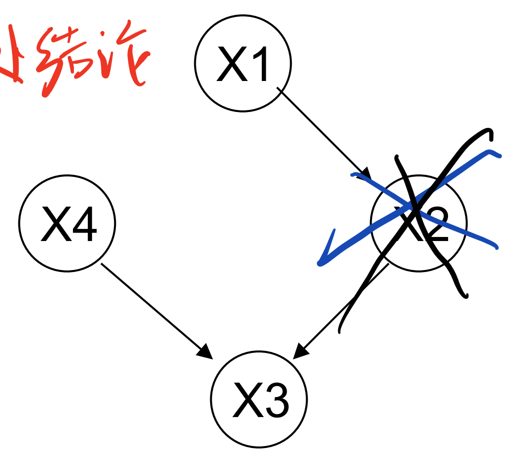
X1 indep of X3 given X2? ✓

X3 indep of X1 given X2? ✓

顺序不影响此处结论

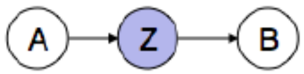
X4 indep of X1 given X2? ✓

只有一条路，且被打断.

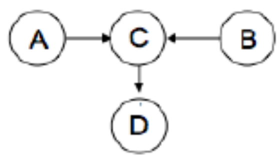


X and Y are **D-separated** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **blocked** by Z

A path from variable A to variable B is **blocked** by Z if it includes a node such that either

1.arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z  

2.the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

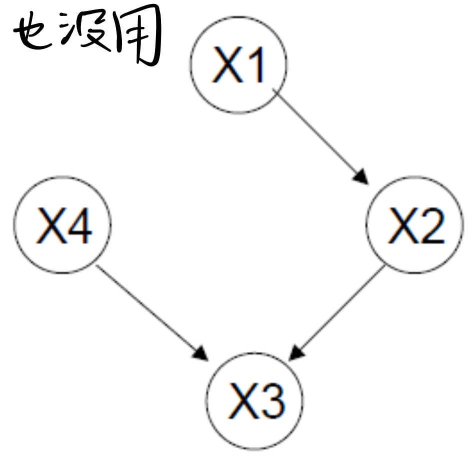


Head2Head.

X4 indep of X1 given X3? ✗ → 原先是有X3故不独立，但选中X3，就不独立

X4 indep of X1 given {X3, X2}? ✓ X2断掉，那么X3连上也没用

X4 indep of X1 given {}? ✓ X3是断的。



X and Y are **<u>D-separated</u>** by Z (and therefore conditionally indep, given Z) iff every path from any variable in X to any variable in Y is **<u>blocked</u>**

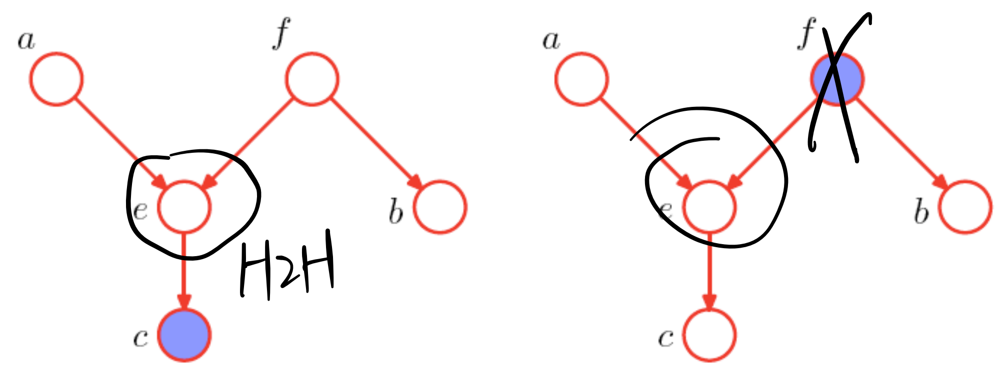
A path from variable A to variable B is **blocked** if it includes a node such that either

1. arrows on the path meet either head-to-tail or tail-to-tail at the node and this node is in Z

2. or, the arrows meet head-to-head at the node, and neither the node, nor any of its descendants, is in Z

a indep of b given c? √

a indep of b given f ? √

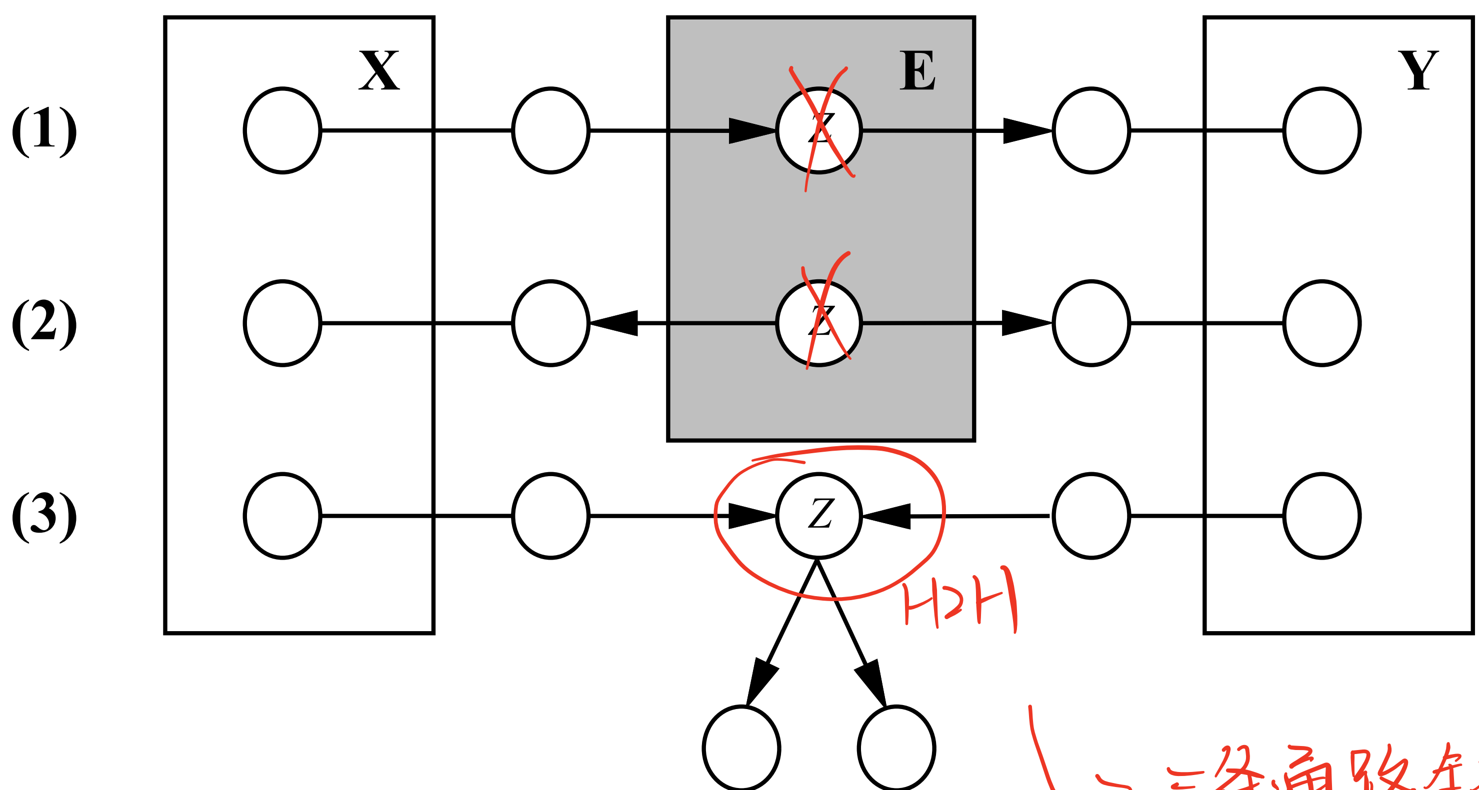


(a) (b)

D-separation

Q: When are nodes X independent of nodes Y given nodes E?

A: When every undirected path from a node in X to a node in Y is **d-separated** by E.

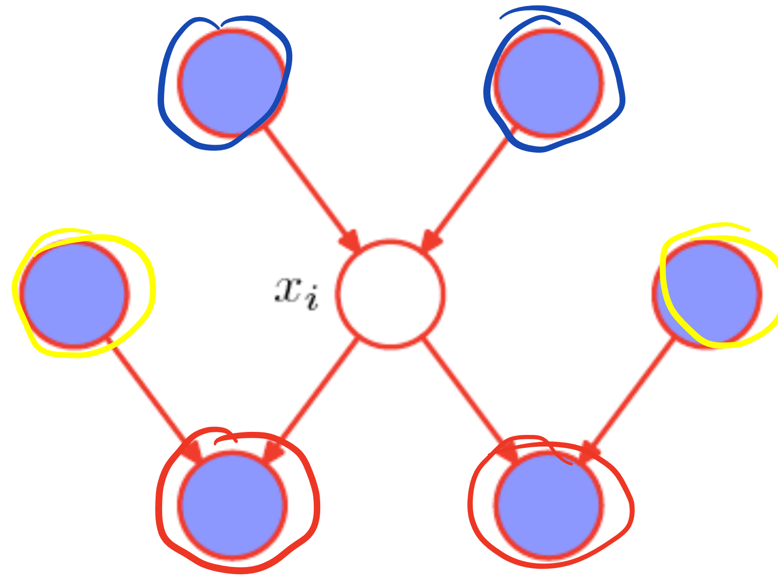


↳ 三条通路全被Block

=> 是独立的

Markov Blanket 马尔可夫毯

The Markov blanket of a node $\mathbf{x}_i$ comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of $\mathbf{x}_i$, conditioned on all the remaining variables in the graph, is dependent only on the variables in the Markov blanket.



定义 x_i，其它所有点为 $x_j, j \neq i$

X_{MB_i}: 马尔可夫毯为　$x_{\{j \neq i\}} = X_{MB_i} \cup X_{\overline{MB_i}}$

$$\Rightarrow P(x_i \mid x_{\{j \neq i\}}) = P(x_i \mid X_{MB_i}, X_{\overline{MB_i}})$$

$$= P(x_i \mid X_{MB_i})$$

from [Bishop, 8.2]

$$\Rightarrow x_i \perp\!\!\!\perp X_{\overline{MB_i}} \mid X_{MB_i}$$

What You Should Know

- Bayes nets are convenient representation for encoding dependencies / conditional independence
- BN = Graph plus parameters of CPD's
  - Defines joint distribution over variables
  - Can calculate everything else from that
  - Though inference may be intractable
- Reading conditional independence relations from the graph
  - Each node is cond indep of non-descendents, given only its parents
  - X and Y are conditionally independent given Z if Z D-separates every path connecting X to Y
  - Marginal independence : special case where Z={}

Inference in Bayes Nets

条件少时复杂度指数上升.

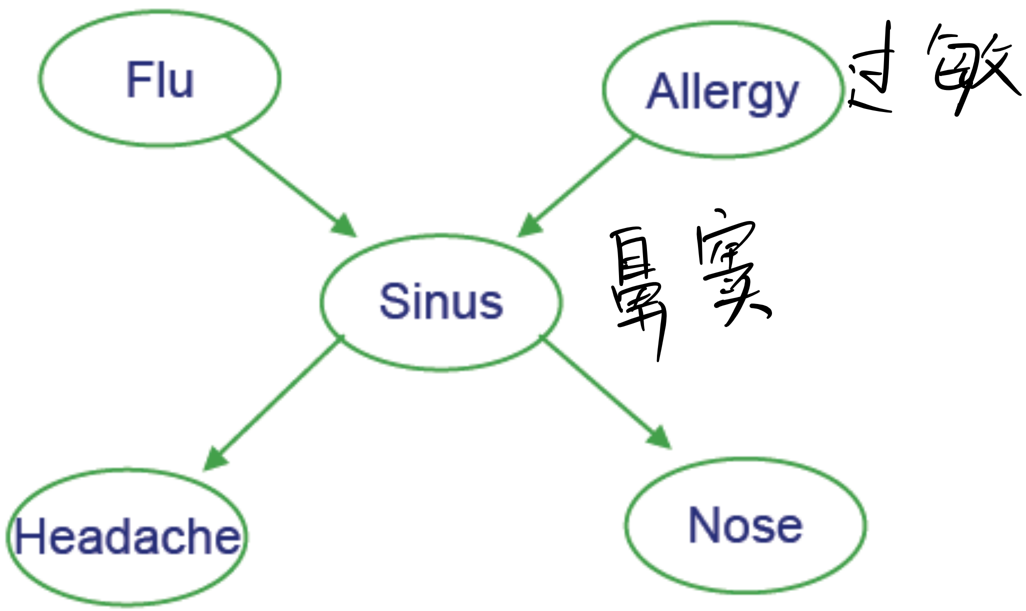
$P(x_1|x_2) = \sum P(x_1, x_2, x_3, \cdots x_n)$

- In general, intractable (NP-Hard. ~~NP-complete~~)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Belief propagation 线性 => 树
- Sometimes use Monte Carlo methods 采样 (但采样也是有代价
  - Generate many samples according to the Bayes Net distribution, then count up the results 无限采样则无偏.
- Variational methods for tractable approximate solutions 变分. 转化成优化问题.

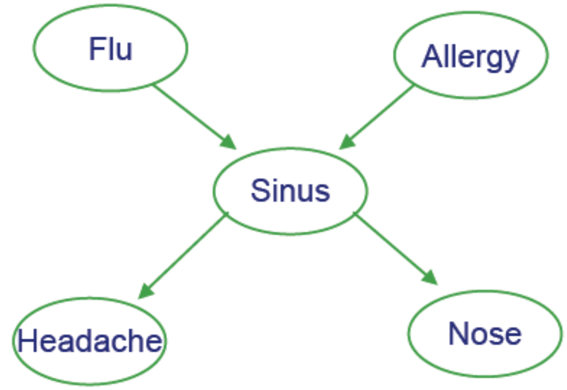
$q_\varphi(x) \Rightarrow \min KL(q_\varphi(x) \| p(x))$

Example

- Bird flu and Allegies both cause Sinus problems
- Sinus problems cause Headaches and runny Nose



Prob. of joint assignment: easy



- Suppose we are interested in joint assignment <F=f,A=a,S=s,H=h,N=n>

What is P(f,a,s,h,n)? $= P(f)P(a)P(s|f,a)P(h|s)P(n|s)$ => 4次乘法.

$P(f,a,s,h) = \sum_n P(f,a,s,h,N=n) = P(f,a,s,h,N=1) + P(f,a,s,h,N=0)$

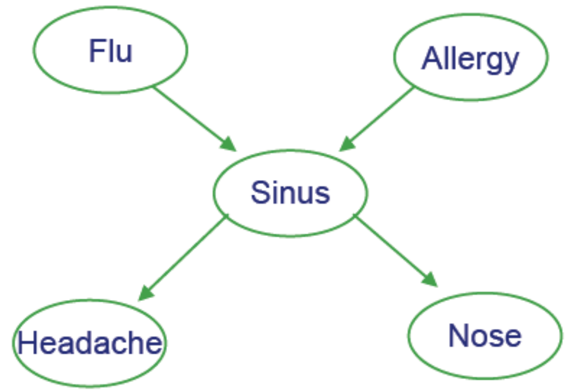
=> 8次乘法

$P(f) = \sum_{a,s,h,n} P(f,A=a,S=s,H=h,N=n)$ => 时间爆炸：$2^4 \times 4$次乘法

有k个值没被观测 => 2^k个式子相加 => $2^k(n-1)$次乘法.

let's use p(a,b) as shorthand for p(A=a, B=b)

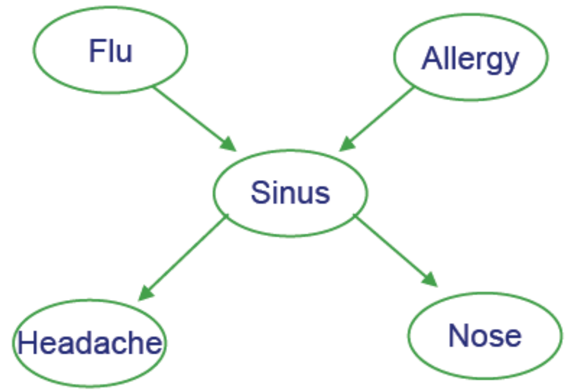
Prob. of marginals: not so easy



- How do we calculate P(N=n) ?

let's use p(a,b) as shorthand for p(A=a, B=b)

Generating a sample from joint distribution: easy



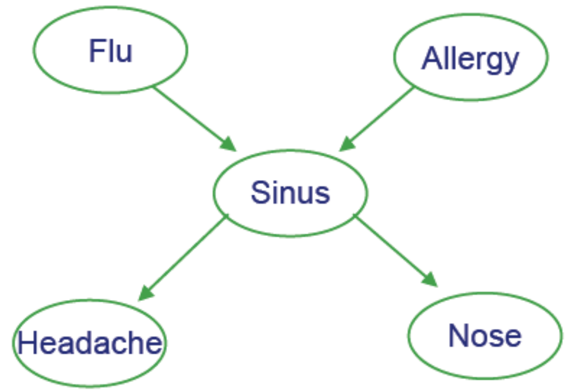
How can we generate random samples drawn according to P(F,A,S,H,N)?

Hint: random sample of F according to $P(F=1) = \theta_{F=1}$:

- draw a value of r uniformly from [0,1]
- if $r<\theta$ then output F=1, else F=0

let's use p(a,b) as shorthand for p(A=a, B=b)

Generating a sample from joint distribution: easy



How can we generate random samples drawn according to P(F,A,S,H,N)?

Hint: random sample of F according to $P(F=1) = \theta_{F=1}$:

- draw a value of r uniformly from [0,1]
- if $r<\theta$ then output F=1, else F=0

Solution:

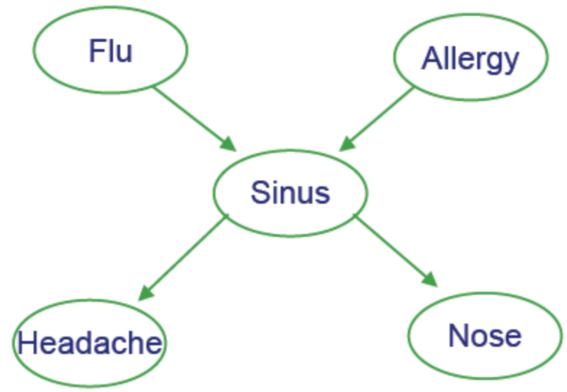
- draw a random value f for F, using its CPD
- then draw values for A, for S|A,F, for H|S, for N|S

	$S=1$	$S=0$
AF	θ_{11}	$1-\theta_{11}$
$A\bar{F}$	θ_{10}	$1-\theta_{10}$
$\bar{A}F$	θ_{01}	$1-\theta_{01}$
$\bar{A}\bar{F}$	θ_{00}	$1-\theta_{00}$

采样.

$\Rightarrow \{a_i, f_i, s_i\}$

Generating a sample from joint distribution: easy



Note we can estimate marginals like P(N=n) by generating many samples from joint distribution, then count the fraction of samples for which N=n

$$\frac{|D_1| / |M|}{|D_2| / |M|} = \frac{|D_1|}{|D_2|}$$

但非常依赖于样本量及分布是否易于采样

Similarly, for anything else we care about

$$P(F=1|H=1, N=0) = \frac{P(F=1, H=1, N=0)}{P(H=1, N=0)}$$

$D_1 : \{F=1, H=1, N=0\}$

$D_2 : \{H=1, N=0\}$

→ weak but general method for estimating <u>any</u> probability term…

Inference in Bayes Nets

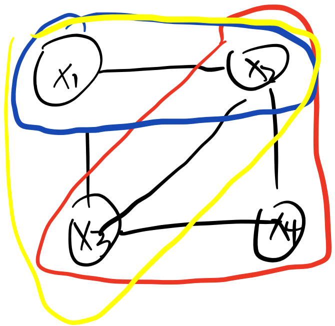
- In general, intractable (NP-complete)
- For certain cases, tractable
  - Assigning probability to fully observed set of variables
  - Or if just one variable unobserved
  - Or for singly connected graphs (ie., no undirected loops)
    - Variable elimination
    - Belief propagation
- Often use Monte Carlo methods
  - e.g., Generate many samples according to the Bayes Net distribution, then count up the results
  - Gibbs sampling
- Variational methods for tractable approximate solutions

see Graphical Models course 10-708

马尔可夫网络：

无向有环图

可以有环



组：可以直接相连点，组合.

x_1、x_4不直接相连

$$\Rightarrow p(x) = \frac{1}{Z}\prod_c \phi_c(x_c) = \frac{1}{Z}(\phi_{1,2}(x_1,x_2)\cdots)$$

$$\phi_c(x_c) = e^{-E(x_c)}$$

$$Z = \sum_x \prod_c \phi_c(x_c)$$

规范系数

$$\Rightarrow p(x) = \frac{1}{Z}\prod_c \phi_c(x_c) = \frac{1}{Z}\prod_c \exp(-E_\theta(x_c))$$

$$= \frac{1}{Z_\theta}\exp\left(-\sum_c E(x_c)\right)$$

EBM

Energy Based Model

Energy Function

$$\min_\theta KL(P_{data}(x)\,\|\,P_\theta(x)), \quad P_\theta(x) = \frac{1}{Z_\theta}\exp(-E_\theta(x))$$

$$\Rightarrow \ell(\theta) = \max_\theta \mathbb{E}_{P_{data}}[\ln P_\theta(x)] = \frac{1}{N}\sum_{i=1}^{N}\ln P_\theta(x_i) = \mathbb{E}_{P_{data}}[-E_\theta(x) - \ln Z_\theta]$$

导数

$$\nabla_\theta \ell(\theta) = \mathbb{E}_{P_{data}}[-\nabla_\theta E_\theta(x)] - \mathbb{E}_{P_{data}}\left[-\int P_\theta(x)\nabla_\theta E_\theta(x)\,dx\right] + \mathbb{E}_{P_\theta}[\nabla_\theta E_\theta(x)]$$

$$\frac{\partial \ln Z_\theta}{\partial \theta} = \frac{\partial \ln Z_\theta}{\partial Z_\theta}\cdot\frac{\partial Z_\theta}{\partial \theta} = -\int \exp(-E_\theta(x))\nabla_\theta E_\theta(x)\,dx$$

Monte Carlo

$$\simeq -\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta E_\theta(x_i) + \frac{1}{K}\sum_{j=1}^{K}\nabla_\theta E_\theta(x_j') \qquad x_j' \sim P_\theta(x)$$

SGLD：采样方法. 第k+1个样本根据x_k样本采样.

高斯分布.

$$x_{k+1} = x_k - \frac{\alpha}{2}\nabla_x E_\theta(x) + \varepsilon. \quad \varepsilon \sim \mathcal{N}(0,\alpha)$$

α是超参数.