Inteligência Artificial

Aula 9

Profª Bianca Zadrozny

http://www.ic.uff.br/~bianca/ia-pos

Raciocínio Probabilístico

Capítulo 14 – Russell & Norvig

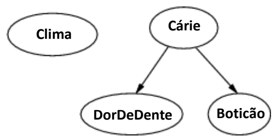
Seções 14.1 a 14.4

Redes Bayesianas

- Estrutura de dados para representar as dependências entre variáveis e fornecer uma especificação **concisa** de *qualquer* distribuição de probabilidade conjunta total.
- Sintaxe:
  - um conjunto de nós, um para cada variável aleatória
  - grafo direcionado e acíclico (seta = "influência direta")
  - cada nó tem uma distribuição condicional **P**(X_i | Pais (X_i)) que quantifica o efeito dos pais sobre o nó
    - tabela de probabilidade condicional (TPC)

Exemplo

- A topologia de uma rede representa relações de independência condicional :

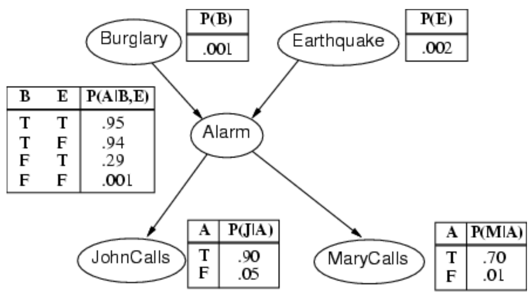


- *Clima* é independente de outras variáveis
- *DorDeDente* e *Boticão* são condicionalmente independentes dado *Cárie*
- Informalmente, a rede representa o fato de que *Cárie* é uma causa direta de DorDeDente e *Boticão*.

Exemplo

- *" Estou no trabalho, meu vizinho João me liga pra dizer que o alarme da minha casa está tocando, mas a minha vizinha Maria não me liga. Às vezes o alarme toca por causa de terremotos leves. Houve um roubo na minha casa? "*
- Variáveis: *Roubo (Burglary)*, *Terremoto (Earthquake)*, *Alarme (Alarm)*, *JoãoLiga (JohnCalls)*, *MariaLiga (MaryCalls)*
- A topologia da rede reflete conhecimento "causal":
  - Um roubo pode ativar o alarme
  - Um terremoto pode ativar o alarme
  - O alarme faz Maria telefonar
  - O alarme faz João telefonar

Exemplo



Da topologia da rede

- Roubos e terremotos afetam diretamente a probabilidade do alarme tocar.
- Mas o fato de João e Maria telefonarem só depende do alarme.
- Desse modo, **a rede representa nossas suposições de que eles não percebem quaisquer roubos diretamente, não notam os terremotos e não verificam antes de ligar!**

As probabilidades...

- ... resumem um conjunto potencialmente infinito de circunstâncias (Maria ouve música alta, João liga quando ouve o telefone tocar; umidade, falta de energia, etc., podem interferir no alarme; João e Maria não estão em casa, etc.
  - **"Preguiça e ignorância"**

Tabelas de probabilidade condicional (TPC)

- Cada linha em uma TPC contém a probabilidade condicional de cada valor de nó para um **caso de condicionamento**;
  - um caso de condicionamento é uma combinação possível de valores para os nós superiores
- Cada linha requer um número p para X_i = *true* (a probabilidade para X_i = *false* é $1-p$)

Tabelas de probabilidade condicional (TPC)

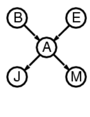
- Um nó sem pais tem apenas uma linha: probabilidade a priori
- Em geral, uma tabela para uma variável booleana com k pais booleanos possui 2^k probabilidades
- Se cada variável não possuir mais do que k pais, a rede completa será $O(n \cdot 2^k)$, para n = número de nós.
  - i.e., cresce linearmente com n, vs. $O(2^n)$ da distribuição total

Semântica das RB

- Duas maneiras equivalentes:
  - **Semântica global** (ou numérica): entender as redes como uma representação da distribuição de probabilidade conjunta.
    - Indica como construir uma rede.
  - **Semântica local** (ou topológica): visualizá-las como uma codificação de uma coleção de declarações de independência condicional.
    - Indica como fazer inferências com uma rede.

Semântica Global

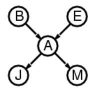
Se uma rede bayesiana for uma representação da distribuição conjunta, então ela poderá ser utilizada para responder qualquer consulta efetuando-se o **produtório** de todas as entradas conjuntas relevantes.



Semântica Global

A semântica global (ou numérica) define a distribuição de probabilidade total como o produto das distribuições condicionais locais:

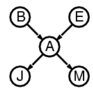
$\mathbf{P}\,(X_1, \ldots, X_n) = \prod_{i=1} \mathbf{P}\,(X_i \mid Pais(X_i))$



Semântica Global

A semântica global (ou numérica) define a distribuição de probabilidade total como o produto das distribuições condicionais locais:

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



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

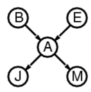
$= \mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}\,(\neg e)$

$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$

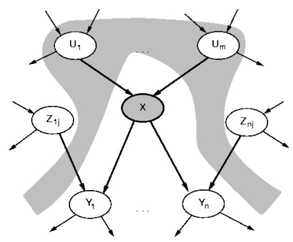
$= 0.00063$

Semântica local

- **Semântica local** (topológica): cada nó é condicionalmente independente de seus não-descendentes, dados seus pais. Ex. **J** é independente de **B** e **E** dado **A**.

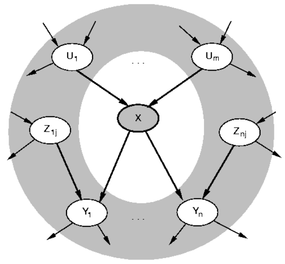


Semântica local



Um nó X é condicionalmente independente de seus não descendentes (Z_{ij}) dados seus pais (U_i).

Semântica local



Um nó X é condicionalmente independente de todos os outros nós dada a sua **cobertura de Markov** (=pais, filhos e pais dos filhos)

Semântica local e global

- A partir dessas asserções sobre a independência condicional e das TPCs, a distribuição conjunta pode ser reconstruída.
  - Deste modo a semântica numérica e topológica são equivalentes.

Construindo uma rede Bayesiana

1. Escolha uma ordem para as variáveis aleatórias $X_1, \dots, X_n$
2. Para i = 1 a n
   - adicione X_i à rede
   - selecione pais para $X_1, \dots, X_{i-1}$ tais que
     $\mathbf{P}(X_i \mid Pais(X_i)) = \mathbf{P}(X_i \mid X_1, \dots X_{i-1})$

Esta escolha de pais garante a semântica global:

$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1})$ (regra da cadeia)

$= \prod_{i=1}^{n} \mathbf{P}(X_i \mid Pais(X_i))$ (por construção)

Ordem para as variáveis

- A ordem correta em que os nós devem ser adicionados consiste em adicionar primeiro as "causas de raiz", depois as variáveis que elas influenciam e assim por diante, até chegarmos às folhas, que não tem nenhuma influência causal direta sobre as outras variáveis.
- E se escolhermos a ordem "errada"?

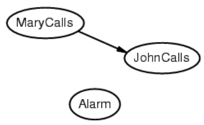
Exemplo

- Assumindo a ordem: *M, J, A, B, E*

$\mathbf{P}(J \mid M) = \mathbf{P}(J)$?

Exemplo

- Assumindo a ordem: *M, J, A, B, E*

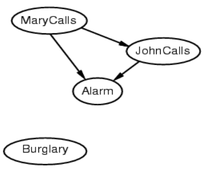


$\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **Não**

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$?

Exemplo

- Assumindo a ordem: *M, J, A, B, E*



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **Não**

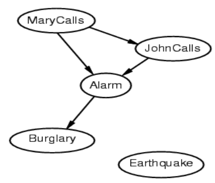
$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **Não**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$?

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$?

Exemplo

- Assumindo a ordem: *M, J, A, B, E*



$\mathbf{P}(J \mid M) = \mathbf{P}(J)$? **Não**

$\mathbf{P}(A \mid J, M) = \mathbf{P}(A \mid J)$? $\mathbf{P}(A \mid J, M) = \mathbf{P}(A)$? **Não**

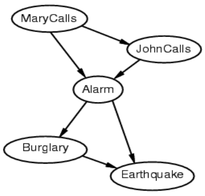
$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B \mid A)$? **Sim**

$\mathbf{P}(B \mid A, J, M) = \mathbf{P}(B)$? **Não**

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A)$?

$\mathbf{P}(E \mid B, A, J, M) = \mathbf{P}(E \mid A, B)$?

Exemplo



P(*J | M*) = ***P***(*J*)? **Não**

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

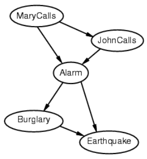
P(*B | A, J, M*) = ***P***(*B | A*)? **Sim**

P(*B | A, J, M*) = ***P***(*B*)? **Não**

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

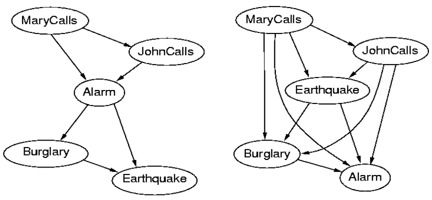
P(*E | B, A, J, M*) = ***P***(*E | A, B*)? **Sim**

Exemplo



- A rede resultante terá dois vínculos a mais que a rede original e exigirá outras probabilidades para serem especificadas
- Alguns dos vínculos apresentam relacionamentos tênues que exigem julgamentos de probabilidade difíceis e antinaturais (prob de *Terremoto*, dados *Roubo* e *Alarme*)
- (Em geral) é melhor pensar de *causas* para *efeitos* (modelo causal) e não do contrário (modelo de diagnóstico)

Exemplo



(a) (b)

- Uma ordenação de nós ruim: *MaryCalls*, *JohnCalls*, *Earthquake*, *Burglary* e *Alarm*
- Entretanto, todas as três redes devem representar a mesma distribuição conjunta. As duas últimas só não expressam todas as independências condicionais

Representação eficiente de distribuições condicionais

- Ainda que o número de pais *k* seja reduzido, o preenchimento da TPC para um nó exige até $O(2^k)$ e muita experiência para decidir os casos condicionais.
  - Esse é o pior caso, em que os relacionamentos entre pais e filhos são arbitrários.
- Em muitos casos podemos utilizar um padrão (***distribuição canônica***) para obter a tabela.

Representação eficiente de distribuições condicionais

- ***Distribuição canônica:***
  - ajustar a distribuição de probabilidades em cada nó a alguma forma padrão.
  - nestes casos a tabela completa pode ser especificada nomeando-se o padrão e fornecendo-se alguns parâmetros.
  - Exemplos:
    - nós determinísticos
    - relacionamentos lógicos ruidosos: *ou-ruidoso*

Representação eficiente: Distribuição canônica

- **Nós determinísticos**: tem seus valores especificados pelos valores de seus pais, sem qualquer incerteza:
  - X = f(Pais(X)) para alguma função f;
  - funções booleanas:
    - Norte Americano $\Leftrightarrow$ Canadense $\vee$ EUA $\vee$ Mexicano
  - relação numérica entre funções contínuas:
    - **pais** *afluentes/escoadouros* **filhos**: Δ*nível da água*
      - Δ*nível da água = afluentes - escoadouros*
    - *Valor mínimo de alguma função*

Representação eficiente: Distribuição canônica

- **Ou-ruidoso**
  - P(¬fever| cold, ¬ flu, ¬ malaria) = 0.6
  - P(¬fever| ¬ cold, flu, ¬ malaria) = 0.2
  - P(¬fever| ¬ cold, ¬ flu, malaria) = 0.1

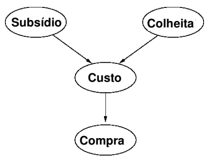
$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	$\mathbf{0.0}$	1.0
F	F	T	0.9	$\mathbf{0.1}$
F	T	F	0.8	$\mathbf{0.2}$
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	$\mathbf{0.6}$
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Representação eficiente: Distribuição canônica

- Ou-ruidoso (em contraste com o ou proposicional)
  - Permite a incerteza sobre a habilidade de cada pai para fazer o filho ser verdadeiro - o relacionamento entre pai e filho pode ser inibido.
    - Todas as causas listadas
    - inibições independentes
    - Assim "febre é falsa se e somente se todos os seus pais verdadeiros são inibidos, e a probabilidade de isso ocorrer é o produto das probabilidades de inibição de cada pai."

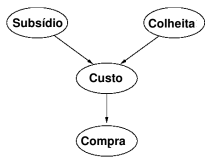
Redes de Bayes Híbridas

- Discretas: *Subsídio?* e *Compra?*



- Dois novos tipos de distribuições condicionais:
  - **variável contínua, com pais contínuos e discretos** (***Custo***)
  - **Variável discreta com pai contínuo (Compra*?*)**

Redes de Bayes Híbridas



- Manipular variáveis contínuas:
  - **Discretização**: repartir os valores possíveis em um conjunto fixo de intervalos
  - **Definir funções de probabilidade padrão** especificadas por um número finito de parâmetros

Variável contínua, com pais contínuos e discretos (*Cost*)

- Para custo: **P**(Custo|Colheita, Subsídio)
  - *O pai discreto (Subsídio) é manipulado por enumeração explícita:*

P(Custo|Colheita, subsídio) e *P(Custo|Colheita, ¬subsídio)*

- Para *Colheita* especificamos como a distribuição sobre o custo **c** depende do valor contínuo **h** de colheita.
  - i.e., os parâmetros da distribuição de custo como função de **h**
  - em geral: **distribuição Gaussiana linear**

Variável contínua, com pais contínuos e discretos *(Custo)*

- **distribuição gaussiana linear:** o filho (*Custo*) tem uma distribuição gaussiana cuja média varia linearmente com o valor do pai (*Colheita*), e cujo desvio padrão é fixo:

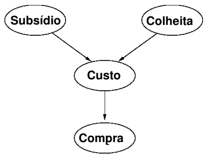
$$
\begin{aligned}
&P(Cost = c | Harvest = h, Subsidy? = true) \\
&= N(a_t h + b_t, \sigma_t)(c) \\
&= \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)
\end{aligned}
$$

- **distribuição gaussiana linear:** o filho (*Custo*) tem uma distribuição gaussiana cuja média varia linearmente com o valor do pai (*Colheita*), e cujo desvio padrão é fixo:

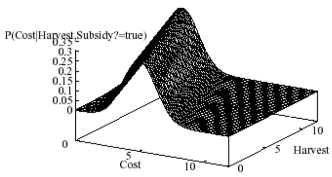
$$P(Cost = c | Harvest = h, Subsidy? = true)$$
$$= N(a_t h + b_t, \sigma_t)(c)$$
$$= \frac{1}{\sigma_t \sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{c-(a_t h + b_t)}{\sigma_t}\right)^2\right)$$

- **distribuição gaussiana linear:** o filho (*Custo*) tem uma distribuição gaussiana cuja média varia linearmente com o valor do pai (*Colheita*), e cujo desvio padrão é fixo:

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



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



A inclinação é negativa, porque o preço diminui à medida que a quantidade oferecida aumenta

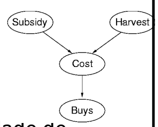
Distribuição gaussiana condicional linear

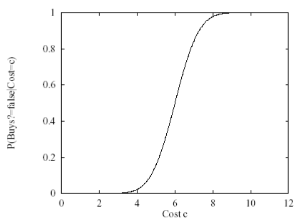
- Uma rede que contém apenas variáveis contínuas com distribuições Gaussianas lineares tem uma distribuição conjunta que é uma **distribuição multivariada** sobre todas as variáveis.
  - superfície em mais de uma dimensão que tem um pico na média (em n dimensões) e decresce para todos os lados.
- Com variáveis discretas (se nenhuma destas é filha de uma var. contínua), a rede define uma **distribuição gaussiana condicional**
  - dada qqr. atribuição às var. discretas, a distribuição sobre as var. contínuas é uma **distribuição gaussiana multivariada.**

variáveis discretas com pais contínuos

- Ex. *Compras*:
  - Podemos supor que o cliente comprará se o preço for baixo e não comprará se for alto e que:
  - A probabilidade de compra varia suavemente em alguma região intermediária
    - A distribuição condicional é semelhante a uma função de limiar "suave" (*soft threshold*)
    - Distribuição **probit** é uma possibilidade...

v.discretas, pais contínuos



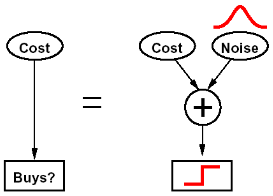


- Probabilidade de *Compra* dado *Custo:* ***limiar suave***
- Distribuição **Probit**:
  - integral da Gaussiana
  - posição precisa do limiar está sujeita a ruído gaussiano

$$\Phi(x) = \int_{-\infty}^{x} N(0,1)(x)dx$$
$$P(Buys? = true \mid Cost = c) = \Phi((-c+\mu)/\sigma)$$

Por que **Probit**?

- Possui mais ou menos o formato desejado
- Pode ser visto como um degrau cuja posição tem ruído.



Inferência Bayesiana

- Dada uma rede bayesiana, queremos computar a distribuição da probabilidade condicional para um conjunto de variáveis de consulta, dado os valores de um conjunto de variáveis de evidência:

P(Variável_consulta|Variáveis_Evidência)

Inferência Exata: Algoritmo de Enumeração

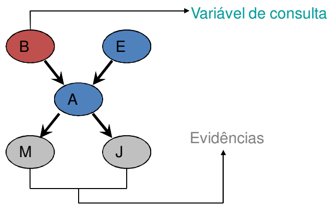
- A idéia básica do algoritmo de Enumeração é avaliar a equação

$$P(X \mid E) = \alpha \sum_{y} P(X, e, y)$$

sem ter que montar explicitamente a tabela de probabilidade conjunta total.

- Apenas, percorrem-se os nós da rede propagando as evidências e extraindo as probabilidades para que sejam feitos os somatórios e multiplicações necessárias.

Algoritmo de Enumeração

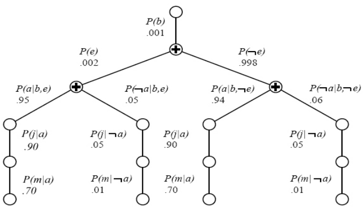


P(B|J,M)?? $\longrightarrow$ $P(B \mid j, m) = \alpha P(B, j, m) = \alpha \sum_{e}\sum_{a} P(B, e, a, j, m)$

$$= \alpha \sum\sum P(B)P(a \mid B, e)P(m \mid a)P(j \mid a)$$

(calcular para cada valor de B (b e ¬b) e normalizar)

Algoritmo de Enumeração



A figura torna explícita as sub-expressões repetidas que são avaliadas pelo algoritmo.
Os produtos P(j | a)P(m | a) e P(j | a)P(m | a) são calculados duas vezes, um para cada valor de *e*.

Algoritmo de Enumeração

```
function ENUMERATION-ASK(X, e, bn) returns a distribution over X
   inputs: X, the query variable
           e, observed values for variables E
           bn, a Bayesian network with variables {X} ∪ E ∪ Y
   Q(X) ← a distribution over X, initially empty
   for each value x_i of X do
       extend e with value x_i for X
       Q(x_i) ← ENUMERATE-ALL(VARS[bn], e)
   return NORMALIZE(Q(X))

function ENUMERATE-ALL(vars, e) returns a real number
   if EMPTY?(vars) then return 1.0
   Y ← FIRST(vars)
   if Y has value y in e
       then return P(y | Pa(Y)) × ENUMERATE-ALL(REST(vars), e)
       else return Σ_y P(y | Pa(Y)) × ENUMERATE-ALL(REST(vars), e_y)
           where e_y is e extended with Y = y
```

Algoritmo de Eliminação de Variáveis

- Elimina os cálculos repetidos do algoritmo de enumeração.
- A idéia é simples: efetuar o cálculo apenas uma vez e guardar os resultados para uso posterior.
- Esta é uma forma de programação dinâmica.

Algoritmo de Eliminação de Variáveis

Problemas do algoritmo de eliminação de variáveis:

- A configuração inicial das variáveis influencia no tempo de execução dos algoritmos.
- Encontrar uma configuração inicial ótima é um problema NP-Completo.