Inteligência Artificial

Luís A. Alexandre

UBI

Ano lectivo 2018-19

Conteúdo

Conteúdo

Introdução

- Vamos ver como se consegue adquirir conhecimento incerto de forma eficiente.
- Depois veremos como fazer inferência probabilística de uma forma eficiente.
- Vamos falar de **redes Bayesianas**.

Introdução

- Vimos na aula anterior que com a distribuição conjunta de probabilidade podemos responder a qualquer questão sobre o domínio.
- O problema é que essa distribuição fica muito difícil de obter com o aumento do número de variáveis.
- Hoje vamos aprender uma nova estrutura de dados: a **rede Bayesiana** (RB), que permite representar a dependência entre variáveis.
- As RBs permitem representar qualquer distribuição conjunta de probabilidade.

Conteúdo

Redes Bayesianas

- Uma RB é um **grafo dirigido** em que cada nodo contém informação probabilística:
  - Cada nodo corresponde a uma variável aleatória (discreta ou contínua).
  - Uma aresta do nodo X para o Y significa que X é o pai de Y.
  - O grafo não contém ciclos.
  - Cada nodo X_i guarda a distribuição probabilística condicional $P(X_i|Pais(X_i))$ que quantifica os efeitos dos pais no nodo em causa.
- Quando temos uma aresta a partir de X e a chegar a Y sabemos que X influencia Y.
- Como as causas estão por detrás dos efeitos, nas RBs, **os nodos causa devem ser pais** dos nodos efeito.

Redes Bayesianas

- Um perito pode facilmente indicar quais são as causas e os respetivos efeitos: permite construir a topologia da RB.
- Depois resta determinar as probabilidades condicionais a colocar em cada nodo.
- Com estes dois elementos, conseguimos achar a distribuição conjunta para todas as variáveis.
- Se um nodo tem m nodos pai que são variáveis booleanas, então a sua probabilidade condicional pode ser representada numa tabela com 2^m entradas.
- Nodos que não estão ligados na RB são nodos **condicionalmente independentes**.

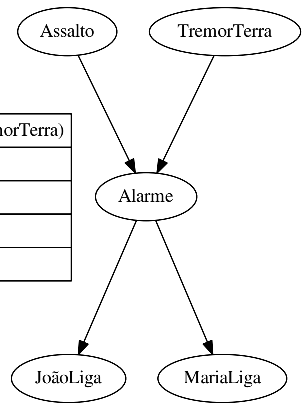
Redes Bayesianas: exemplo

- Vejamos um exemplo: um alarme instalado em casa.
- Deteta bem assaltantes mas por vezes dispara com tremores de terra.
- Existem 2 vizinhos, João e a Maria, que prometeram que nos ligavam se o alarme disparasse.
- O João liga quase sempre quando ouve o alarme mas por vezes confunde o som do telefone a tocar com o do alarme e liga também nesse caso.
- A Maria ouve música alto e, muitas das vezes, acaba por não ouvir o alarme a tocar.
- Dada a informação sobre as chamadas recebidas (ou não) dos dois vizinhos, queremos saber qual a probabilidade de existir um assalto.

Redes Bayesianas: exemplo

P(Assalto)
0.001

P(TremorTerra)
0.002



Assalto	TremorTerra	P(Alarme|Assalto,TremorTerra)
V	V	0.95
V	F	0.94
F	V	0.29
F	F	0.001

Alarme	P(JoãoLiga|Alarme)
V	0.90
F	0.05

Alarme	P(MariaLiga|Alarme)
V	0.70
F	0.01

Adaptada de Russel & Norvig, p.512

Redes Bayesianas: exemplo

- Os nodos *Assalto* e *TremorTerra* só têm associados a probabilidade à priori de ocorrerem visto não dependerem de mais nenhum nodo no nosso problema.
- Os restantes têm a probabilidade de o respetivo acontecimento ter lugar, face aos acontecimentos de que eles dependem.
- Exemplo: a probabilidade do alarme tocar dado que não houve assalto e houve um tremor de terra é de 0.29.
- Os problemas que referimos com o João a confundir o toque do telefone com o toque do alarme e da Maria não ouvir o alarme por causa da música, aparecem condensados nas probabilidades de eles telefonarem dado que o alarme tocou.

Redes Bayesianas: exemplo

- A probabilidade conjunta será obtida com
$$P(x_1, \dots, x_n) = \prod_{i=1}^{n} P(x_i|Pais(X_i)) \tag{1}$$
onde os x_i representam $X_i = x_i$.
- Exemplo: achar a probabilidade de o alarme tocar sem ter havido assalto nem tremor de terra e o João e a Maria ligarem a avisar que o alarme tocou:
$$P(JoãoLiga, MariaLiga, Alarme, \neg Assalto, \neg TremorTerra) =$$
$$P(JoãoLiga|Alarme)P(MariaLiga|Alarme)P(Alarme|\neg Assalto, \neg TremorTerra)P(\neg Assalto)P(\neg TremorTerra) =$$
$$0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.000628$$

Construir Redes Bayesianas

- Vejamos como construir uma RB.
- **Nodos**: determinar o conjunto de variáveis necessárias para modelar o domínio, $\{X_1, \ldots, X_n\}$. Ordenar de modo a que as causas apareçam antes dos efeitos.
- **Arestas**: Para cada i de 1 a n, fazer:
  - Escolher em $\{X_1, \ldots, X_{i-1}\}$ os pais que tornem a seguinte equação válida:
  $$P(X_i|X_1, \ldots, X_{i-1}) = P(X_i|Pais(X_i))$$
  sendo que $Pais(X_i) \subset \{X_1, \ldots, X_{i-1}\}$.
  - Criar uma aresta entre cada pai e X_i.
  - Escrever a $P(X_i|Pais(X_i))$.

Construir Redes Bayesianas

- Exemplo: imaginemos que faltava colocar o nodo *MariaLiga* na figura acima.
- Sabemos que este nodo é influenciado por *Assalto* e *TremorTerra*, mas não diretamente, através apenas de *Alarme*.
- Também sabemos que o *JoãoLiga* não tem influência em *MariaLiga*.
- Isto traduz-se em

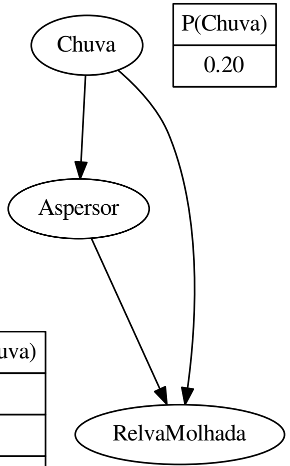
$$P(MariaLiga|JoãoLiga, Alarme, TremorTerra, Assalto) =$$

$$P(MariaLiga|Alarme)$$

- Isto implica que *Alarme* é o nodo pai de *MariaLiga*.
- Uma propriedade importante das RB é que **não têm valores redundantes**, logo não há inconsistências numa RB: nunca são violados os axiomas das probabilidades.

Redes Bayesianas: exercício

- Dada a seguinte RB, cálcule a probabilidade de estar a chover (C) dado que a relva está molhada (R).



P(Chuva)
0.20

Chuva	P(Aspersor|Chuva)
V	0.01
F	0.40

Aspersor	Chuva	P(RelvaMolhada|Aspersor,Chuva)
V	V	0.99
V	F	0.90
F	V	0.80
F	F	0.00

Adaptado de en.wikipedia.org/wiki/File:SimpleBayesNet.svg

Redes Bayesianas: exercício

- Queremos saber $P(C|R)$.
- Sabemos que
$$P(C|R) = \frac{P(C, R)}{P(R)}$$
- Precisamos então de obter $P(C, R)$ e $P(R)$.
- Da equação (1) sabemos que
$$P(C, A, R) = P(C)P(A|C)P(R|A, C)$$
- $P(C, R) = \sum_{A \in \{V,F\}} P(C, A, R) = P(C, A, R) + P(C, \neg A, R) =$
$$0.2 \times 0.01 \times 0.99 + 0.2 \times 0.99 \times 0.8 = 0.1604$$

Redes Bayesianas: exercício

- Também sabemos que

$$P(R) = \sum_{A,C \in \{V,F\}} P(R, A, C) =$$

$$P(R, A, C) + P(R, \neg A, C) + P(R, A, \neg C) + P(R, \neg A, \neg C)$$

- $P(R, A, C) = P(R|A, C)P(A|C)P(C) = 0.99 \times 0.01 \times 0.2 = 0.00198$
- $P(R, \neg A, C) = P(R|\neg A, C)P(\neg A|C)P(C) = 0.8 \times 0.99 \times 0.2 = 0.1584$
- $P(R, A, \neg C) = P(R|A, \neg C)P(A|\neg C)P(\neg C) = 0.9 \times 0.4 \times 0.8 = 0.288$
- $P(R, \neg A, \neg C) = P(R|\neg A, \neg C)P(\neg A|\neg C)P(\neg C) = 0$
- Temos então $P(R) = 0.44838$.
- Finalmente, $P(C|R) = 0.1604/0.44838 = 0.3577$.

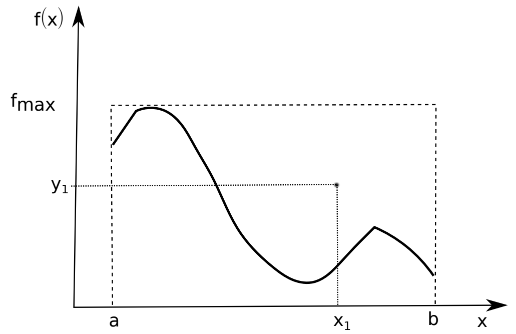
Conteúdo

Inferência Aproximada

- Dado que para casos reais o cálculo da **inferência exata** não é possível devido ao custo computacional, usa-se em geral **inferência aproximada**.
- Chamam-se **algoritmos de Monte Carlo** a métodos de amostragem aleatória, que fornecem aproximações baseadas na geração de muitas amostras aleatórias.
- Estamos interessados em algoritmos que permitam estimar probabilidades à posteriori.
- Vamos ver primeiro um exemplo simples de uso do método de Monte Carlo e depois a inferência baseada em cadeias de Markov.

Método de Monte Carlo

- Vejamos como usar o método de MC para determinar uma quantidade determinística, o valor de um integral.
- Consideremos a seguinte função



- Queremos achar o valor de $I = \int_a^b f(x)dx$.
- A área do rectângulo tracejado é $A = (b - a) \times f_{max}$.
- A probabilidade escolher um ponto ao acaso dentro do retângulo e ele pertencer a zona por baixo da função $f(x)$ é $P = I/A$.

Método de Monte Carlo

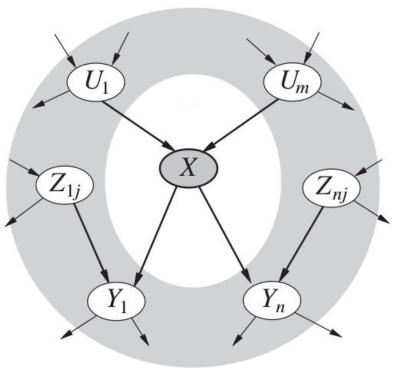
- Podemos determinar P experimentalmente, gerando muitos pontos aleatórios e verificando se estão ou não na área debaixo da curva:
  - 1- Criar um contador, $c = 0$.
  - 2- Repetir n vezes os passos 3 a 5:
  - 3- Escolher 2 valores aleatórios de uma distribuição uniforme entre $[0, 1]$, v_1 e v_2.
  - 4- Criar um ponto no retângulo com as coordenadas $x_1 = a + v_1 \times (b - a)$ e $y_1 = v_2 \times f_{max}$.
  - 5- Se $y_1 \leq f(x_1)$ incrementar o contador c.
  - 6- A estimativa fica $P = c/n$.
- Quando tivermos achado P, torna-se imediato achar o valor do integral: $I = P \times A$.

Amostragem por cadeias de Markov

- Vejamos então como usar as cadeias de Markov para estimar probabilidades à posteriori, no âmbito das redes Bayesianas.
- A familia de algoritmos de amostragem por cadeias de Markov, normalmente chamados métodos MCMC (Markov Chain Monte Carlo) geram cada nova amostra fazendo uma alteração aleatória na amostra anterior.
- Há vários algoritmos dentro dos MCMC, mas vamos ver apenas um: a Amostragem de Gibbs.

Amostragem de Gibbs

- Comecemos por definir o **lençol de Markov** (LM) de uma variável aleatória: os seus pais, filhos e os outros pais dos seus filhos.



Russel & Norvig, p.518

- A distribuição de probabilidade de um nodo só depende da distribuição conjunta das variáveis no seu LM.

Amostragem de Gibbs

- Para obtermos uma amostra para X_i vamos levar em conta apenas os valores das variáveis que pertencem ao seu LM.
- O que a AG faz é:
  - partir de um estado em que as variáveis observadas têm os seus valores fixos.
  - gerar um novo estado mudando de forma aleatória o valor das restantes variáveis.
  - Em cada iteração:
    - Escolher uma variável X
    - Calcular P(X=verdade | as restantes variáveis)
    - Atribuir a X ser verdade a probabilidade calculada
  - Repetir o processo muitas vezes.
  - A frequência com que uma variável é verdade é a sua probabilidade à posteriori.
  - Este algoritmo converge para as verdadeiras probabilidades quando esses valores deixam de variar significativamente.

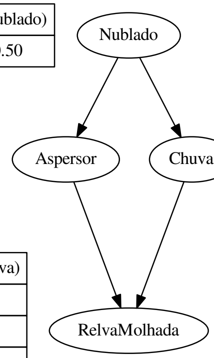
Amostragem de Gibbs

- Como achar P(X=verdade | as restantes variáveis)?
- Usamos o LM: sabemos que cada variável só depende dos pais, filhos e outros pais dos seus filhos.
- Então o problema passa a ser achar P(X=verdade | LM(X)).

Amostragem de Gibbs: exemplo

- Imaginemos que estamos a estimar probabilidades no seguinte problema, e que a ordem topológica das variáveis é definida como [*Nublado*, *Aspersor*, *Chuva*, *RelvaMolhada*].

P(Nublado)
0.50



Nublado	P(Aspersor|Nublado)
V	0.10
F	0.50

Nublado	P(Chuva|Nublado)
V	0.80
F	0.20

Aspersor	Chuva	P(RelvaMolhada|Aspersor,Chuva)
V	V	0.99
V	F	0.90
F	V	0.90
F	F	0.00

Adaptada de Russel & Norvig, p.529

Amostragem de Gibbs: exemplo

- Queremos achar $P(Chuva|Aspersor, RelvaMolhada)$.
- Qual é o LM de *Chuva*?
- As variáveis *Aspersor* e *RelvaMolhada* ficam fixas com os valores que foram observados, ou seja, as variáveis [*Nublado*, *Aspersor*, *Chuva*, *RelvaMolhada*] ficam inicialmente com os valores [?,V,?,V].
- Depois inicializamos aleatóriamente as restantes variáveis do LM, logo poderíamos obter o estado inicial $[V, V, F, V]$.
- Agora a AG o que faz é amostrar as variáveis não observadas (*Nublado* e *Chuva*), repetidamente.

Amostragem de Gibbs: exemplo

- Exemplo:
  - *Nublado* é amostrado, dados os valores do seu LM, ou seja, $P(Nublado|Aspersor, \neg Chuva)$. Se $Nublado = F$ então o novo estado é $[F, V, F, V]$.
  - *Chuva* é amostrado, dados os valores do seu LM, ou seja, $P(Chuva|\neg Nublado, Aspersor, RelvaMolhada)$. Se $Chuva = V$ então o novo estado é $[F, V, V, V]$.
- Cada estado visitado durante este processo é uma amostra que permite estimar a questão original sobre a variável *Chuva*.
- Se este processo resultar em 20 estados com $Chuva = V$ e 60 com $Chuva = F$ então a resposta será que a $P(Chuva|Aspersor, RelvaMolhada) = 0.25$.

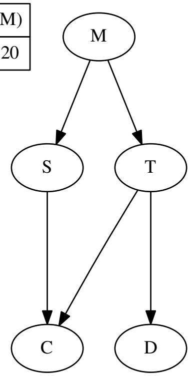
RB: exercício

- O cancro metastático (M) é uma possível causa de tumores no cérebro (T) e também aumenta o cálcio serum (S). Tanto um tumor cerebral como o aumento do cálcio serum podem explicar um coma (C). Sabemos também que os tumores cerebrais podem causar dores de cabeça fortes (D).
- Represente esta informação numa RB (sem atribuição de probabilidades).
- Dada a seguinte RB para o problema descrito, ache o valor da probabilidade de um doente ter um tumor cerebral sabendo que tem fortes dores de cabeça.

P(M)
0.20

M	P(S|M)
V	0.80
F	0.20

M	P(T|M)
V	0.20
F	0.05



S	T	P(C|S,T)
V	V	0.80
V	F	0.80
F	V	0.80
F	F	0.05

T	P(D|T)
V	0.80
F	0.60

Adaptado de Bayesian AI, 2nd edition, p.44

- A **probabilidade conjunta** de n eventos:

$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i|A_1, \ldots, A_{i-1}) \tag{2}$$

- A **probabilidade conjunta** de n **eventos independentes** é o produto das suas probabilidades:

$$P(A_1, A_2, \ldots, A_n) = \prod_{i=1}^{n} P(A_i) \tag{3}$$

- **Probabilidade condicional** relativa a duas variáveis A e B:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{4}$$

- Se Z representa o conjunto de todos os valores das variáveis envolvidas no problema, então podemos obter $P(A)$ **marginalizando**:

$$P(A) = \sum_{B \in Z} P(A,B) \tag{5}$$

- Se as probabilidades são condicionais em vez de conjuntas (aplicar a equação (4) à anterior):

$$P(A) = \sum_{B \in Z} P(A|B)P(B) \tag{6}$$

- **Regra de Bayes**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{7}$$

- Se duas variáveis A e B são **independentes** então

$$P(A|B) = P(A) \text{ e } P(B|A) = P(B) \text{ e } P(A,B) = P(A)P(B) \tag{8}$$

- Se duas variáveis A e B são **independentes condicionalmente**, dada a variável C então

$$P(A,B|C) = P(A|C)P(B|C) \tag{9}$$

- Cada nodo duma RB guarda a seguinte probabilidade condicional:

$$P(A|Pais(A)) \tag{10}$$

- Numa RB, a **probabilidade conjunta** obtém-se através do produto das distribuições de todos os nodos.

Leitura recomendada

- Russell e Norvig, cap. 14.