

REDES BAYESIANAS

ANO LECTIVO 2013/2014 - 1º SEMESTRE (14.1 - 14.3) –
ADAPTADO DE
[HTTP://AIMA.EECS.BERKELEY.EDU/SLIDES-TEX/](http://aima.eecs.berkeley.edu/slides-tex/)

Resumo

- ◇ Sintaxe
- ◇ Semântica
- ◇ Distribuições parametrizadas

Redes Bayesianas

Notação gráfica simples para asserções de independência condicional e portanto uma especificação compacta para distribuições conjuntas totais

Sintaxe:

- um conjunto de nós, um por variável

- um grafo acíclico dirigido (arco $\approx$ “influencia directamente”)

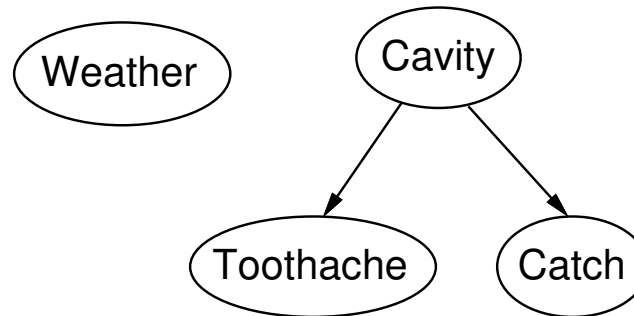
- uma distribuição condicional para cada nó dados os seus pais:

$$\mathbf{P}(X_i | \text{Parents}(X_i))$$

No caso mais simples, distribuição condicional representada como uma **tabela de probabilidade condicionada** (CPT) especificando a distribuição de X_i para cada combinação dos valores dos pais

Exemplo

Topologia da rede codifica asserções de independência condicional:



Weather é independente das outras variáveis

Toothache e *Catch* são condicionalmente independentes dado *Cavity*

Exemplo

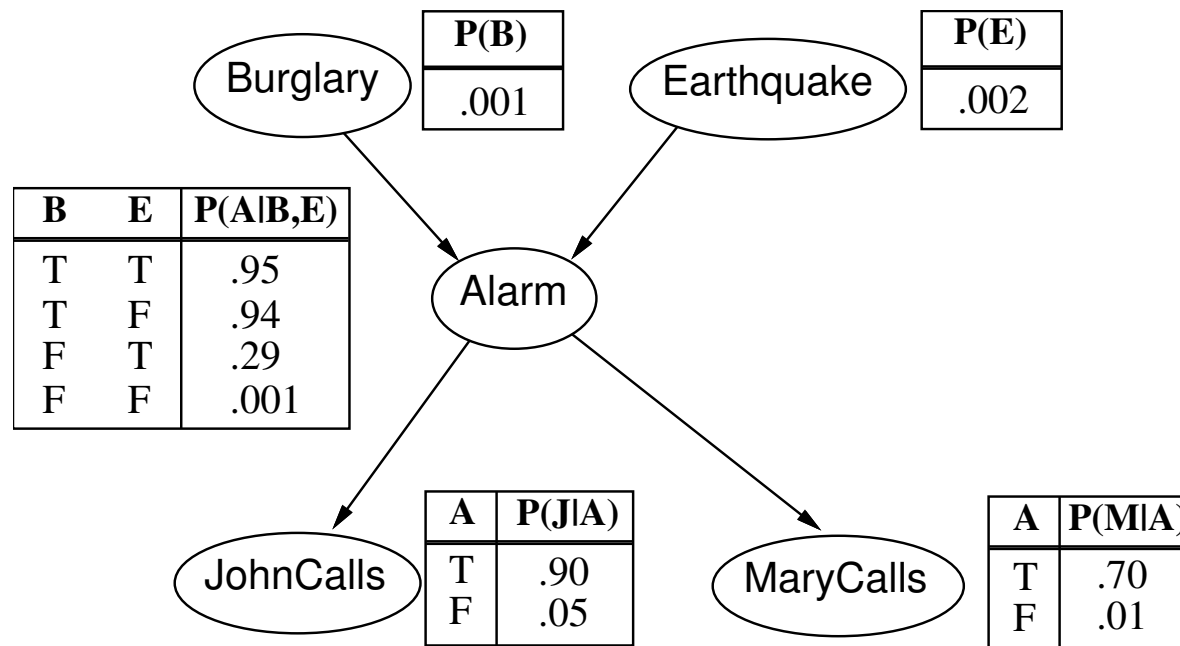
Estou no trabalho, o vizinho John telefona-me para dizer que o meu alarme está a tocar, mas a vizinha Mary não telefona. Ocasionalmente dispara por causa de pequenos tremores de terra. A minha casa está a ser assaltada?

Variáveis: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

Topologia da rede reflecte conhecimento “causal”:

- Um assaltante pode disparar o alarme
- Um tremor de terra pode disparar o alarme
- O alarme pode fazer com que a Mary telefone
- O alarme pode fazer com que o John telefone

Exemplo (cont.)



Representação Compacta

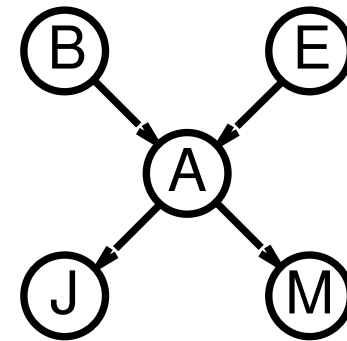
Uma CPT para X_i Booleana com k pais Booleanos requer 2^k linhas para as combinações de valores dos pais

Cada linha requer um número p para $X_i = \text{true}$ (o número para $X_i = \text{false}$ é simplesmente $1 - p$)

Se cada variável não tem mais de k pais, a rede precisa no total de $O(n \cdot 2^k)$ números

I.e., cresce linearmente com n , vs. $O(2^n)$ para a distribuição conjunta total

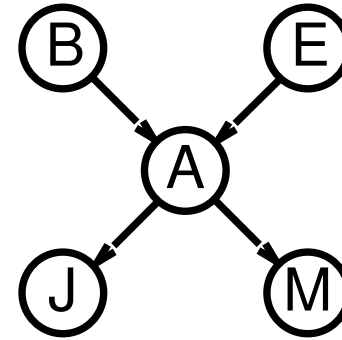
Para a rede do assaltante, $1 + 1 + 4 + 2 + 2 = 10$ números (vs. $2^5 - 1 = 31$)



Semântica Global

Semântica **global** define a distribuição conjunta total como o produto distribuições condicionais locais:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i))$$



Nesta situação dizemos que a distribuição P é compatível com a rede G .

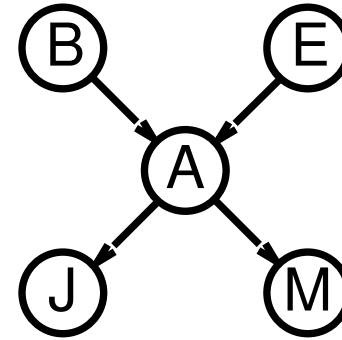
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

=

Semântica Global

Semântica **global** define a distribuição conjunta total como o produto distribuições condicionais locais:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i))$$



Nesta situação dizemos que a distribuição P é compatível com a rede G .

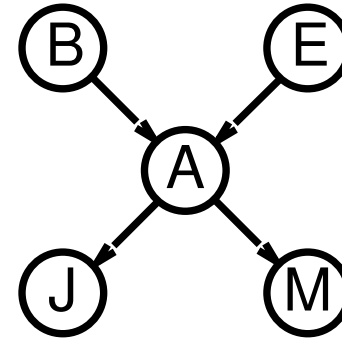
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

Semântica Global

Semântica **global** define a distribuição conjunta total como o produto distribuições condicionais locais:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i))$$



Nesta situação dizemos que a distribuição P é compatível com a rede G .

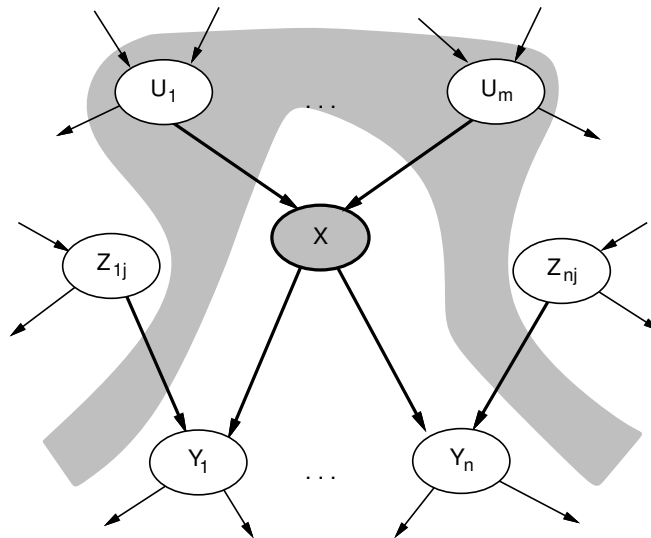
e.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e)$$

$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

Semântica Local

Semântica **Local**: cada nó é condicionalmente independente dos seus não-descendentes dado os seus pais

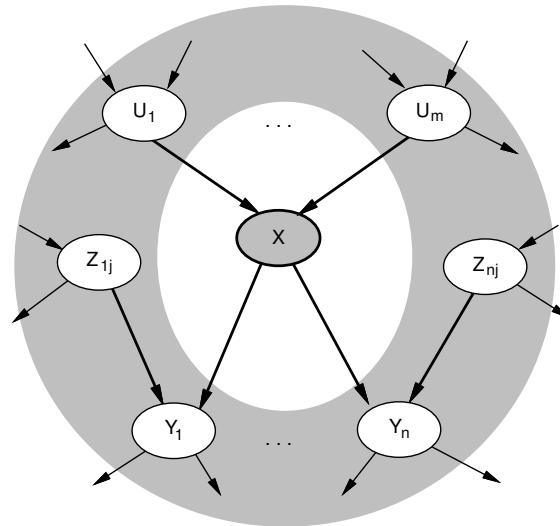


Para todo o nó X assume-se que $P(X|Z_{1j}, \dots, Z_{nj}, U_1, \dots, U_m) = P(X|U_1, \dots, U_m)$.
Concluiu-se que:

Teorema: **Semântica Local** $\Leftrightarrow$ **Semântica Global**

Markov blanket

Cada nó é condicionalmente independente de todos os outros dado o seu
Markov blanket: pais + filhos + pais dos filhos



Seja $W_1, \dots, W_p$ um qualquer conjunto de nós da rede disjunto do Markov Blanket de X tem-se $P(X|W_1, \dots, W_p, U_1, \dots, U_m, Y_1, \dots, Y_n, Z_{1j}, \dots, Z_{nj}) = P(X|U_1, \dots, U_m, Y_1, \dots, Y_n, Z_{1j}, \dots, Z_{nj})$

Grafos morais e d -separação

Sejam 3 subconjuntos de nós X, Y, Z de uma rede Bayesiana G . O grafo moral é construído da seguinte forma:

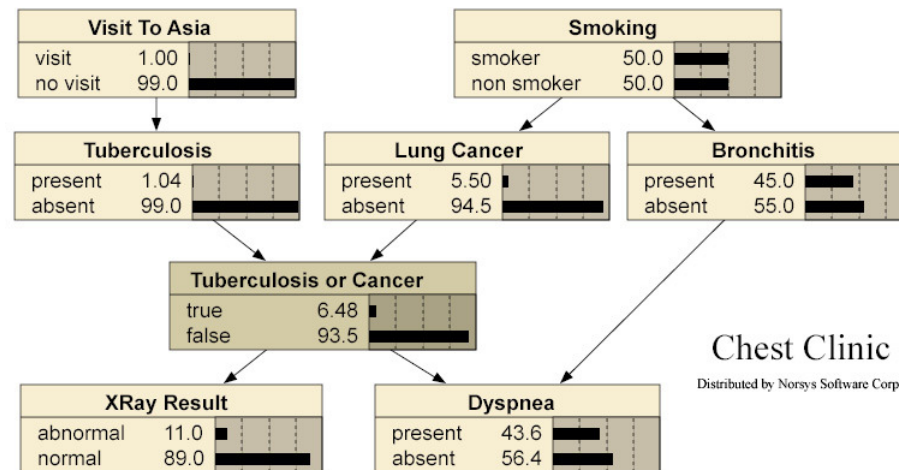
- Removem-se todos os nós de G exceptuando aqueles que se encontram em $X \cup Y \cup Z$ e todos os seus antecessores.
- Liga-se com uma aresta todo o par de nós que têm um filho comum.
- Removem-se as setas dos restantes arcos.

Dizemos que $(X \perp\!\!\!\perp Y|Z)_G$ (Z d -separa X de Y) quando todo o caminho entre X e Y é interceptado por Z no grafo moral.

Teorema: Sejam X, Y e Z quaisquer subconjuntos disjuntos de uma rede G e para toda a função de probabilidade P compatível com G , tem-se

1. Se $(X \perp\!\!\!\perp Y|Z)_G$ então X é condicional. indepen. de Y dado Z em P ,
2. Se X é condicionalmente independente de Y dado Z para todas as distribuições P compatíveis com G , então $(X \perp\!\!\!\perp Y|Z)_G$.

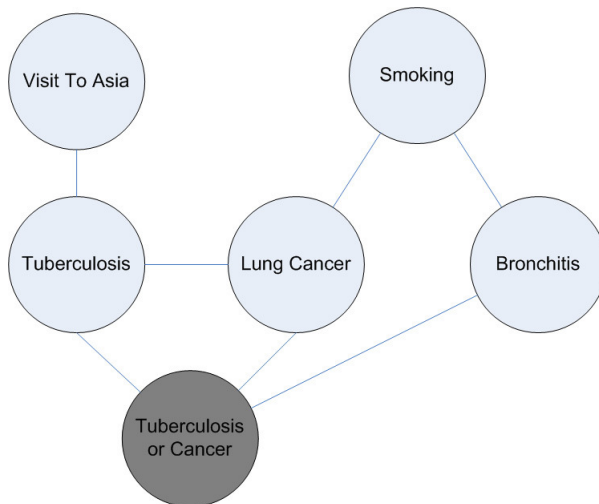
Exemplo de d-separação



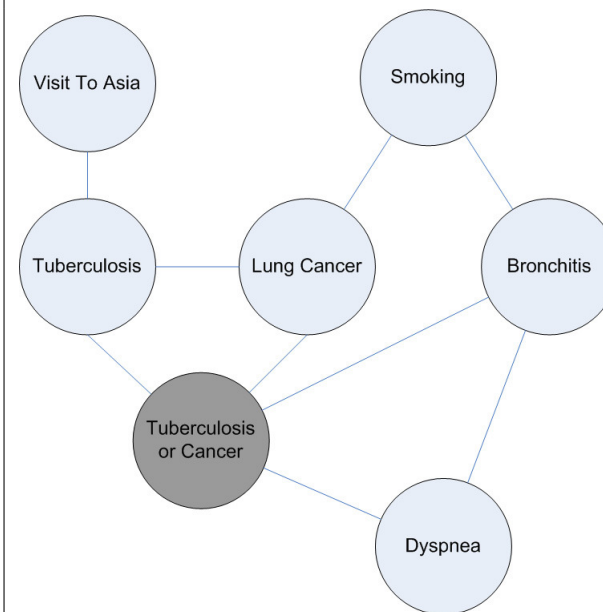
- ◇ “Tuberculosis or Cancer” d-separa “Tuberculosis” de “Bronchitis”?
- ◇ “Tuberculosis or Cancer” d-separa “Tuberculosis” de “Dyspnea”?
- ◇ “Lung Cancer” d-separa Tuberculosis de “Bronchitis”?
- ◇ “Tuberculosis or Cancer” e “Lung Cancer” d-separam “Tuberculosis” de “Dyspnea”?

Exemplo de d-separação (cont).

“Tuberculosis or Cancer” d-separa “Tuberculosis” de “Bronchitis”? **Não**

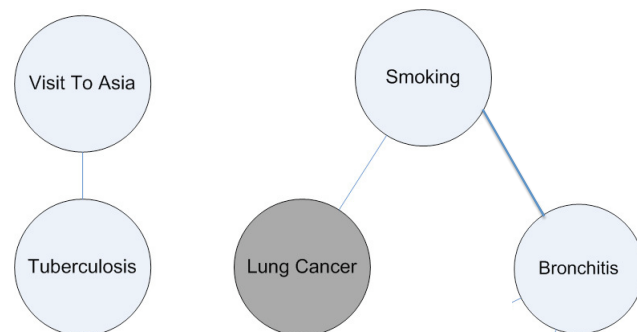


“Tuberculosis or Cancer” d-separa “Tuberculosis” de “Dyspnea”? **Não**

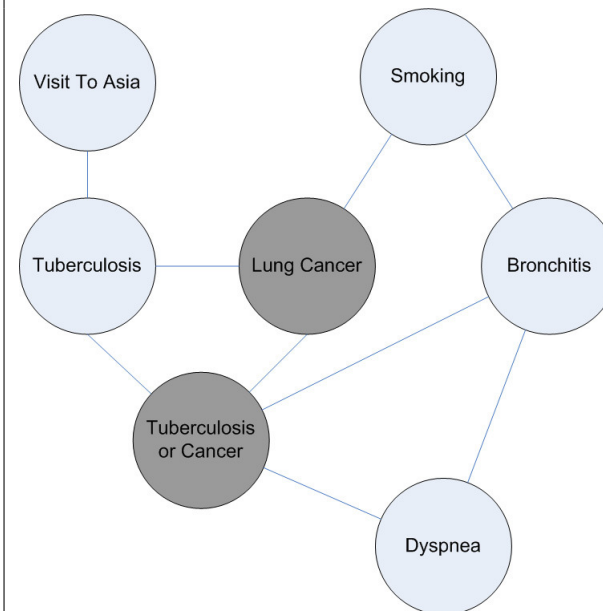


Exemplo de d-separação (cont).

“Lung Cancer” d-separa Tuberculosis de “Bronchitis”? **Sim**



“Tuberculosis or Cancer” e “Lung Cancer” d-separam “Tuberculosis” de “Dyspnea”? **Sim**



Construção de redes Bayesianas

É necessário um método tal que uma série de asserções testadas localmente de independência condicional garantam a semântica global pretendida

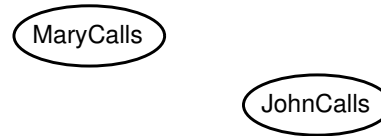
1. Escolher uma ordenação das variáveis $X_1, \dots, X_n$
2. For $i = 1$ to n
 adicionar X_i à rede
 seleccionar pais de $X_1, \dots, X_{i-1}$ tal que
 $\mathbf{P}(X_i | \text{Parents}(X_i)) = \mathbf{P}(X_i | X_1, \dots, X_{i-1})$

Esta escolha de pais garante a semântica global:

$$\begin{aligned}\mathbf{P}(X_1, \dots, X_n) &= \prod_{i=1}^n \mathbf{P}(X_i | X_1, \dots, X_{i-1}) \quad (\text{regra da cadeia}) \\ &= \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i)) \quad (\text{por construção})\end{aligned}$$

Exemplo

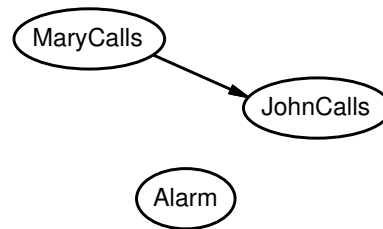
Suponhamos a seguinte ordenação M, J, A, B, E



$$P(J|M) = P(J)?$$

Exemplo

Suponhamos a seguinte ordenação M, J, A, B, E

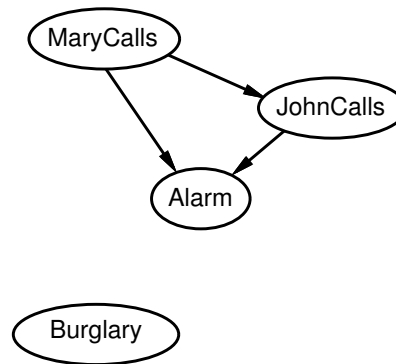


$P(J|M) = P(J)$? Não

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

Exemplo

Suponhamos a seguinte ordenação M, J, A, B, E



$P(J|M) = P(J)$? Não

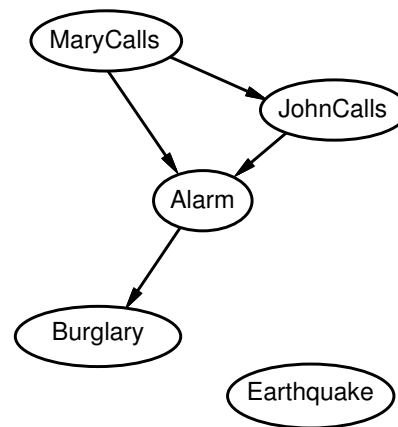
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? Não

$P(B|A, J, M) = P(B|A)$?

$P(B|A, J, M) = P(B)$?

Exemplo

Suponhamos a seguinte ordenação M, J, A, B, E



$P(J|M) = P(J)$? Não

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? Não

$P(B|A, J, M) = P(B|A)$? Sim

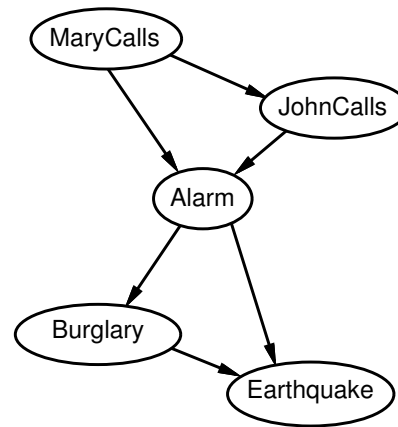
$P(B|A, J, M) = P(B)$? Não

$P(E|B, A, J, M) = P(E|A)$?

$P(E|B, A, J, M) = P(E|A, B)$?

Exemplo

Suponhamos a seguinte ordenação M, J, A, B, E



$P(J|M) = P(J)$? Não

$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$? Não

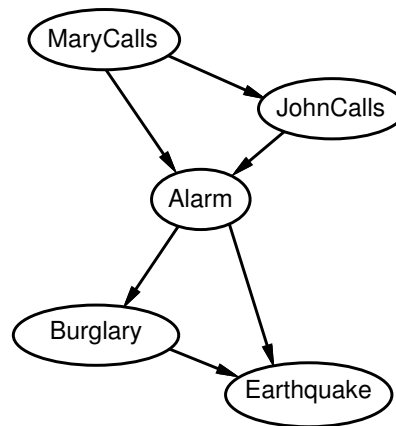
$P(B|A, J, M) = P(B|A)$? Sim

$P(B|A, J, M) = P(B)$? Não

$P(E|B, A, J, M) = P(E|A)$? Não

$P(E|B, A, J, M) = P(E|A, B)$? Sim

Exemplo (cont.)



Decidir sobre independência condicional é difícil nas direcções não causais: modelos causais e independência condicional parecem ser inatos nos humanos!

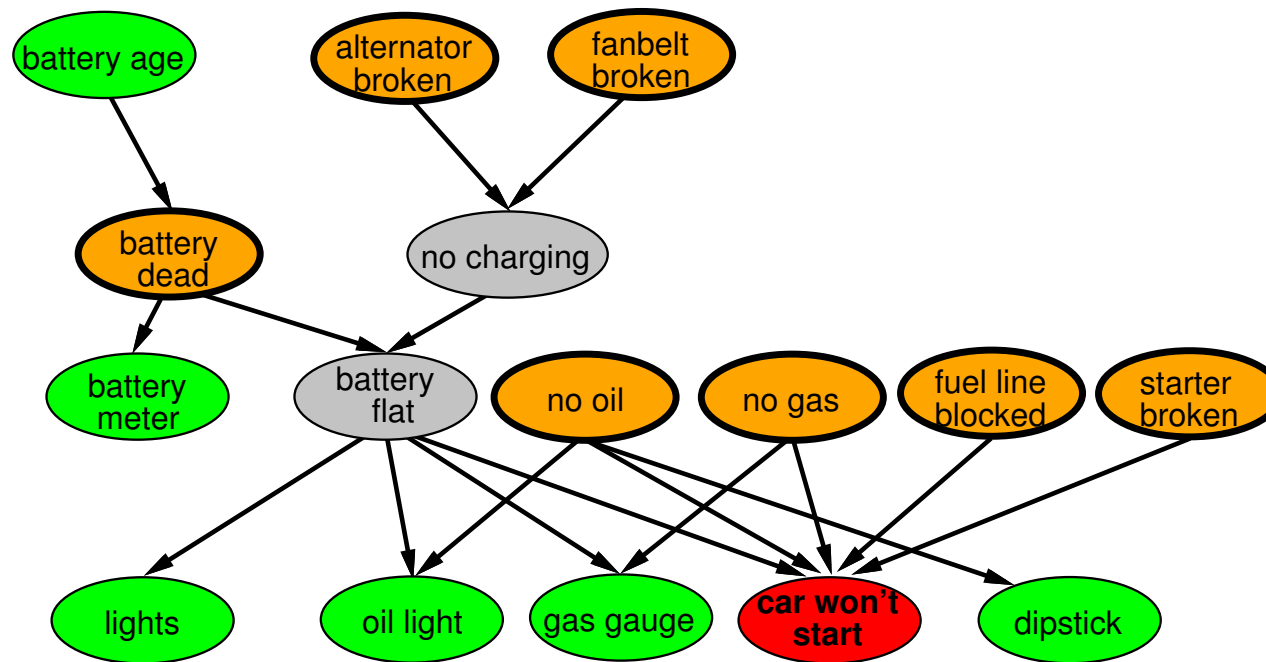
Determinação de probabilidades condicionais é difícil nas direcções não causais. Rede é menos compacta do que a inicial: necessários $1 + 2 + 4 + 2 + 4 = 13$ números

Exemplo: Diagnóstico de avarias

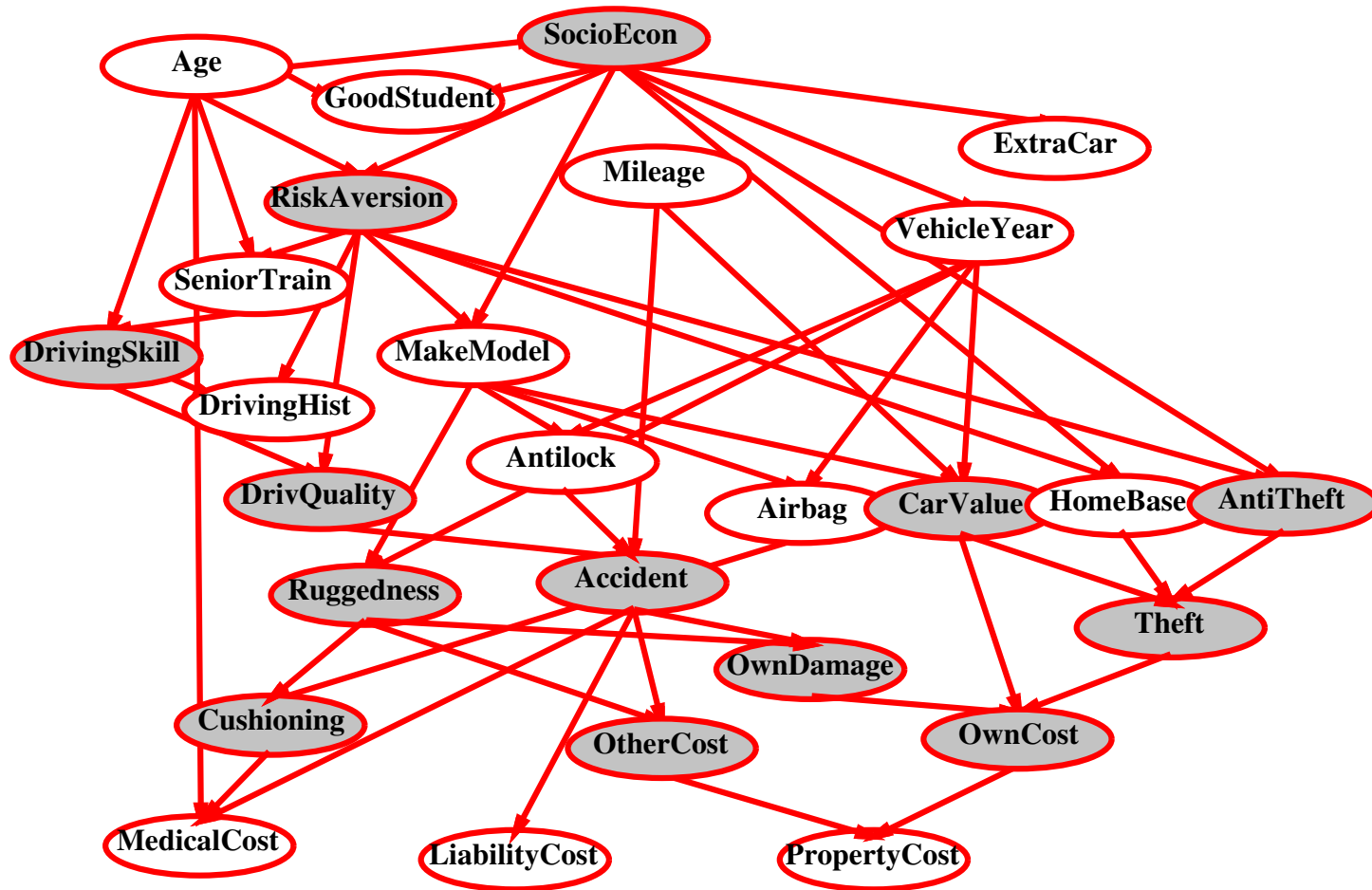
Evidência inicial evidência: carro não pega

Variáveis observáveis (verde), variáveis “avariado, arranja-o” (laranja)

Variáveis ocultas (cinzento) garantem estrutura esparsa, reduzem parâmetros



Exemplo: Seguro do carro



Distribuições condicionais compactas

CPT cresce exponencialmente com o número de pais

CPT fica infinita com pai ou filho tomando valores contínuos

Solução: distribuições *canónicas* que são definidas compactamente

Nós *deterministas* são o caso mais simples:

$$X = f(\text{Parents}(X)) \text{ para alguma função } f$$

E.g., funções Booleanas

$$\text{NorthAmerican} \Leftrightarrow \text{Canadian} \vee \text{US} \vee \text{Mexican}$$

E.g., relações numéricas entre variáveis contínuas

$$\frac{\partial \text{Level}}{\partial t} = \text{inflow} + \text{precipitation} - \text{outflow} - \text{evaporation}$$

Distribuições condicionais compactas (cont.)

Distribuições **Noisy-OR** modelam múltiplas causas que não interagem

- 1) Pais $U_1 \dots U_k$ incluem todas as causas (pode-se adicionar nó)
- 2) Probabilidade de falha independentes q_i para cada causa isoladamente

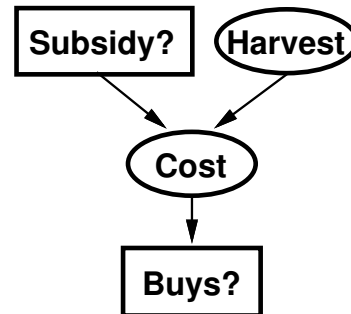
$$\Rightarrow P(X|U_1 \dots U_j, \neg U_{j+1} \dots \neg U_k) = 1 - \prod_{i=1}^j q_i$$

<i>Constip.</i>	<i>Gripe</i>	<i>Malaria</i>	$P(\text{Febre})$	$P(\neg \text{Febre})$
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	0.6
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Número de parâmetros **linear** no número de pais

Redes híbridas (discretas+contínuas)

Discretas (*Subsidy?* e *Buys?*); contínuas (*Harvest* and *Cost*)



Opção 1: discretização—erros podem ser grandes, CPTs grandes

Opção 2: famílias canónicas com número finito de parâmetros

1) Variável contínua, pais discretos+contínuos (e.g., *Cost*)

2) Variável discreta, pais contínuos (e.g., *Buys?*)

Variáveis filho contínuas

Necessária uma função de **densidade condicional** para cada variável filho dados pais contínuos, para cada possível atribuição de pais discretos

Mais habitual é o modelo **linear Gaussiano**, e.g.,:

$$\begin{aligned} P(Cost = c | Harvest = h, Subsidy? = true) \\ &= N(a_t h + b_t, \sigma_t)(c) \\ &= \frac{1}{\sigma_t \sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{c - (a_t h + b_t)}{\sigma_t} \right)^2 \right) \end{aligned}$$

Média *Cost* varia linearmente com *Harvest*, variância fixa

Variação linear não é razoável para todo o domínio

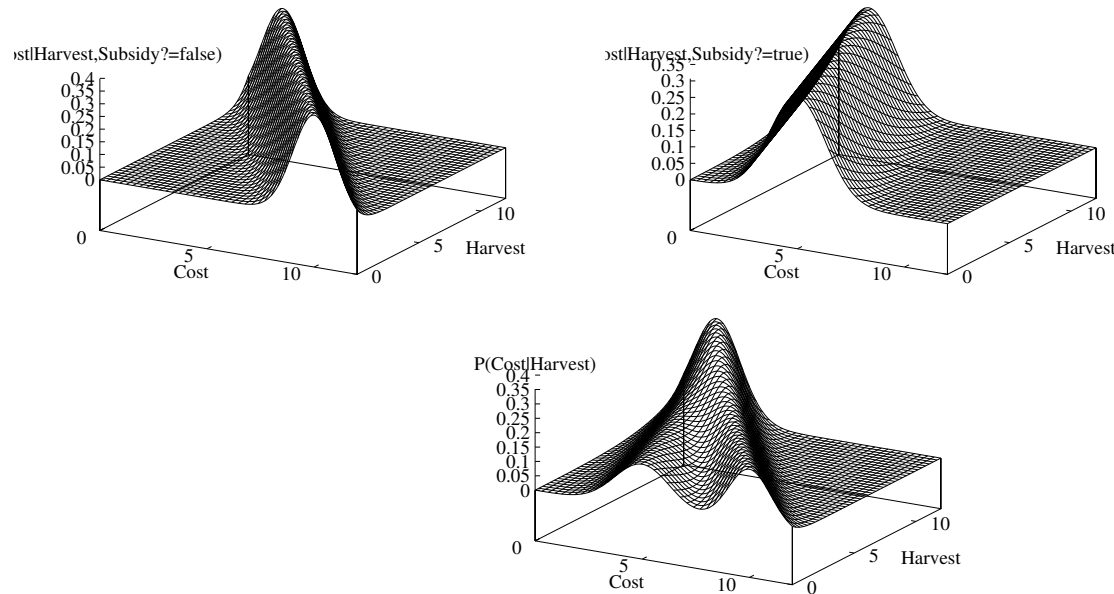
mas funciona bem se os valores **prováveis** de *Harvest* estão limitados

Variáveis filho contínuas

Rede só com variáveis contínuas com distribuição LG

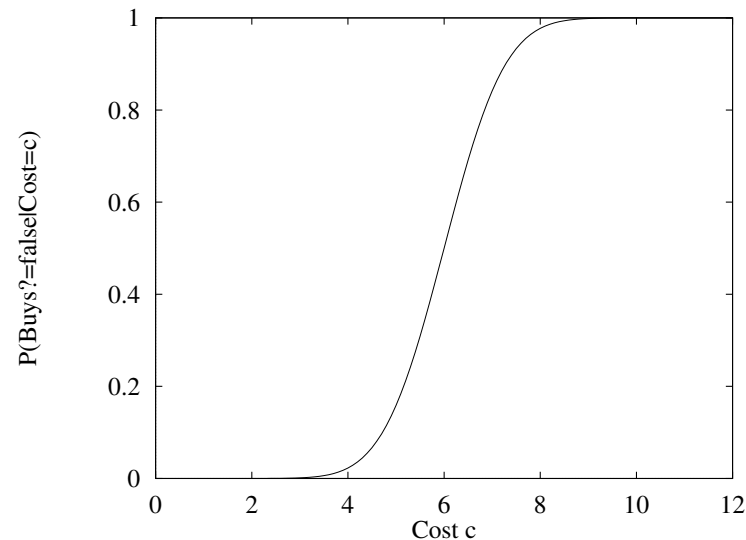
⇒ distribuição conjunta tem distribuição Gaussiana multivariada

Rede discreta+contínua LG é uma rede **Gaussiana condicional**, i.e., uma Gaussiana multivariada para todas as variáveis contínuas para cada combinação de valores de variáveis discretas



Variáveis discretas com pais contínuos

Probabilidade de *Buys?* dado *Cost* deve ser um limiar “suave”:



Distribuição **Probit** utiliza integral de uma Gaussiana:

$$\Phi(x) = \int_{-\infty}^x N(0, 1)(x)dx$$

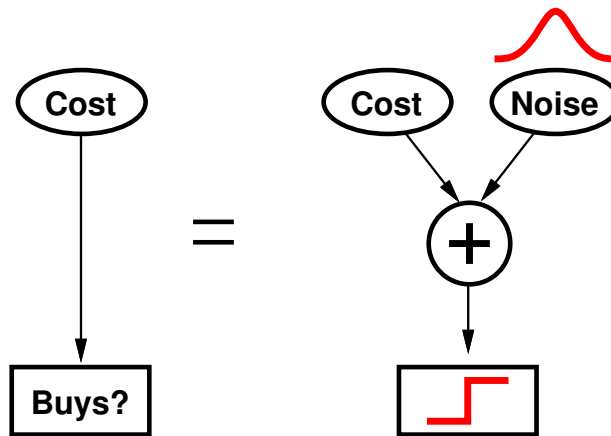
$$P(\text{Buys?} = \text{true} \mid \text{Cost} = c) = \Phi((-c + \mu)/\sigma)$$

$$P(\text{Buys?} = \text{false} \mid \text{Cost} = c) = 1 - \Phi((-c + \mu)/\sigma) = \Phi((c - \mu)/\sigma)$$

Em que μ é o local onde ocorre o limiar e σ um parâmetro que controla a largura do limiar..

Porquê a probit?

1. Tem a forma correcta
2. Pode ser entendida como um limiar cuja localização está sujeita a ruído

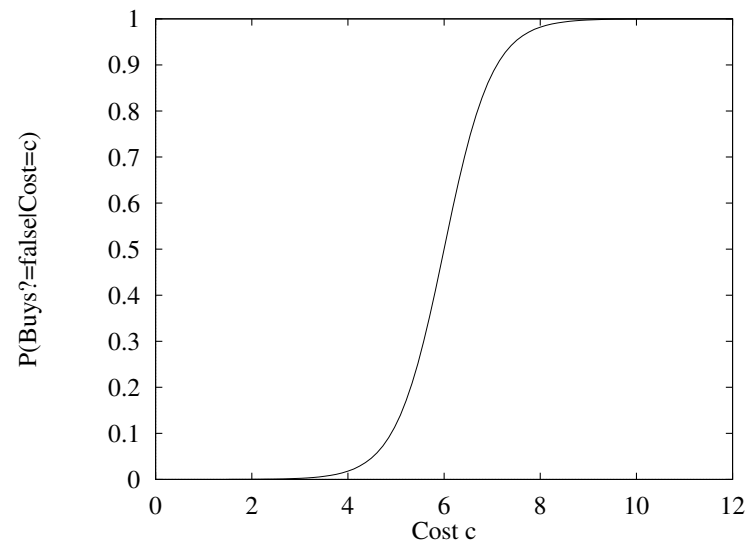


Variável discreta (cont.)

Distribuição **Sigmóide** (ou **logit**) também utilizada em redes neuronais:

$$P(Buys? = true \mid Cost = c) = \frac{1}{1 + \exp(-2\frac{-c+\mu}{\sigma})}$$

Sigmóide tem forma semelhante à da probit mas com caudas maiores:



Sumário

Redes de Bayes são uma representação natural para independência condicional (induzidas causalmente)

Topologia + CPTs = representação compacta de distribuição conjunta

Geralmente fácil de construir por (não)peritos

Distribuições canónicas (e.g., noisy-OR) = representação compacta de CPTs

Variáveis contínuas $\Rightarrow$ distribuições parametrizadas (e.g., Gaussiana linear)