Aprendizagem Bayesiana

Francisco Carvalho
DI-UFPE

Métodos Bayesianos

Fornece algoritmos práticos de aprendizagem

- Aprendizagem Bayesiana ingenua
- Aprendizagem de Redes Bayesianas
- Combina conhecimento a priori (probabilidade a priori ou incondicional) com dados de observação
- Requer probabilidades à priori

Teorema de Bayes

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)}$$

- P(h): probabilidade a priori da hipótese h
- P(D): probabilidade a priori dos dados de treinamento D
- P(h/D): probabilidade de h dado D
- P(D/h): probabilidade de D dado h

Escolha de hipóteses

- Geralmente deseja-se a hipótese mais provável observados os dados de treinamento

- Hipótese de maior probabilidade a posteriori h_{MAP}

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(h/D) \\
&= \arg\max_{h \in H} \frac{P(D/h)P(h)}{P(D)} \\
&= \arg\max_{h \in H} P(D/h)P(h)
\end{aligned}
$$

- Hipótese de máxima verossimilhança h_{ML}

$$h_{ML} = \arg\max_{h_i \in H} P(D/h_i)$$

Aplicação do Teorema de Bayes: Diagnóstico Médico

- Seja

M=doença meningite

S= dor de cabeça

- Um Doutor sabe:

P(S/M)=0.5

P(M)=1/50000

P(S)=1/20

$$\mathbf{P}(M/S)=\frac{\mathbf{P}(S/M)\mathbf{P}(M)}{\mathbf{P}(S)}$$

$$=\frac{0,5*(1/50000)}{1/20}=0,002$$

- A probabilidade de uma pessoa ter meningite dado que ela está com dor de cabeça é 0,02% ou ainda 1 em 5000.

Fórmulas Básicas de Probabilidade

- Regra do Produto: Probabilidade de uma conjunção de dois eventos *A* e *B*

$$P(A \wedge B) = P(A/B)P(B) = P(B/A)P(A)$$

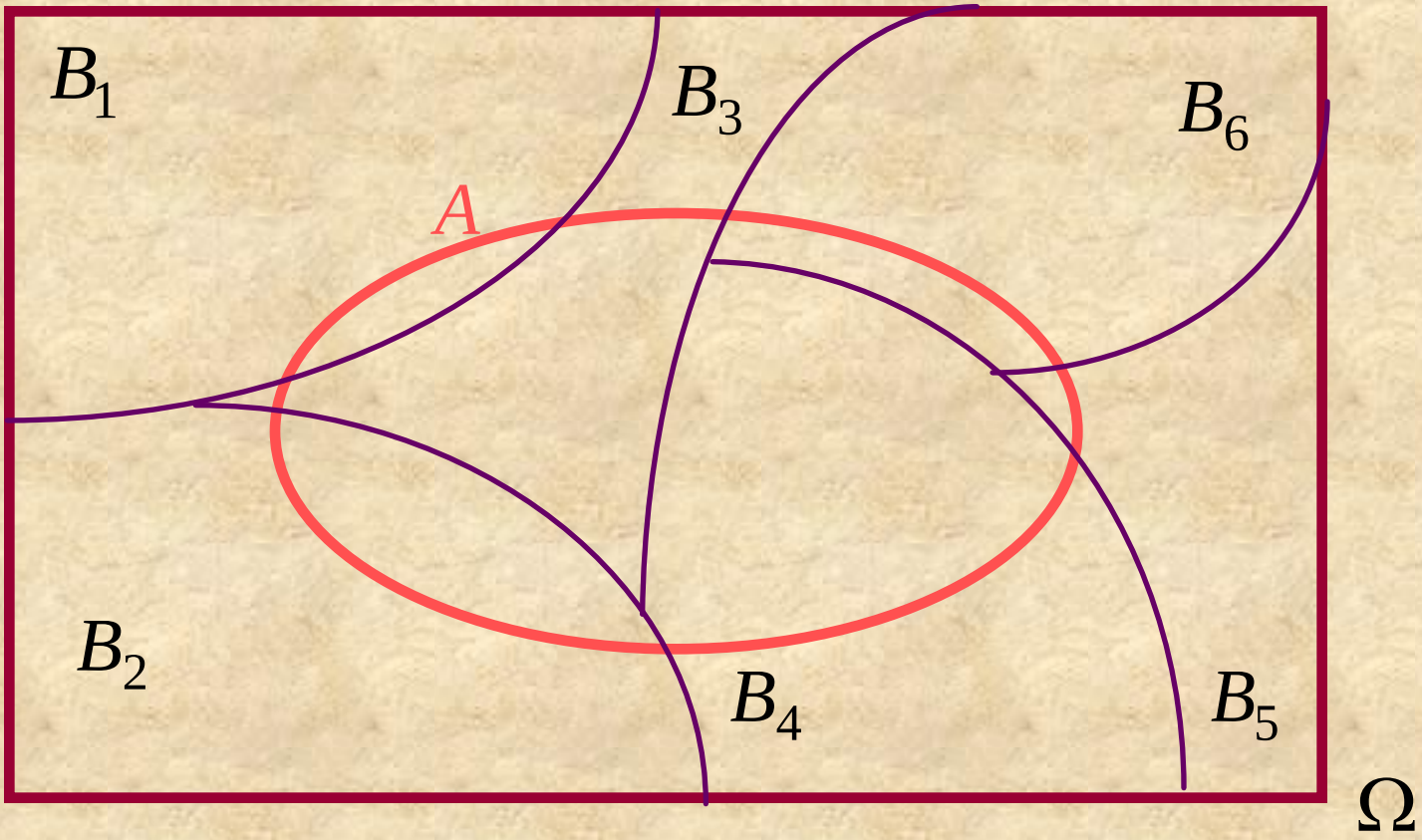
- Regra da Soma: Probabilidade de uma disjunção de dois eventos *A* e *B*

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- Teorema da Probabilidade Total: Se os eventos $A_1,\ldots,A_n$ são mutuamente exclusivos e formam uma partição do evento certo

$$P(B) = \sum_{i=1}^{n} P(B/A_i)P(A_i)$$

Teorema da Probabilidade Total



$$P(A) = \sum_k P(A/B_k)P(B_k)$$

Teorema da Multiplicação de Probabilidades

$$P(A_1 \cap \cdots \cap A_n) = P(A_n / A_1 \cap \cdots \cap A_{n-1}) P(A_1 \cap \cdots \cap A_{n-1})$$

$$\Downarrow$$

$$P(A_1 \cap \cdots \cap A_n) = P(A_n / A_1 \cap \cdots \cap A_{n-1}) \cdots P(A_2 / A_1) P(A_1)$$

Esse resultado permite calcular a probabilidade de ocorrência simultânea de vários eventos a partir das probabilidades condicionais.

Algoritmo de aprendizagem da Probabilidade Máxima à Posteriori - MAP

1. Para cada hipótese $h \in H$, calcule a probabilidade a posteriori

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)}$$

2. Escolha a hipótese h_{MAP} de maior probabilidade à posteriori

$$h_{MAP} = \underset{h \in H}{\arg\max}\, P(h/D)$$

Classificação mais provável de uma nova instância

- Dada uma nova instância x, qual é a sua classificação mais provável?
  - $h_{MAP}(x)$ não é a classificação mais provável

Considere,

- Três hipóteses:
  - $P(h_1/D) = 0.4$, $P(h_2/D) = 0.3$ e $P(h_3/D) = 0.3$
- Dada uma nova instância x,
  - Suponha: $h_1(x) = +$, $h_2(x) = -$ e $h_3(x) = -$
- A classificação mais provável de x: -

Classificador Bayesiano Ótimo

- Um novo exemplo pode ser classificado como $v_j \in V$, a probabilidade de que a classificação correta seja v_j

$$P(v_j / D) = \sum_{h_i \in H} P(v_j / h_i) P(h_i / D)$$

Classificação Bayesiana ótima

$$\underset{v_j \in V}{\mathbf{argmax}} \sum_{h_i \in H} P(v_j / h_i) P(h_i / D)$$

Classificador Bayesiano Ótimo

- Exemplo.
  - $P(h_1/D) = 0.4$, $P(-/h_1) = 0$, $P(+/h_1) = 1$
  - $P(h_2/D) = 0.3$, $P(-/h_1) = 1$, $P(+/h_1) = 0$
  - $P(h_3/D) = 0.3$, $P(-/h_3) = 1$, $P(+/h_1) = 0$

- Portanto $$\sum_{h_i \in H} P(+/h_i)P(h_i/D) = 0.4$$

$$\sum_{h_i \in H} P(-/h_i)P(h_i/D) = 0.6$$

  - e $$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j/h_i)P(h_i/D) = -$$

Classificador Bayesiano Ingênuo

- Junto com as árvores de decisão, redes neurais, vizinhos mútuos, um dos métodos de aprendizagem mais práticos

- Quando usa-lo
  - Quando disponível um conjunto de treinamento médio ou grande
  - Os atributos que descrevem as instâncias forem condicionalmente independentes dada uma classificação

Aplicações de Sucesso
  - Diagnóstico
  - Classificação de documentos textuais

Classificador Bayesiano Ingênuo

- Suponha uma função de classificação f: X → V, onde cada instância x é descrita pelos atributos $\{a_1, \ldots, a_n\}$
- O valor mais provável de f(x) é

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j / a_1, \cdots, a_n)$$

$$= \arg\max_{v_j \in V} \frac{P(a_1, \cdots, a_n / v_j) P(v_j)}{P(a_1, \cdots, a_n)}$$

$$= \arg\max_{v_j \in V} P(a_1, \cdots, a_n / v_j) P(v_j)$$

- Suposição Bayesiana Ingênua $P(a_1, \cdots, a_n / v_j) = \prod_i P(a_i / v_j)$
- Classificador Bayesiano Ingênuo

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i / v_j)$$

Algoritmo Bayesiano Ingênuo

- Aprendizagem_Bayesiana_Ingênua(exemplos)
  - Para cada v_j
    - $P'(v_j) \leftarrow$ estimativa de $P(v_j)$
    - Para cada valor a_i de cada atributo a
      - $P'(a_i/v_j) \leftarrow$ estimativa de $P(a_i/v_j)$
- Classificador_Novas_Instancias(x)

$$v_{NB} = \underset{v_j \in V}{\arg\max}\, P'(v_j) \prod_{a_i \in x} P'(a_i / v_j)$$

Classificador bayesiano ingênuo: exemplo

Dia	Tempo	Temp.	Humid.	Vento	Jogar
D1	Sol	Quente	Alta	Fraco	Não
D2	So	Quente	Alta	Forte	Não
D3	Coberto	Quente	Alta	Fraco	Sim
D4	Chuva	Normal	Alta	Fraco	Sim
D5	Chuva	Frio	Normal	Fraco	Não
D6	Chuva	Frio	Normal	Forte	Não
D7	Coberto	Frio	Normal	Forte	Sim
D8	Sol	Normal	Alta	Fraco	Não
D9	Sol	Frio	Normal	Fraco	Sim
D10	Chuva	Normal	Normal	Fraco	Sim
D11	Sol	Frio	Alta	Forte	?

- P(Sim) = 5/10 = 0.5
- P(Não) = 5/10 = 0.5
- P(Sol/Sim) = 1/5 = 0.2
- P(Sol/Não) = 3/5 = 0.6
- P(Frio/Sim) = 2/5 = 0.4
- P(Frio/Não) = 2/5 = 0.4
- P(Alta/Sim) = 2/5 = 0.4
- P(Alta/Não) = 3/5 = 0.6
- P(Forte/Sim) = 1/5 = 0.2
- P(Forte/Não) = 2/5 = 0.4
- P(Sim)P(Sol/Sim) P(Frio/Sim)
- P(Alta/Sim) P(Forte/Sim) = 0.0032
- P(Não)P(Sol/Não)P(Frio/Não)
- P(Alta/Não) P(Forte/Não) = 0.0288
- $\Rightarrow$ **Jogar_Tenis(D11) = Não**

Algoritmo Bayesiano Ingênuo : Dificuldades

- Suposição de independência condicional quase sempre violada

$$P(a_1,\cdots,a_n / v_j) = \prod_i P(a_i / v_j)$$

  - Mas funciona surpreendentemente bem

- O que acontece se nenhuma das instancias classificadas como v_j tiver o valor a_i?

$$P'(a_i / v_j) = 0 \Rightarrow P'(v_j)\prod_i P'(a_i / v_j) = 0$$

Algoritmo Bayesiano Ingênuo : Dificuldades

- Solução típica

$$P'(a_i / v_j) \leftarrow \frac{n_c + m \times p}{n + m}$$

- Número de exemplos para os quais v = v_j
- N_c número de exemplos para os quais v = v_j e a = a_i
- P é a estimativa à priori para $P'(a_i/v_j)$
- M é o peso dado à priori (número de exemplos "virtuais")

Redes Bayesianas

- Interesse
  - Suposição de independência condicional muito restritiva
  - Mas sem esse tipo de suposição em algum nível o problema se torna intratável

- Redes Bayesianas descrevem independência condicional entre subconjuntos de variáveis
  - Permite a combinação do conhecimento a priori sobre a independência entre variáveis com os dados observados

Independência Condicional

- X é condicionalmente independente de Y dado Z se a distribuição de probabilidade de X é independente do valor de Y dado o valor de Z

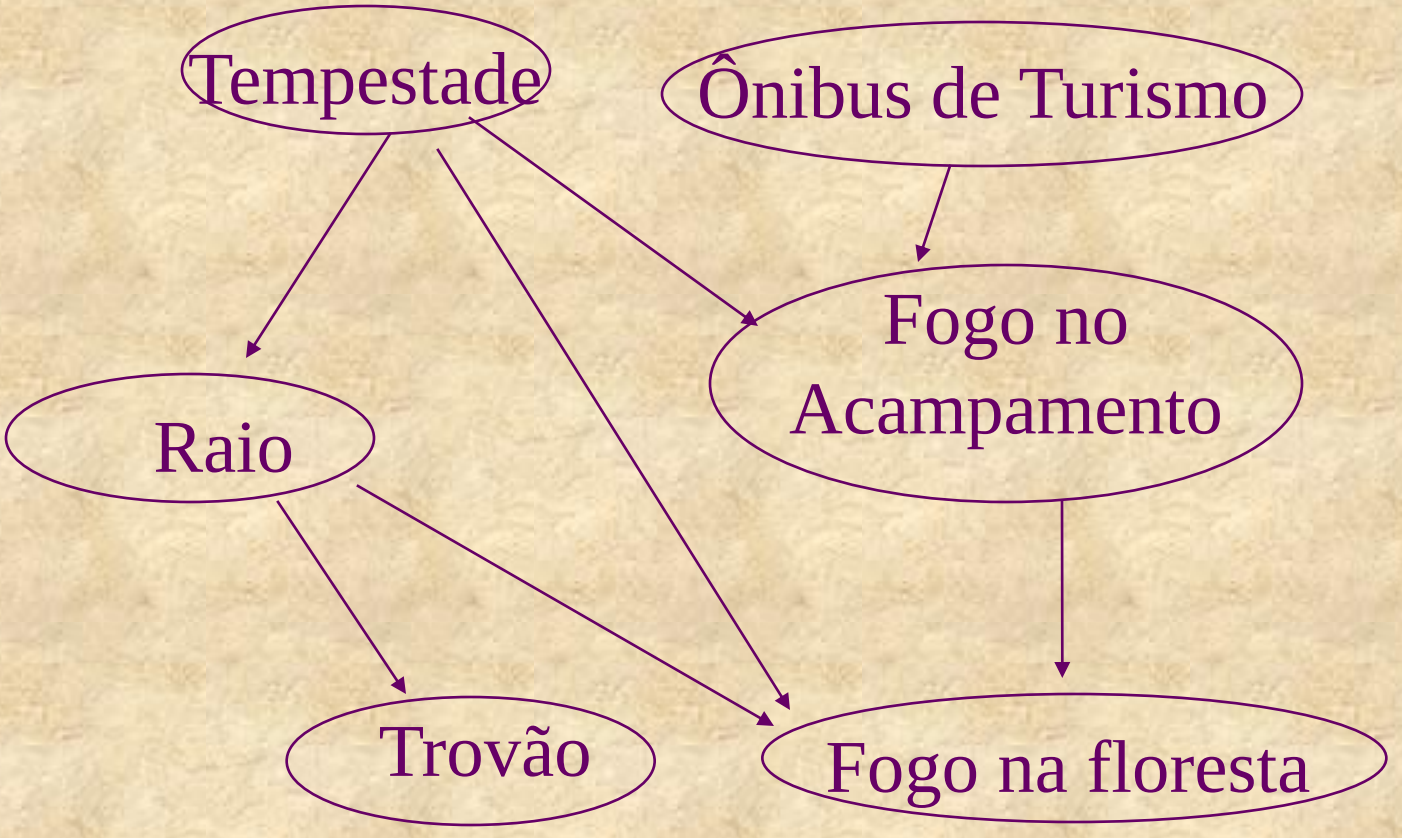
$$(\forall x_i, y_i, z_i) P(X = x_i / Y = y_j, Z = z_k) = P(X = x_i / Z = z_k)$$

$$P(X / Y, Z) = P(X / Z)$$

- Exemplo: Trovão é condicionalmente independente de Chuva, dado Relâmpago
  - P(Trovão/ Chuva, Relâmpago) = P(Trovão/ Relâmpago)
  - Regra do Produto:

$$\begin{aligned} P(X, Y / Z) &= P(X / Y, Z) P(Y / Z) \\ &= P(X / Z) P(Y / Z) \end{aligned}$$

Redes Bayesianas



Redes Bayesianas

	T,O	T,¬O	¬T, O	¬T, ¬O
FC	0.4	0.1	0.8	0.2
¬FC	0.6	0.9	0.2	0.8

Fogo no Acampamento

- A rede representa um conjunto de asserções de independência condicional
  - Cada nó é condicionalmente independente dos seus não descendentes, dados os seus predecessores imediatos
  - Grafo acíclico direto

Redes Bayesianas

- Representa a distribuição de probabilidade conjunta entre todas as variáveis
  - Exemplo, P(Tempestade, …, Fogo na Floresta)
  - Em geral

$$P(y_1, \cdots, y_n) = \prod_{i=1}^{n} P(y_i / \text{Predecessores}(Y_i))$$

Onde Predecessores(Y_i) significa predecessores imediatos de Y_i no grafo

- A distribuição conjunta é definida pelo gafo mais os $P(y_i / \text{Predecessores}(Y_i))$

Redes Bayesianas: representação do conhecimento para raciocínio com incerteza

- Representa 3 tipos de conhecimento do domínio:
  - relações de independência entre variáveis aleatórias (graficamente);
  - probabilidades *a priori* de algumas variáveis;
  - probabilidades condicionais entre variáveis dependentes.
- Permite calcular eficientemente:
  - probabilidades *a posteriori* de qualquer variável aleatória(inferência);
  - usando para isso uma definição recursiva do teorema de Bayes.
- Conhecimento representado:
  - pode ser aprendido a partir de exemplos
  - reutilizando parte dos mecanismos de raciocínio

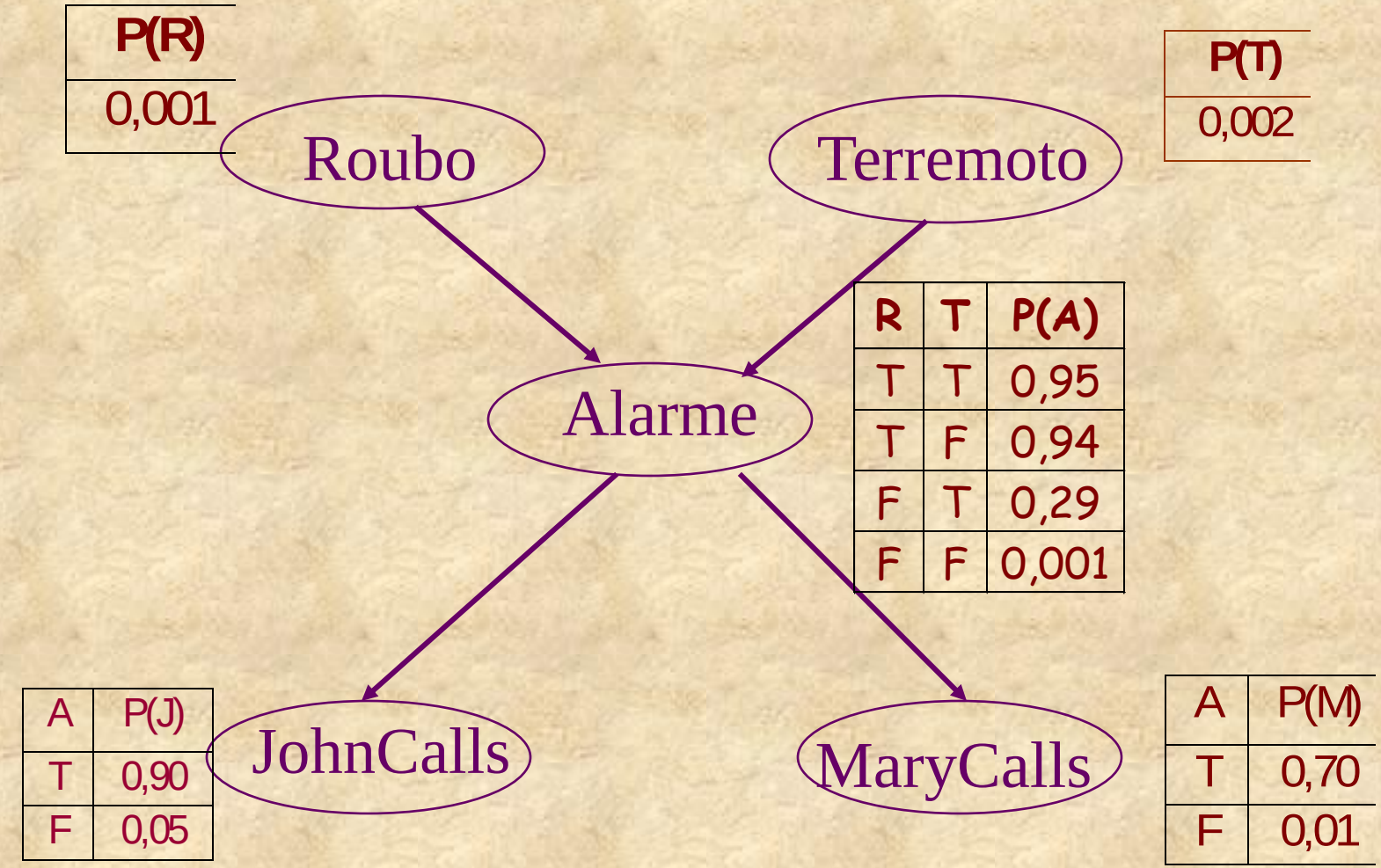
Estrutura de uma rede bayesiana

- Cada variável aleatória (VA) é representada por um nó da rede
- Cada nó (VA) recebe conexões dos nós que têm influência direta (seus pais) sobre ele. (Tarefa fácil para o especialista)
- Cada nó possui uma tabela de Probabilidades Condicionais que quantifica a influência dos seus pais sobre ele. (Difícil para o especialista)
- O grafo é acíclico

Construção (manual) de uma rede bayesiana

- Escolher variáveis **relevantes** que descrevam o domínio;
- Escolher uma ordem para as variáveis;
- Enquanto tiver variáveis sobrando:
  - pegar uma variável e adicionar um nó na rede para ela;
  - criar links dos nós anteriormente inseridos que satisfaçam a independência condicional;
  - definir a tabela de probabilidade condicional para a variável.

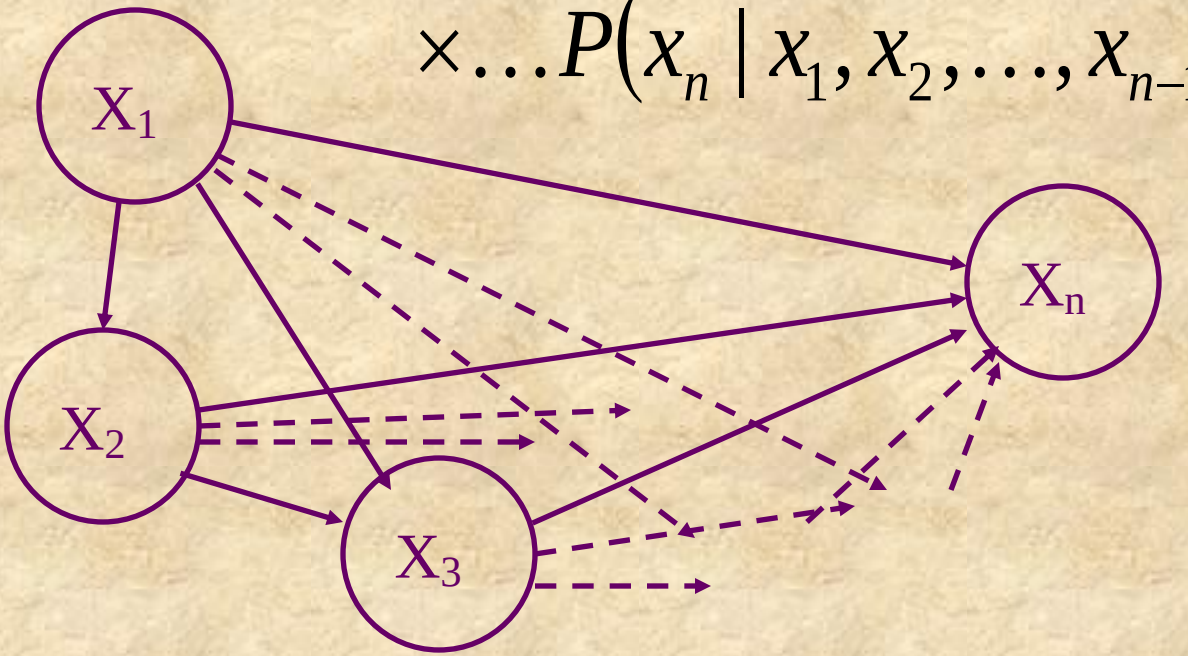
Exemplo simples de rede bayesiana (cont.)



Decomposição da Probabilidade Conjunta

$$P(x_1, x_2, \ldots, x_n) = P(x_1) \times P(x_2 \mid x_1) \times P(x_3 \mid x_1, x_2)$$

$$\times \ldots P(x_n \mid x_1, x_2, \ldots, x_{n-1})$$



- Essa decomposição deixa clara a necessidade de a rede bayesiana ser um grafo acíclico

Tipos de conhecimento

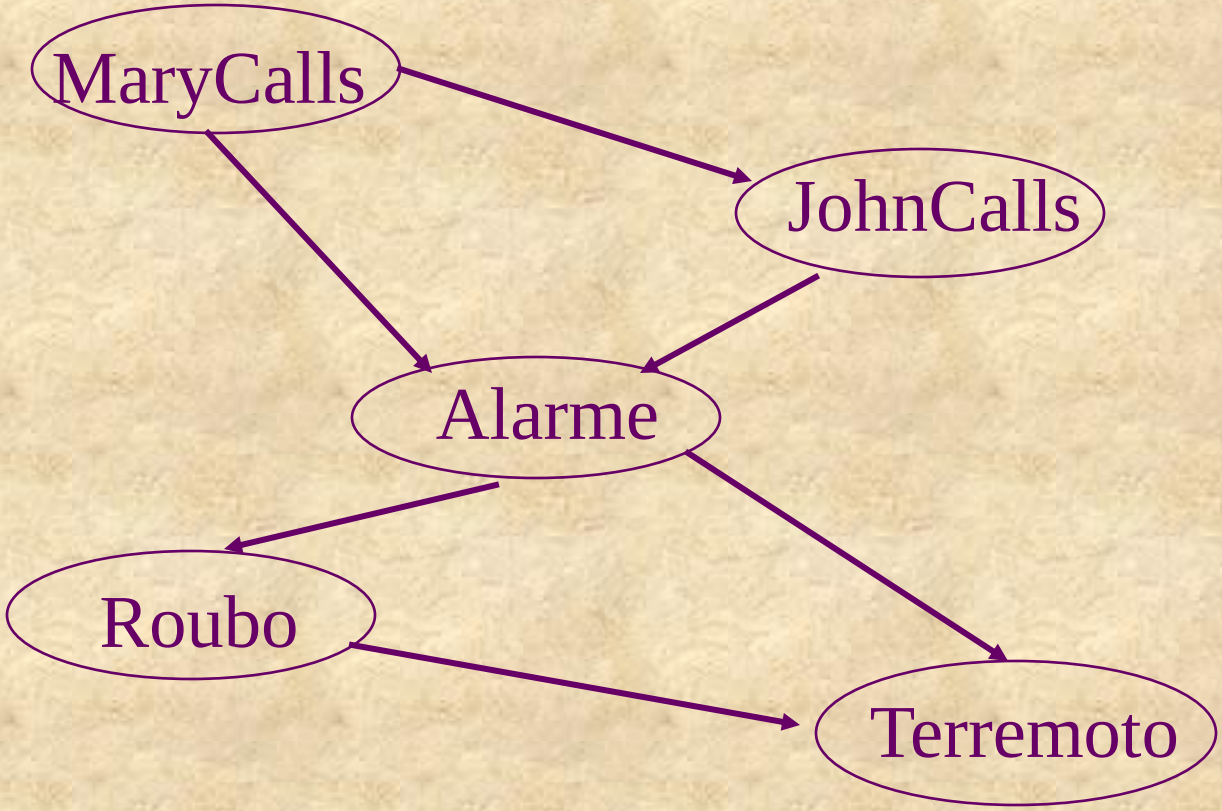
- Causal
  - Refletem a direção conhecida de causalidade no mundo: para algumas propriedades do mundo percepções são geradas.
  - ex, P(DorDeDente|Cárie), P(MaryCalls|Alarme)
- Diagnóstico
  - Infere a presença de propriedades escondidas diretamente da percepção.
  - Produzem conclusões fracas.
  - ex, P(Ca'rie|DorDeDente), P(Alarme|MaryCalls)

Ordenar nós de uma rede bayesiana

- Algoritmo de construção apresentado especifica a ordem
- Raízes sempre causais, folhas sem influência causal sobre nenhuma outra variável
- Se não perde:
  - concisão da rede
  - eficiência computacional (pior caso volta a distribuição de probabilidade conjunta)

Exemplo de rede bayesiana não puramente causal

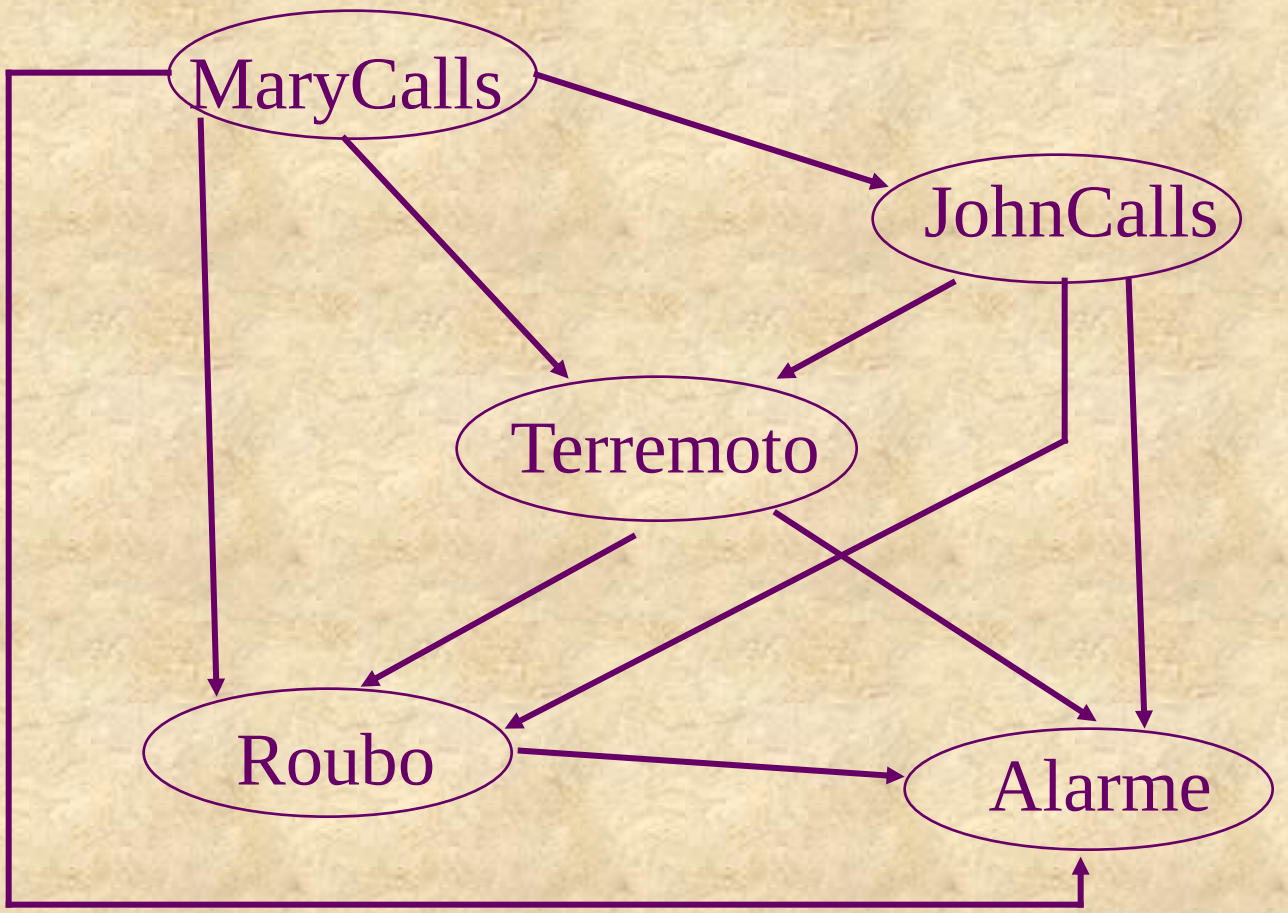
- Vamos usar o exemplo do alarme com a seguinte ordem de inserção dos nós:
  - MaryCalls, JohnCalls, Alarme, Roubo e Terremoto.



Exemplo de rede bayesiana não puramente causal (cont.)

- Problemas:
  - A figura possui duas conexões a mais;
  - julgamento não natural e difícil das probabilidades;
- Tendo uma rede puramente causal, teríamos um número menor de conexões
- Podemos piorar ainda mais a nossa configuração da rede, seguindo a seguinte ordem de criação:
  - MaryCalls, JohnCalls, Terremoto, Roubo e Alarme.

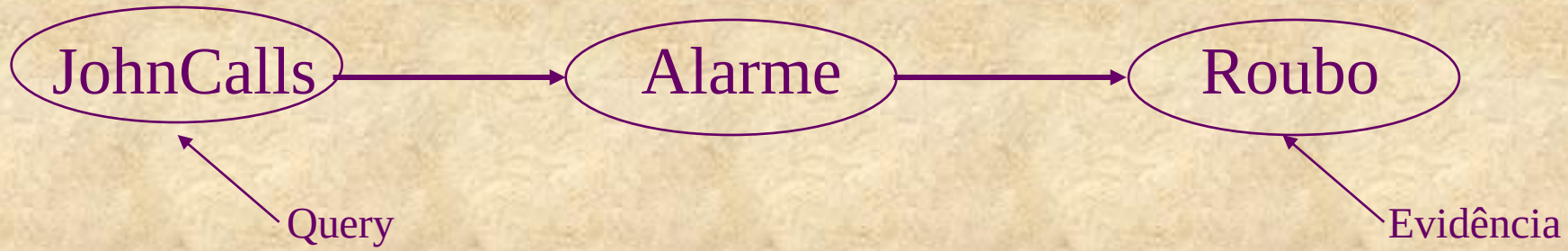
Exemplo de rede bayesiana não puramente causal (cont.)



Versatilidade das redes bayesianas

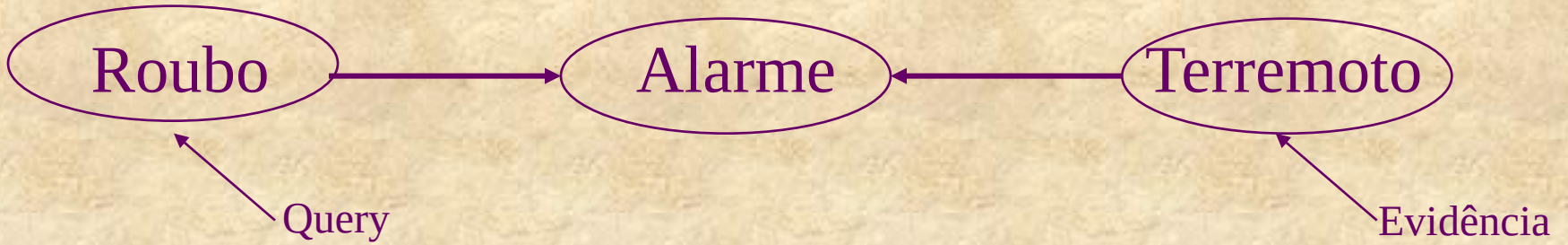
- Redes Bayesianas oferecem 4 tipos de inferência:
  - **Causal** (da causa para o efeito)
    - P(JohnCalls/Roubo) = 0,86

  - **Diagnóstico** (do efeito para a causa)
    - P(Roubo/JohnCalls) = 0,016

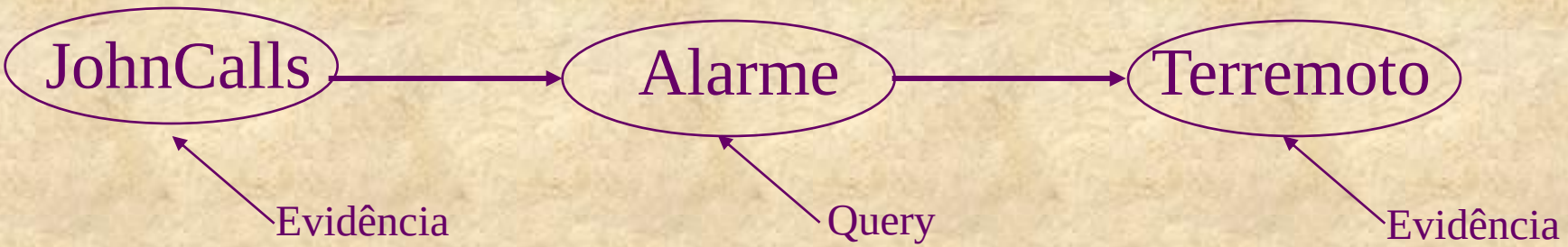


Versatilidade das redes bayesianas

- **Intercausal** (entre causas com um efeito comum)
  - P(Roubo/Alarme) = 0,376
  - P(Roubo/Alarme $\wedge$Terremoto) = 0,373



- **Mista** (combinando duas ou mais das de cima)
  - P(Alarme/JohnCalls $\wedge\neg$Terremoto) = 0,03
  - Este é um uso simultâneo de inferência causal e diagnóstico.



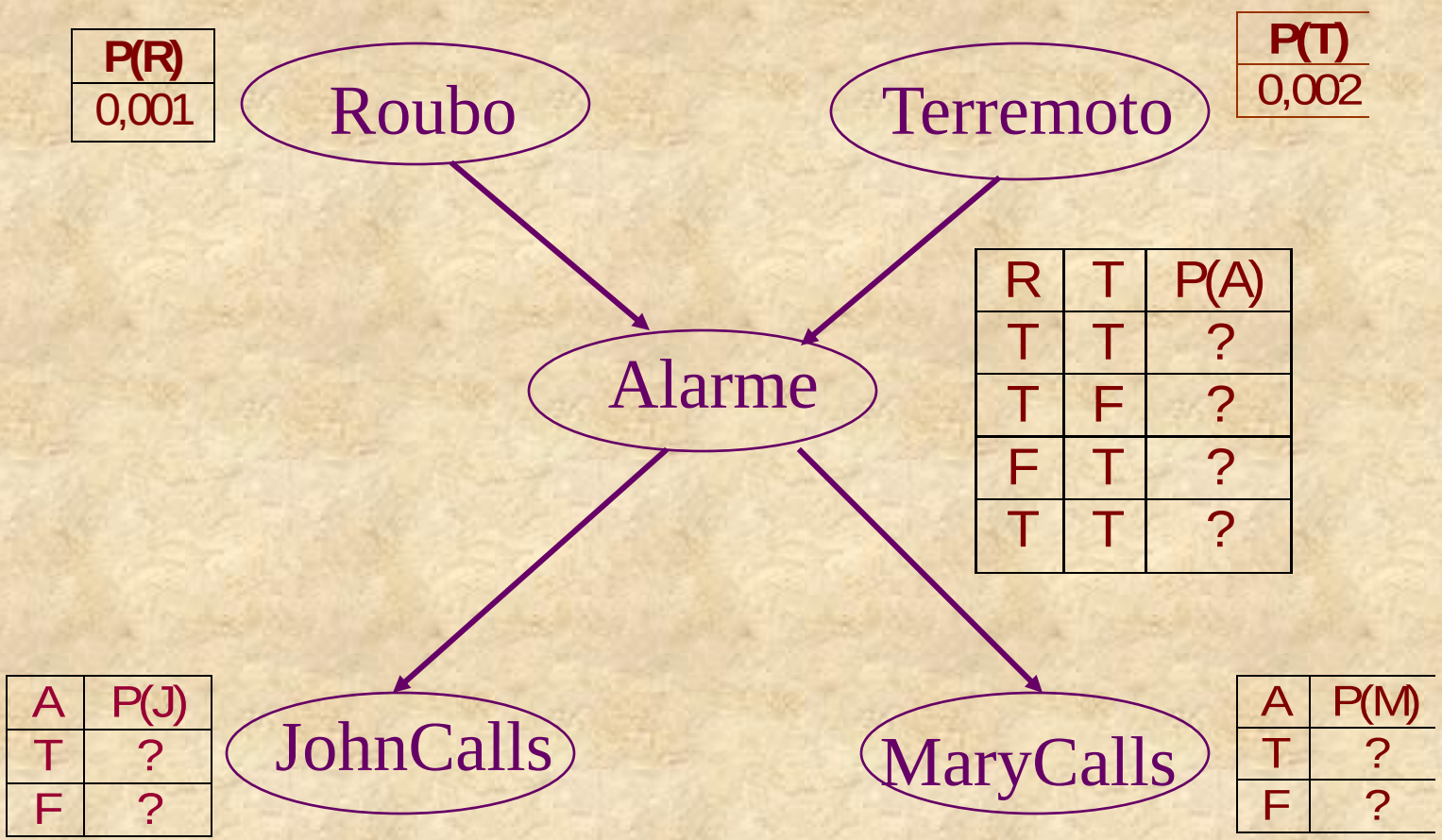
Inferência em Redes Bayesianas

- Como inferir as probabilidades dos valores de uma ou mais variáveis na rede, à partir das probabilidades dos valores das outras variáveis
  - A rede Bayesiana contém toda a informação necessária para essa inferência
  - Quando se trata de apenas uma variável, a inferência é trivial
  - No caso geral, o problema é NP hard
- Na prática, pode-se alcançar-la de várias formas
  - Métodos exatos de inferência funcionam bem para algumas estruturas de rede
  - Métodos de tipo Monte Carlo "simulam" a rede aleatoriamente para obter soluções aproximadas

Aprendizado em redes bayesianas

- 4 problemas de aprendizagem:
  - Estrutura conhecida, completamente observável
    - as tabelas de probabilidade condicionada podem ser estimadas usando o conjunto de exemplos com classificador *ingênuo?* de Bayes
  - Estrutura desconhecida, completamente observável
    - o problema é construir a topologia da rede. Busca no espaço de estruturas.
  - Estrutura conhecida, variáveis escondidas
    - caso parecido com aprendizado em redes neurais
  - Estrutura desconhecida, variáveis escondidas
    - não se conhece algoritmos para este tipo de problema

Exemplo da tarefa de aprendizagem



Aprender probabilidades com estrutura fixa

- Humanos acham fácil dizer o que causa o que, mas acham difícil colocar números nos links.
- Tarefa de aprendizagem
  - Dados:
    - relações de independência entre variáveis aleatórias (estrutura)
    - probabilidades *a priori* das variáveis "de entrada"
    - probabilidades *a posteriori* de variáveis "de saída"
  - Calcular:
    - probabilidades condicionais das variáveis dependentes
- 2 algoritmos principais:
  - ***gradiente ascendente*** de P(D|Hi) - muito parecido com aprendizagem de pesos em redes neurais
  - ***algoritmo EM*** (Estimação Média)
  - ambos iterativos e sujeito a encontrar mínimo local

Aprendizagem de Redes Bayesianas

- Variantes da tarefa de aprendizagem
  - A estrutura da rede pode ser conhecida ou desconhecida
  - O conjunto de treinamento pode fornecer valores para todas as variáveis da rede ou para somente algumas
- Se a estrutura é conhecida e todas as variáveis observadas
  - Então é tão fácil como treinar um classificador Bayesiano ingênuo

Aprendizagem de Redes Bayesianas

- Suponha a estrutura conhecida e variáveis parcialmente observáveis
  - Exemplo, observa-se *fogo na Floresta, Tempestade, Ônibus de turismo*, mas não *Raio, Fogo no Acampamento*
  - Problema similar o treinamento de uma rede neural com variáveis ocultas
  - Aprende-se a tabela de probabilidades condicionais de cada nó usando o algoritmo do gradiente ascendente
  - O sistema converge para a rede h que maximiza localmente P(D/h)

Gradiente Ascendente p/ Redes Bayesianas

- Seja w_{ijk} uma entrada na tabela de probabilidade condicional para a variável Y_i na rede
- $W_{ijk} = P(Y_i = y_{ij}/\text{Predecessores}(Y_i) = \text{lista } u_{ik} \text{ de valores})$
- Exemplo, se Y_i = Fogo no Acampamento, então u_{ik} pode ser {*Tempestade = T, Ônibus de Turismo = F*}
- Aplicar o gradiente ascendente repetidamente
  1. Atualizar todos os w_{ijk} usando os dados de treinamento D

$$w_{ijk} \leftarrow w_{ijk} + \eta \sum_{d \in D} \frac{P_h(y_{ij}, u_{ik} / d)}{w_{ijk}}$$

  2. Normalizar os w_{ijk} para assegurar

$$\sum_j w_{ijk} = 1 \qquad e \qquad 0 \le w_{ijk} \le 1$$

Aprendizagem em Redes Bayesianas

- O algoritmo EM também pode ser usado. Repetir:
  1. Calcule as probabilidades das variáveis não observadas, supondo h verdadeira
  2. Calcule novo w_{ijk} que maximize E[ln P(D/h)], onde D agora inclui tanto as variáveis observadas como as probabilidades calculadas das não observadas
- Quando a estrutura é desconhecida
  - Tópico de pesquisa ativo ...

Sumário de Redes Bayesianas

- Combina conhecimento a priori com dados observados
- O impacto do conhecimento a priori (quando correto) é a redução da amostra de dados necessários
- Área de pesquisa ativa
  - Passar de variáveis Booleanas para variáveis numéricas
  - Distribuições em vez de tabelas
  - Lógica de primeira ordem no lugar de proposicional
  - Métodos de inferência mais efetivos

Expectation Maximization (EM)

- Quando usar:
  - Os dados são observáveis apenas parcialmente
  - Aprendizagem não supervisionada (Clustering, os grupos são desconhecidos)
  - Aprendizagem Supervisionada (alguns valores de algumas variáveis não são observados
- Alguns usos
  - Treinamento das redes Bayesianas
  - Clustering
  - etc

Geração de Dados a partir de k Gaussianas

- Cada instancia x é gerada
  1. Escolhendo uma das k Gaussianas com probabilidade uniforme
  - Gerando uma instancia aleatoriamente de acordo com essa Gaussiana

EM para a estimação de k médias

- Dado
  - Instancias de x geradas pela mistura de k distribuições Gaussianas
  - Médias desconhecidas $\langle \mu_1, \ldots, \mu_k \rangle$ das k Gaussianas
  - Não se sabe que instancia x_i foi gerada por qual Gaussiana
- Determine
  - Estimativas de Máxima Verossimilhança de $\langle \mu_1, \ldots, \mu_k \rangle$
- Considere a descrição completa de cada instancia como $y_i = \langle x_i, z_{i1}, z_{i2} \rangle$, onde
  - z_{ij} é 1 se x_i for gerado pela j-ésima Gaussiana
  - x_i observável
  - z_{ij} não observável

EM para a estimação de k médias

- EM algoritmo: Selecione aleatoriamente uma hipótese inicial h = < μ_1, μ_2>

- Passo E: Calcule o valor esperado E[z_{ij}] de cada variável oculta z_{ij}, supondo válida a hipótese atual h = < μ_1, μ_2>

$$E[z_{ij}] = \frac{p(x = x_i / \mu = \mu_j)}{\sum_{n=1}^{2} p(x = x_i / \mu = \mu_n)} = \frac{\exp\left[-\frac{1}{2\sigma^2}(x_i - \mu_j)^2\right]}{\sum_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu_n)^2\right]}$$

EM para a estimação de k médias

- Passo M: Calcule a nova hipótese de Máxima Verossimilhança $h' = \langle \mu'_1, \mu'_2 \rangle$, supondo que o valor assumido por cada variável oculta z_{ij} é o valor esperado $E[z_{ij}]$ já calculado. Troque $h = \langle \mu_1, \mu_2 \rangle$ por $h' = \langle \mu'_1, \mu'_2 \rangle$.

$$\mu_j \leftarrow \frac{\sum_{i=1}^{m} E[z_{ij}] x_i}{\sum_{i=1}^{m} E[z_{ij}]}$$

- Converge para um máximo de verossimilhança h local e fornece estimativas para as variáveis ocultas z_{ij}

Exemplo de aplicação bem sucedida de redes bayesianas: PathFinder

- Sistema especialista p/ diagnóstico de doenças nós linfáticos.
  - PathFinder I - sistema baseado em regras sem incerteza.
  - PathFinder II
    - comparando vários métodos de representação da incerteza

    (teoria de crença de Dempster-Shafer, coeficientes de incerteza,etc)
    - modelo bayesiano simplificado (todas as doenças assumidas independentes) melhor de todos (10% de erro)
  - PathFinder III - modelo bayesiano melhorado com aquisição mais cuidadosa das probabilidades a priori com especialistas
  - PathFinder IV - Redes Bayesianas
    - melhor do que os especialistas cujo conhecimento foi codificado

Bibliografia

- Russel, S, & Norvig, P. (1995). Artificial Intelligence: a Modern Approach (AIMA) Prentice-Hall. Pages 436-458, 588-593
- An Introduction to Baysean Networks
- Mitchell, T. & (1997): Machine Learning, McGraw-Hill. Cap.6
- Fayyad et al. (1996): Advances in knowledge discovery and data mining, AAAI Press/MIT Press. Cap.11
- Pearl, J. (1988) Probabilistic Reasoning in Inteligent Systems