REDES BAYESIANAS CAP 14 (14.1 - 14.3)

Parcialmente adaptado de

http://aima.eecs.berkeley.edu

Resumo

- Sintaxe
- Semântica
- Distribuições parametrizadas

Redes Bayesianas

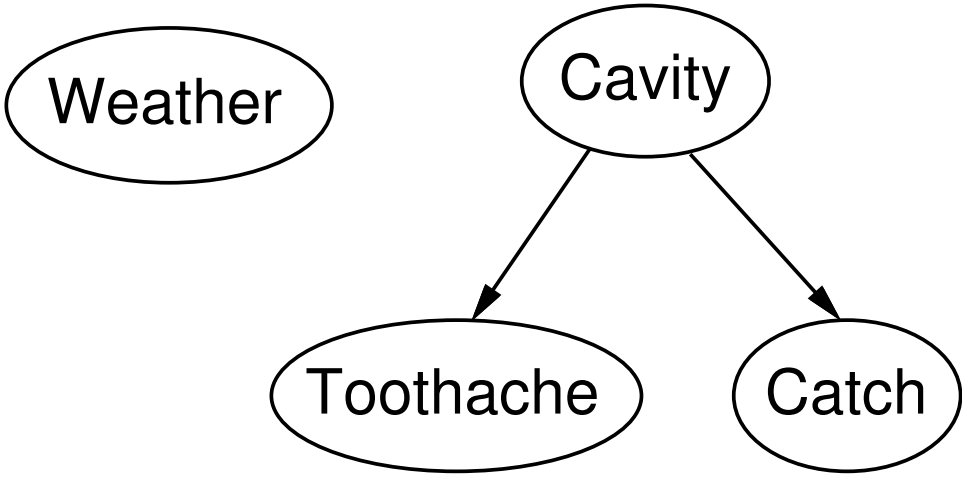
- Notação gráfica simples para asserções de independência condicional, e portanto para uma especificação compacta de distribuições conjuntas totais
- Sintaxe:
  - um conjunto de nós, um por variável
  - um grafo acíclico dirigido (arco ≈ “influencia directamente")
  - uma distribuição condicional para cada nó dados os seus pais:

$$\mathbf{P}(X_i \mid Parents(X_i))$$

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

Exemplo

- A topologia da rede representa asserções de independência condicional:



- *Weather* é independente das outras variáveis
- *Toothache* e *Catch* são condicionalmente independentes dado *Cavity*

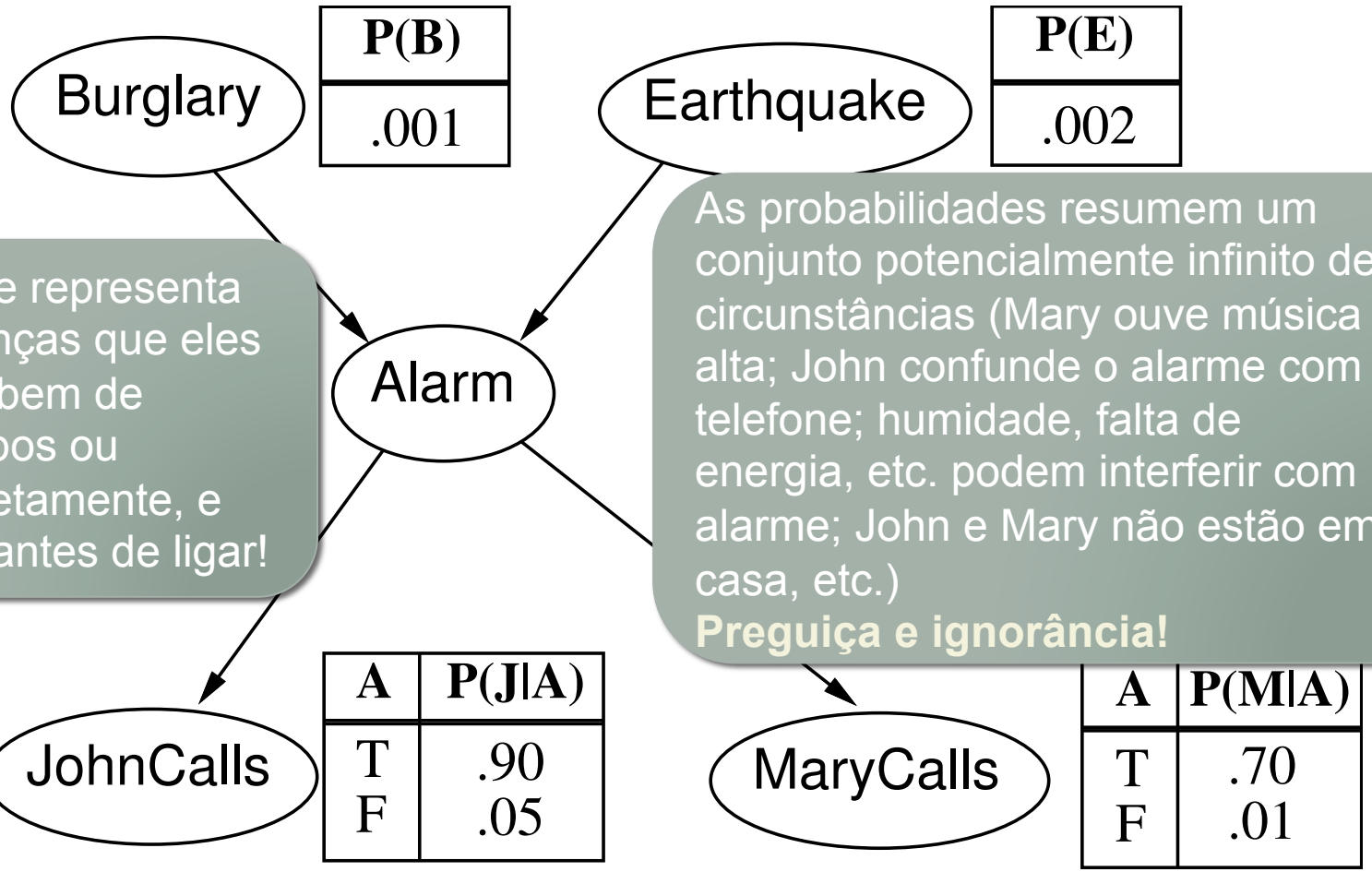
Exemplo

- 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 esta a ser assaltada?
- Variáveis: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Topologia da rede reflecte conhecimento "causal":
  - Um assaltante pode fazer disparar o alarme
  - Um tremor de terra pode fazer disparar o alarme
  - O alarme pode fazer com que a Mary telefone
  - O alarme pode fazer com que o John telefone

Exemplo

Roubos e terramotos afectam diretamente a probabilidade do alarme tocar

O Facto de John e Mary telefonarem só depende diretamente do alarme.

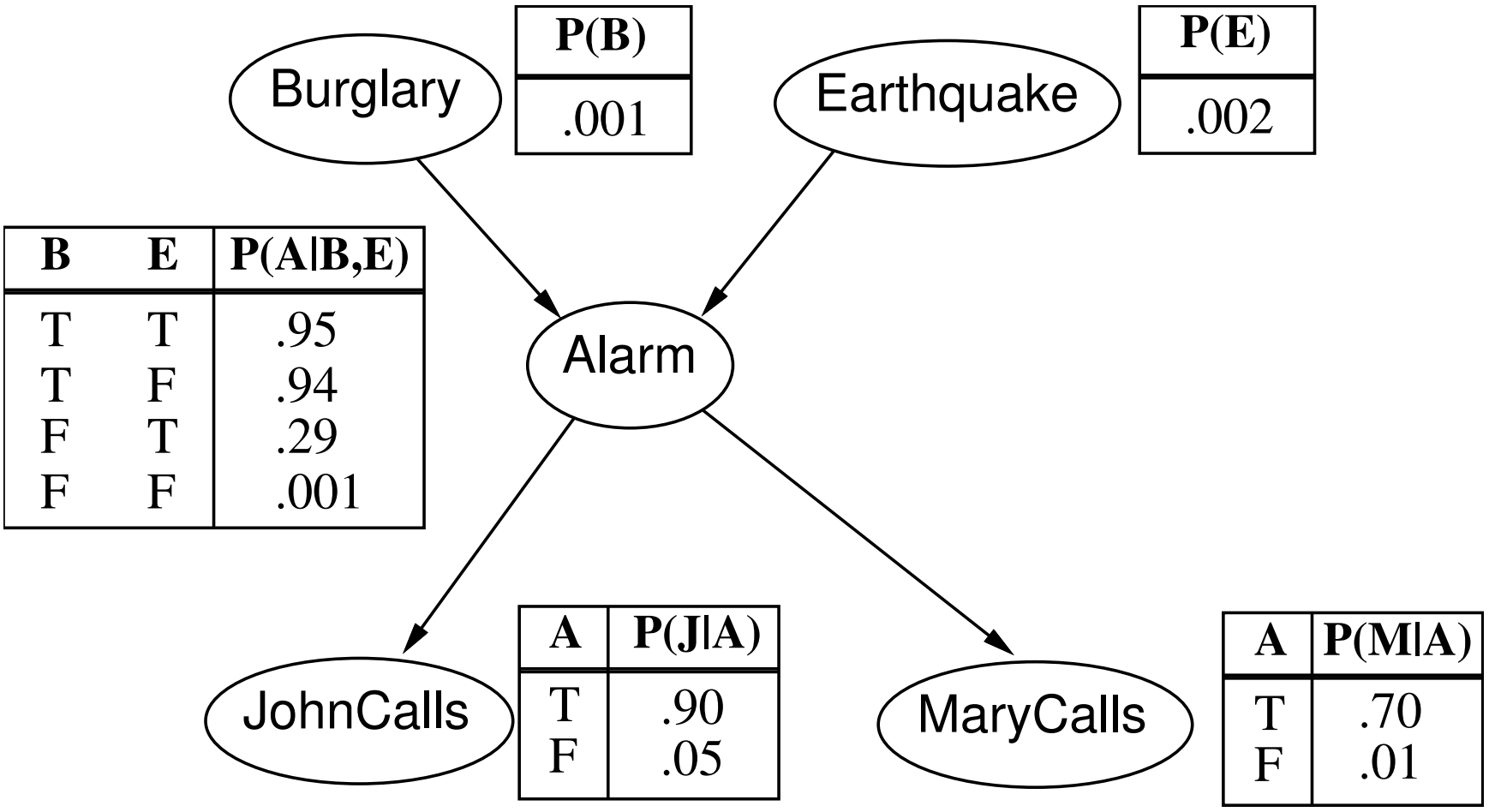


Ou seja, a rede representa as nossas crenças que eles não se apercebem de quaisquer roubos ou terramotos diretamente, e não verificam antes de ligar!

As probabilidades resumem um conjunto potencialmente infinito de circunstâncias (Mary ouve música alta; John confunde o alarme com o telefone; humidade, falta de energia, etc. podem interferir com alarme; John e Mary não estão em casa, etc.)

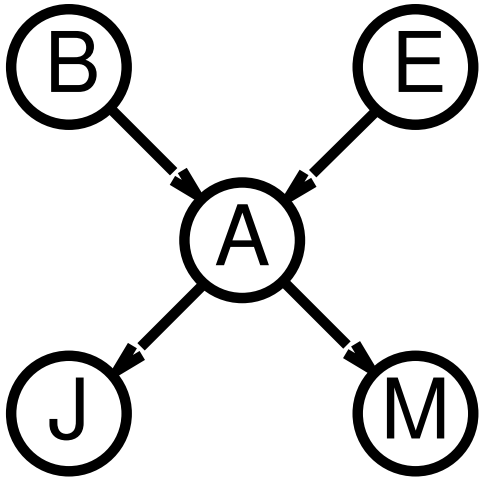
Preguiça e ignorância!

Exemplo


Burglary
P(B)
.001
Earthquake
P(E)
.002
Alarm
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
JohnCalls
A P(J|A)
T .90
F .05
MaryCalls
A P(M|A)
T .70
F .01

Representação compacta

- Cada linha numa CPT contém a probabilidade condicional de cada valor do nó para cada combinação de valores dos pais.
- 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 $X_i = true$ (o número para $X_i = false$ 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. $2^5-1 = 31$)

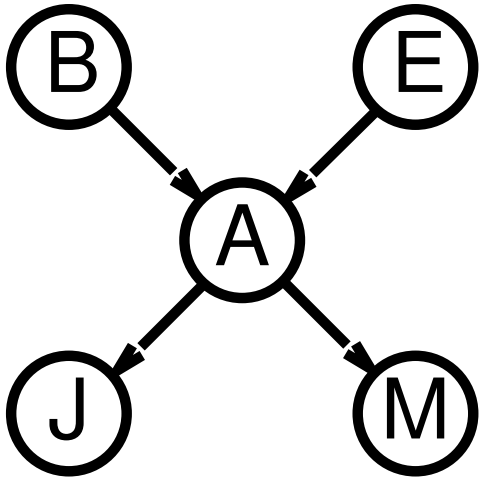
Semântica

- Duas formas equivalentes:
  - Semântica global (ou numérica): interpretar as redes como uma representação da distribuição de probabilidade conjunta
    - Indica como construir uma rede
  - Semântica local (ou topológica): interpretar as redes como uma representação de uma coleção de declarações de independência condicional
    - Indica como fazer inferências com uma rede

Semântica Global

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

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1..n} \mathbf{P}(X_i \mid Parents(X_i))$$

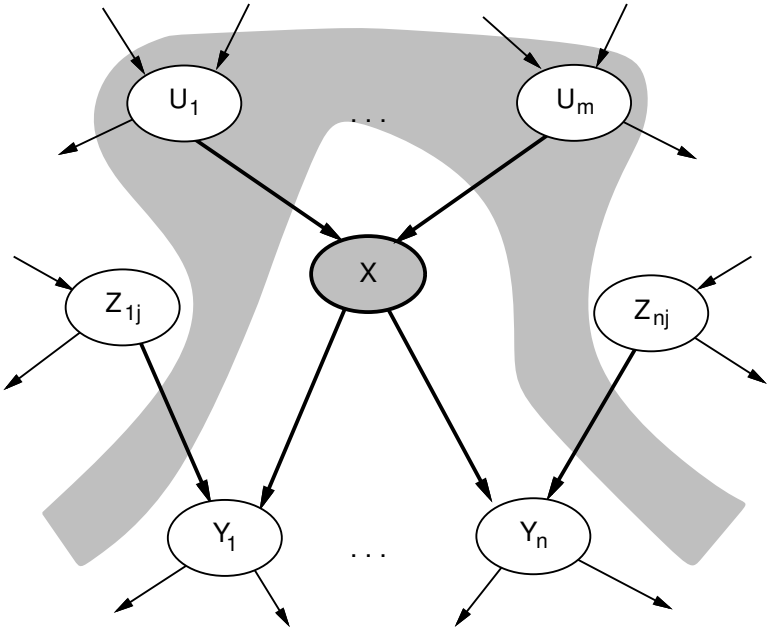


- Nesta situação dizemos que a distribuição **P** é compatível com a rede G.
- E.g.

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) =$$
$$= P(j|a)\, P(m|a)\, P(a|\neg b, \neg e)\, P(\neg b)\, P(\neg e) =$$
$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00063$$

Semântica Local

- Semântica local (topológica): cada nó é condicionalmente independente dos seus não-descendentes, dado os seus pais



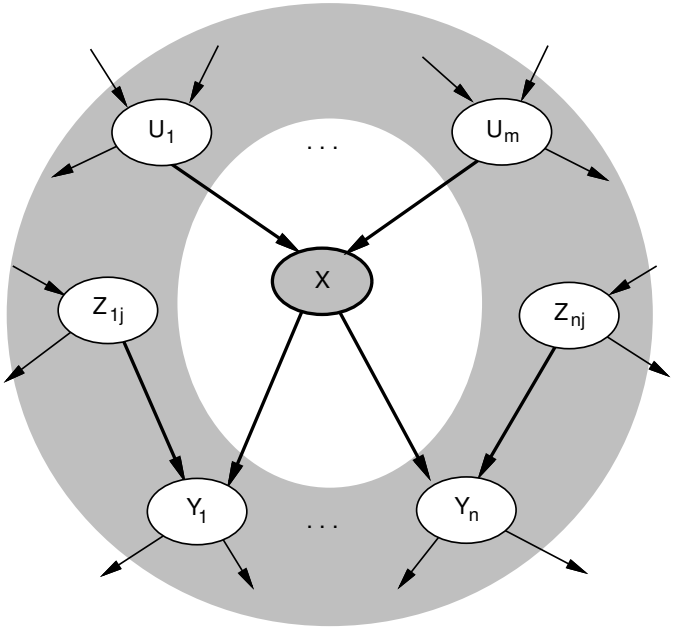
- Para todo o nó X assume-se que

$$P(X|Z_{1j},...,Z_{nj},U_1,...,U_m) = P(X|U_1,...,U_m)$$

- Conclui-se que: Semântica Local ⇔ Semântica Global

Cobertura de Markov

- Cada nó é condicionalmente independente de todos os outros dada a sua cobertura de Markov (Markov blanket): pais + filhos + pais dos filhos



- Seja $W_1,\ldots,W_p$ um qualquer conjunto de nós da rede, disjunto da cobertura de Markov de X. Tem-se:

$$P(X|W_1,\ldots,W_p,U_1,\ldots,U_m,Y_1,\ldots,Y_n,Z_{1j},\ldots,Z_{nj}) = P(X|U_1,\ldots,U_m,Y_1,\ldots,Y_n,Z_{1j},\ldots,Z_{nj})$$

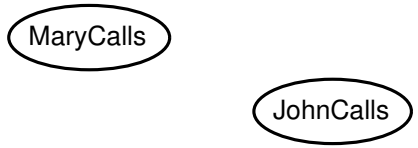
Construção de Redes Bayesianas

1. Escolher uma ordenação das variáveis $X_1, \ldots, X_n$
2. Para $i=1$ até n
   1. Adicionar X_i à rede
   2. Escolher de entre $X_1, \ldots, X_{i-1}$ um conjunto minimal de pais para X_i tal que $\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$ e acrescentar ligações de cada pai para X_i
   3. Escrever a tabela de probabilidade condicional (CPT) para $\mathbf{P}(X_i \mid Parents(X_i))$

- Esta escolha de pais garante a semântica global:
- $\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1..n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$ (regra da cadeia)
  $= \prod_{i=1..n} \mathbf{P}(X_i \mid Parents(X_i))$ (por construção)

Exemplo

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

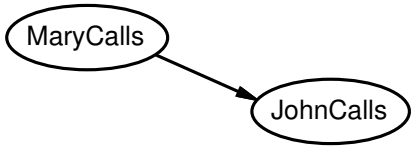


Se a Mary tiver telefonado, isso provavelmente significa que o alarme disparou, o que tornaria mais provável que John também tivesse telefonado...

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

Exemplo

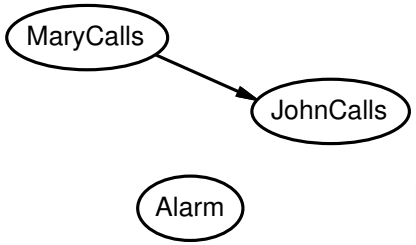
- Suponhamos a seguinte ordenação M, J, A, B, E



- $P(J|M) = P(J)$? Não

Exemplo

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

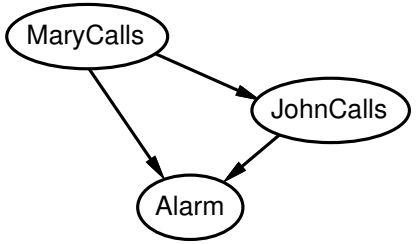


Claramente que se ambos tiverem telefonado, é mais provável que o alarme tenha disparado do que se apenas um deles tiver telefonado...

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

Exemplo

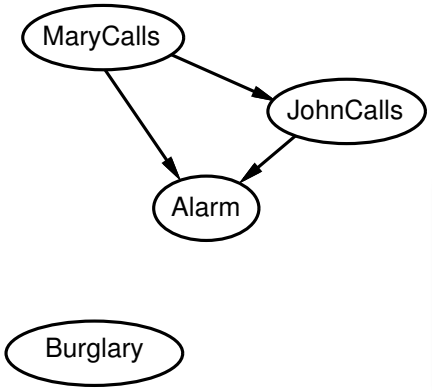
- Suponhamos a seguinte ordenação M, J, A, B, E



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

Exemplo

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

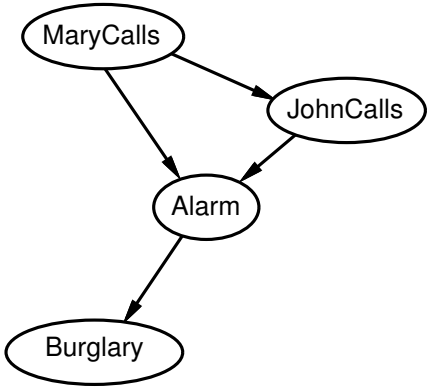


Se soubermos o estado do alarme, então os telefonemas não acrescentam nada em relação ao roubo...
Apenas precisamos do alarme...

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

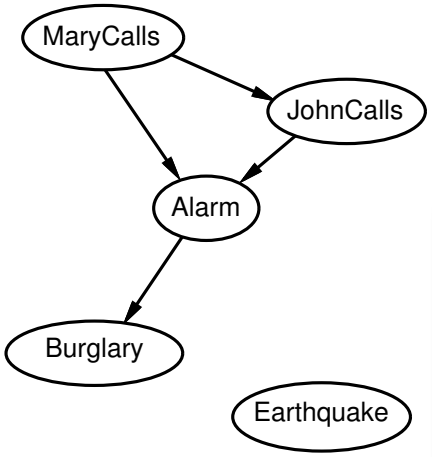
- Suponhamos a seguinte ordenação M, J, A, B, E



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

Exemplo

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

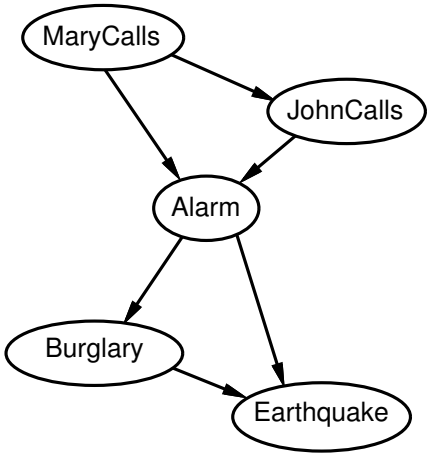


- $P(J|M) = P(J)$? Não
- $P(A|J,M) = P(A|J)?\ P(A|J,M) = P(A)$? Não
- $P(B|A,J,M) = P(B|A)$? Sim
- $P(B|A,J,M) = P(B)$? Não
- $P(E|B,A,J,M) = P(E|A)$?
- $P(E|B,A,J,M) = P(E|A,B)$?

Se o alarme tiver disparado, é mais provável que tenha havido um tremor de terra. Mas se soubermos que houve um roubo, então isso explica o alarme, e a probabilidade de ter havido um tremor de terra é mais baixa...

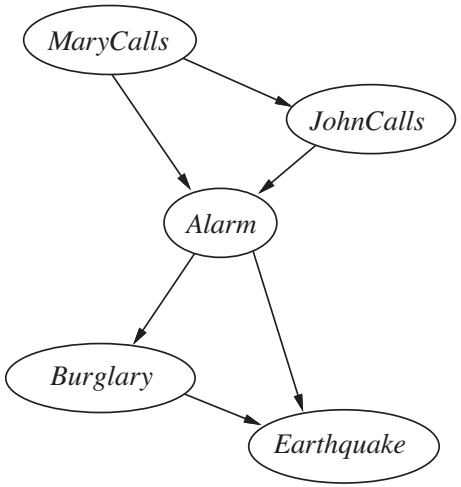
Exemplo

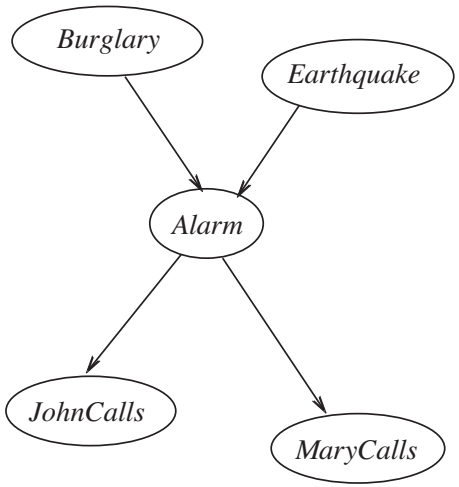
- Suponhamos a seguinte ordenação M, J, A, B, E



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

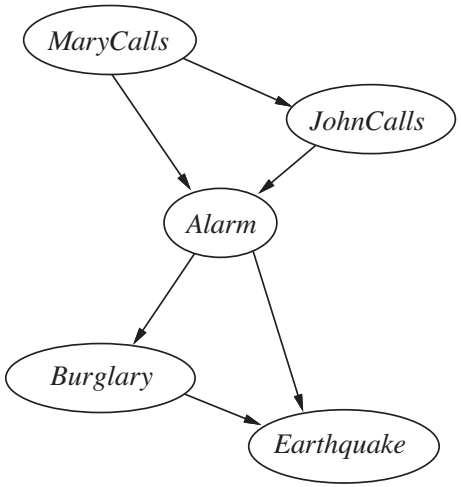
Exemplo

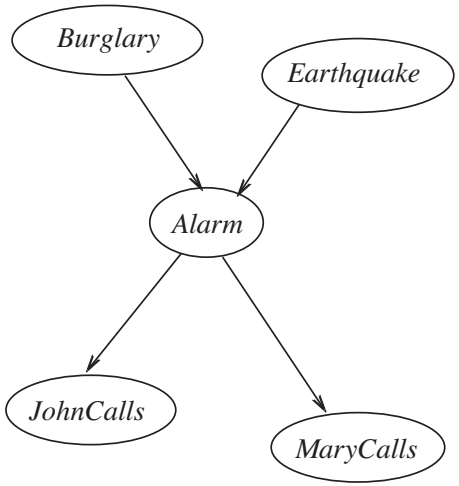


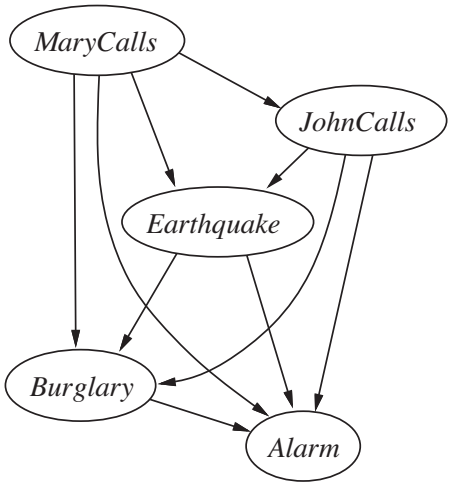


- Decidir sobre independência condicional é difícil nas direções não causais: modelos causais e independência condicional parecem ser inatos nos humanos!
- Determinação de probabilidades condicionais é difícil nas direções não causais.
  - Na rede Bayesiana da esquerda será necessário determinar a probabilidade condicional e.g. de *Earthquake* dado *Alarm* e *Burglary*, que não é fácil nem natural...
- Rede (esquerda) é menos compacta do que a inicial (direita). Necessários
  - Esquerda: 1 + 2 + 4 + 2 + 4 = 13 valores
  - Direita: 1 + 1 + 4 + 2 + 2 = 10 valores

Exemplo



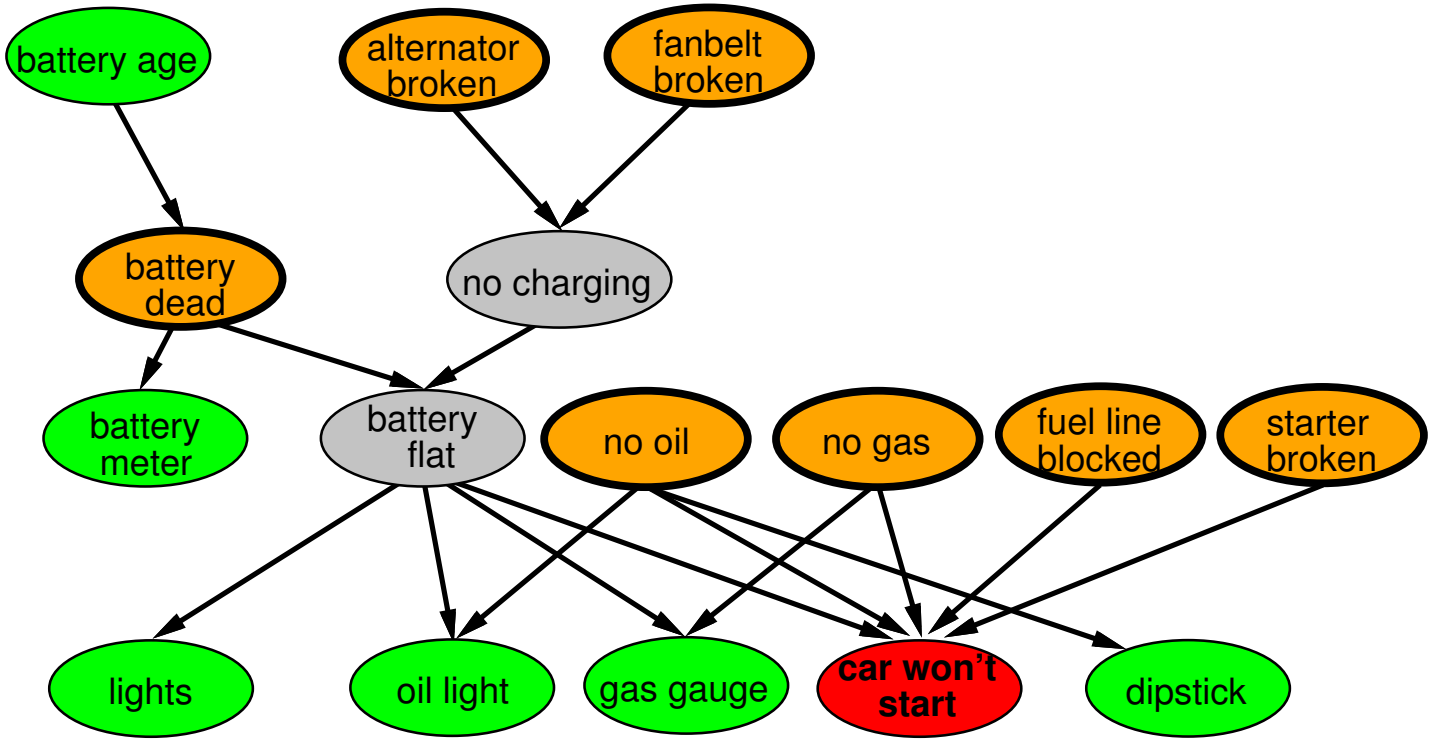




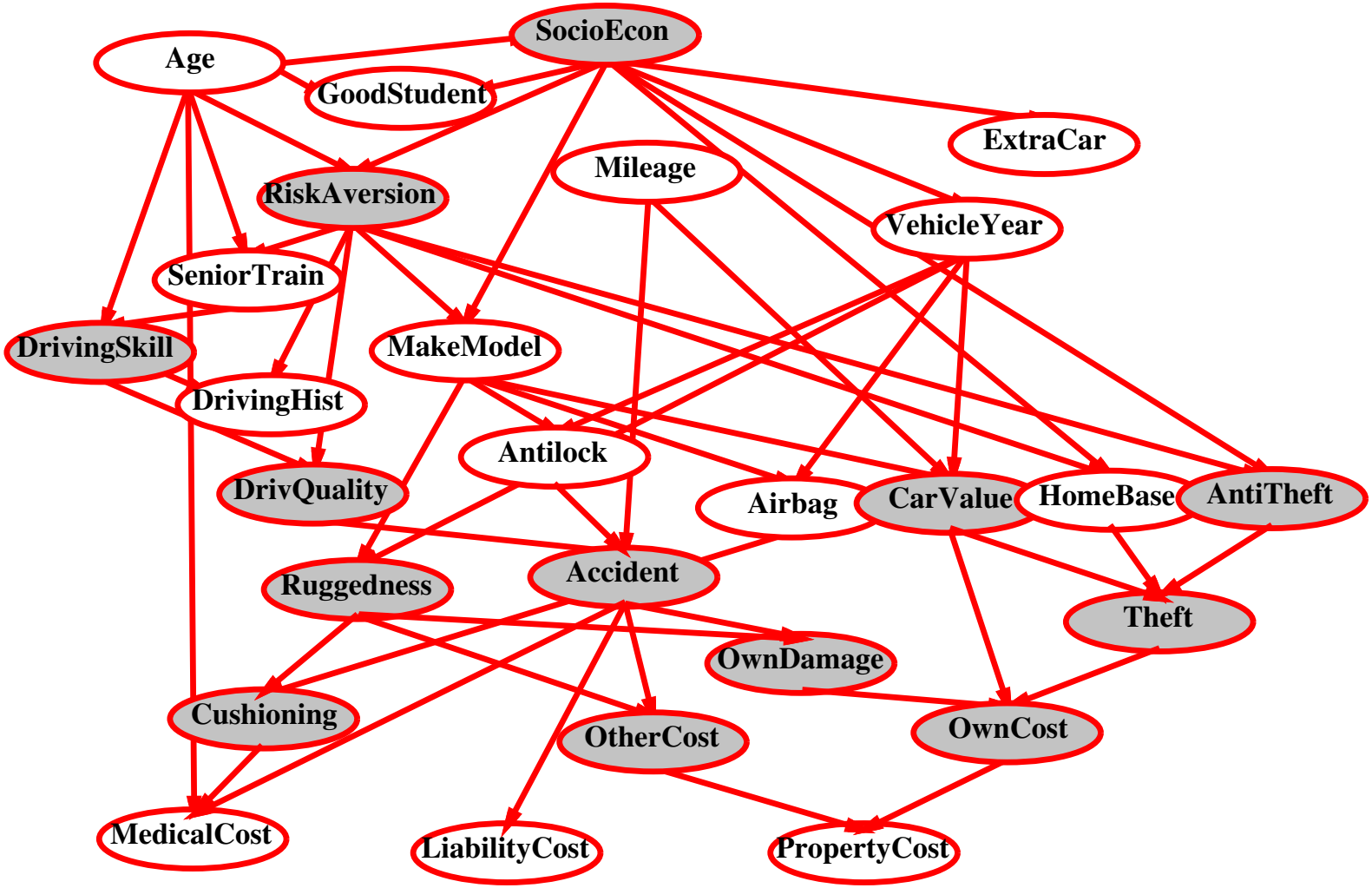
- À direita, uma muito má ordenação das variáveis.
  - Requer a especificação de 31 valores (vs. 13 da da esquerda e 10 da do meio)
- As três redes representam a mesma distribuição conjunta de probabilidade.
- No entanto, apenas a do meio expressa todas as independências condicionais, levando a uma representação mais compacta.

Exemplo: diagnóstico de avarias

- Evidência inicial evidência: carro não pega
- Variáveis observáveis (verde), variáveis “avarias” (laranja)
- Variáveis ocultas (cinzento) garantem estrutura esparsa, reduzem parâmetros



Exemplo: seguro do carro



Distribuições condicionais compactas

- CPT cresce exponencialmente com o numero de pais
- CPT fica infinita com pai ou filho tomando valores contínuos
- Solução: distribuições canónicas que são definidas de forma compacta
- Nos determinísticos são o caso mais simples:
  - $X = f(Parents(X))$ para alguma função f
- E.g., funções Booleanas
  - $NorthAmerican \Leftrightarrow Canadian \vee US \vee Mexican$
- E.g., relações numéricas entre variáveis continuas
  - $\partial Level/\partial t = inflow + precipitation - outflow - evaporation$

Distribuições condicionais compactas

- Distribuições Noisy-OR modelam múltiplas causas que não interagem

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

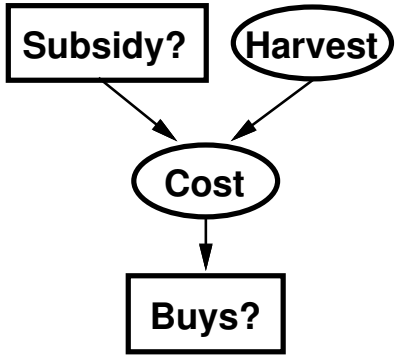
$$P(X|U_1,\ldots,U_j,\neg U_{j+1},\ldots,\neg U_k) = 1-\prod_{i=1..j} q_i$$

Constipação	Gripe	Malária	P(Febre)	P(¬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=0.2×0.1
T	F	F	0.4	**0.6**
T	F	T	0.94	0.06=0.6×0.1
T	T	F	0.88	0.12=0.6×0.2
T	T	T	0.988	0.012=0.6×0.2×0.1

- Número de parâmetros linear no número de pais

Redes Híbridas (discretas + contínuas)

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



- Opção 1: discretização – erros podem ser grandes, CPTs grandes
- Opção 2: famílias canónicas com numero finito de parâmetros
  - Variável contínua, pais discretos+contínuos (e.g., *Cost*)
  - 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.,:

$$P\left(Cost = c \mid Harvest = h, Subsidy? = true\right)$$

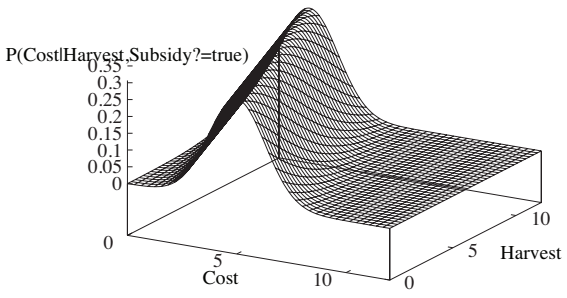
$$= N\left(a_t h + b_t, \sigma_t\right)\left(c\right)$$

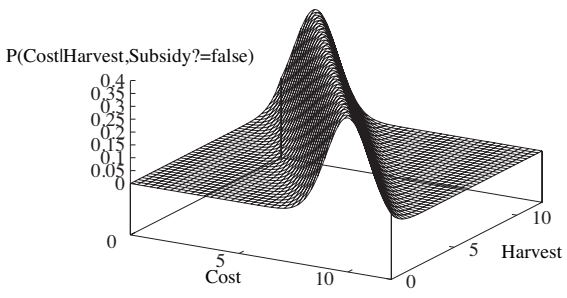
$$= \frac{1}{\sigma_t \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{c - \left(a_t h + b_t\right)}{\sigma_t}\right)^2\right)$$

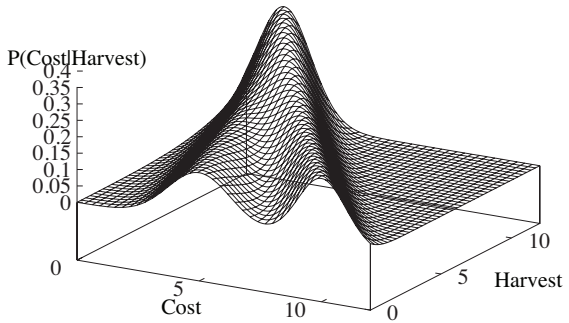
- Média *Cost* varia linearmente com *Harvest*, variância fixa
- Variação linear não é razoável para todo o domínio mas funciona bem se os valores prováveis de *Harvest* estiverem limitados

Variáveis filho contínuas

- Rede só com variáveis contínuas com distribuição Linear Gaussiana (LG)
  - distribuição conjunta tem distribuição Gaussiana multivariada
- Rede discreta+contínua LG é uma rede Gaussiana condicional, i.e., uma Gaussiana multivariada para todas as variaveis contínuas, para cada combinação de valores de variáveis discretas

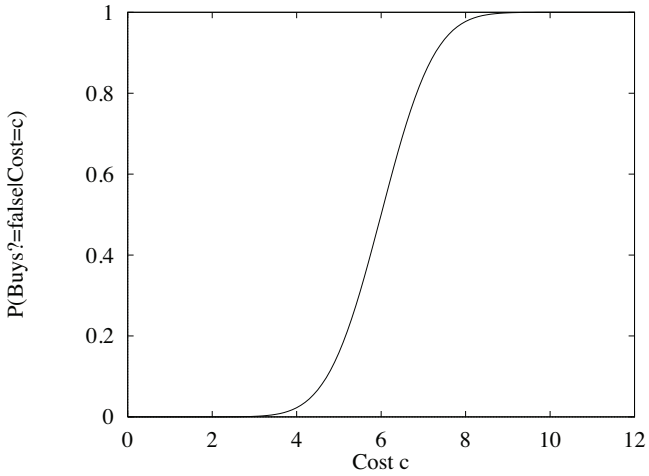






Variáveis discretas com pais contínuos

- Probabilidade de *Buys*? dado *Cost* deve ser um limiar "suave":



- Distribuição Probit utiliza integral de uma Gaussiana:

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

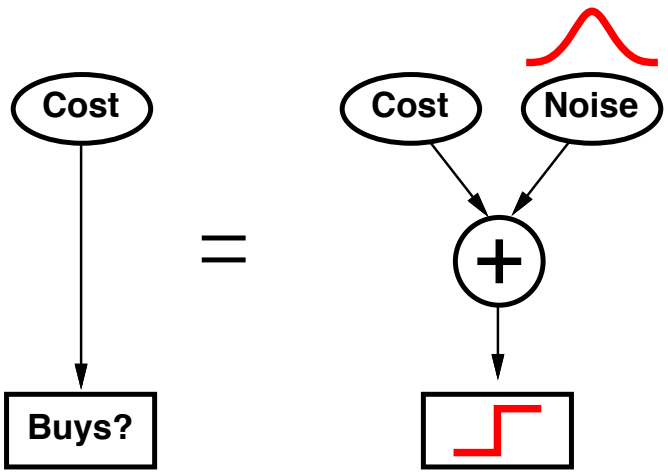
$$P(Buys? = true \mid Cost = c) = \Phi((-c + \mu)/\sigma)$$

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

- 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 esta sujeita a ruído

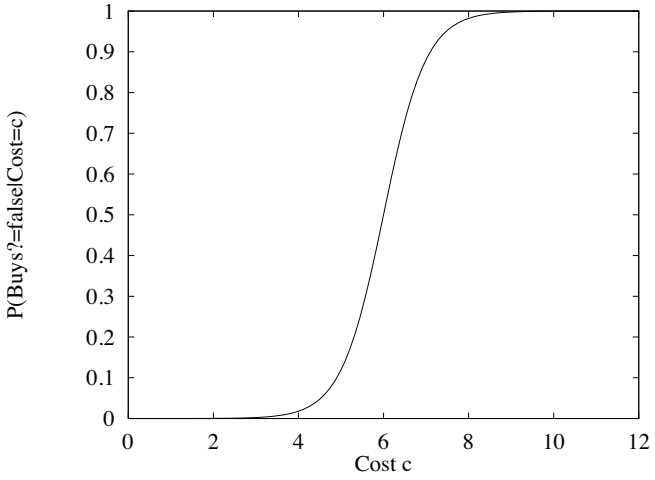


Variável discreta (cont.)

- Distribuição Sigmoide (ou logit) também utilizada em redes neuronais:

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

- Sigmoide tem forma semelhante a 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 continuas ⇒ distribuições parametrizadas (e.g., Gaussiana linear)