Inteligência Artificial

Raciocínio Probabilístico – Introdução a Redes Bayesianas

Uncertainty (Partially observed or stochastic) environments?

1. **Ignorance.** The limits of our knowledge lead us to be uncertain about many things. Does our poker opponent have a flush or is she bluffing?
2. **Physical randomness or indeterminism.** Even if we know everything that we might care to investigate about a coin and how we impart spin to it when we toss it, there will remain an inescapable degree of uncertainty about whether it will land heads or tails when we toss it. A die-hard determinist might claim otherwise, that some unimagined amount of detailed investigation might someday reveal which way the coin will fall; but such a view is for the foreseeable future a mere act of scientistic faith. We are all practical indeterminists.
3. **Vagueness.** Many of the predicates we employ appear to be vague. It is often unclear whether to classify a dog as a spaniel or not, a human as brave or not, a thought as knowledge or opinion.

Referências Adicionais

- Russel e Norvig cap. 14 e 15

- Pearl, Judea. **Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference.** Morgan Kaufmann, San Mateo, California. 1988.

- Witten, I.H., Frank, E. Data Mining Practical Machine Learning Tools and Techniques. 2a. Ed. Elsevier. 2005

Revisão de Conceitos Básicos de Probabilidade

P(A | K) – probabilidade condicional ou posterior.
Crença em *A*, dado o corpo de informação *K.*

P(A) – probabilidade *a priori*: Crença em A, na falta de informação adicional proveniente de K.

Uma Variável aleatória tem um domínio (conjunto de valores)e associada a cada um a probabilidade de ocorrência daquele valor. Essa função é chamada de distribuição de Probabilidade.
Exemplo:
Variável Tempo = {Sol, Chuva, Nublado}
P*(Tempo) – é uma distribuição de probabilidade*
P*(Tempo) = <0,7;0,2;0,1>*

P*(Tempo=sol) = 0.7*
P*(Tempo=chuva) = 0.2*
P*(Tempo=nublado) = 0.1*

No caso contínuo, usa-se o termo função de densidade de probabilidade. Vamos focar no caso discreto.

Probabilidade condicional

Probabilidade condicional ou posterior, e.g., **P**(cárie|dordedente) = 0.8

i.e., dado que dordedente é tudo que conheço, a chance de cárie (vista por mim) é de 80%.

P(Cárie | Dordedente) = Vetor de 2 elementos cada um com dois elementos. Por Exemplo: P(Cárie | Dordedente) = $<<0,8;0,2>;<0,01;0,99>>$

Se sabemos mais, e.g., *cárie é também observada*, então

P(cárie|dordedente, *cárie*) = 1

OBS:

1) A crença menos específica permanece válida, mas pode ficar inútil.

2) A nova evidência pode ser inútil:

P(cárie|dordedente, Corinthians derrotado) = P(cárie|dordedente) = 0.8

NOTE A IMPORTÂNCIA DO CONHECIMENTO DO DOMÍNIO PARA QUALQUER PROCESSO DE INFERÊNCIA.

O Axioma Básico

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

- Ou:

$$P(A,B) = P(A \mid B)\,P(B)$$

Corolário:

$$P(A) = \sum_i P(A, B_i)$$

$$P(A) = \sum_i P(A \mid B_i)\,P(B_i)$$

Regra da Cadeia

Regra da Cadeia:

$$P(E_1, E_2, \ldots E_n) = P(E_n \mid E_{n-1}, \ldots E_2, E_1) \ldots P(E_2 \mid E_1) P(E_1)$$

Prova:

$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1, \ldots, X_{n-1})\, \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})$

$= \mathbf{P}(X_1, \ldots, X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-2})\, \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})$

$= \ldots$

$= \Pi\, \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$

Inversão Bayesiana (Regra de Bayes)

- P(H|e): Probabilidade posterior

$$P(H|e) = \frac{P(e\,|\,H)P(H)}{P(e)}$$

P(H): Probabilidade a priori

Por quê a fórmula é importante?

Muitas vezes P(e|H) é fácil de calcular, ao contrário de P(H|e)?

Exemplo.No cassino, um croupier fala 12! Ele jogou os dados ou estava comandando um jogo de roleta?

P(*12|dados*), **P**(*12|roleta*): fácil de modelar. **P**(*dados*), **P**(*roleta*): fácil, basta ver número de mesas de dado ou roleta no cassino. **P**(*dados|12*), **P**(*roleta|12*): não é tão fácil estimar . . .

Outro Exemplo?: Pense sobre as probabilidades a priori e posterior no caso Meningite x DordePescoço.

Rede Bayesiana ou Rede de Crença (Belief Network)

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

Syntax:

a set of nodes, one per variable

a directed, acyclic graph (link $\approx$ "directly influences")

a conditional distribution for each node given its parents:

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

In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

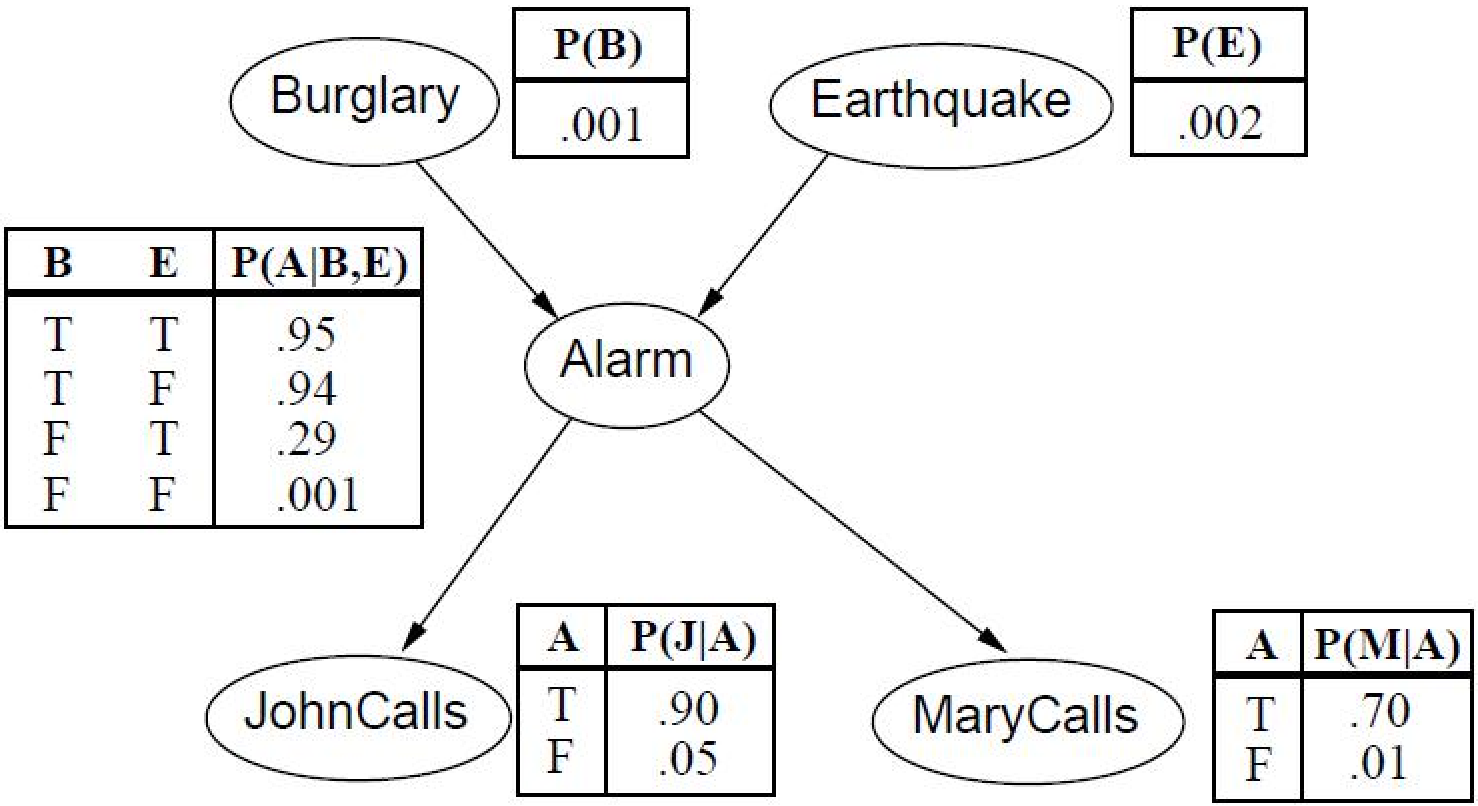
Exemplo: Terremoto ou ladrão?

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

Variables: $Burglar$, $Earthquake$, $Alarm$, $JohnCalls$, $MaryCalls$
Network topology reflects "causal" knowledge:

- A burglar can set the alarm off
- An earthquake can set the alarm off
- The alarm can cause Mary to call
- The alarm can cause John to call

Exemplo – 2



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Método para construção de uma rede

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

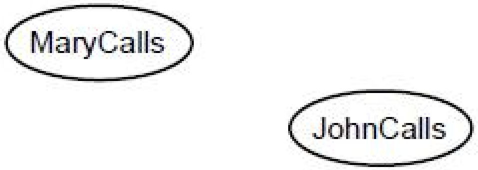
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \prod_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1}) \quad \text{(chain rule)} \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i)) \quad \text{(by construction)}\end{aligned}$$

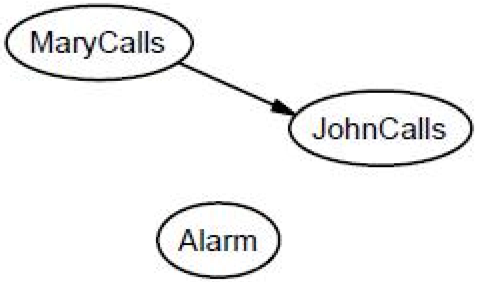
Exemplo

Suppose we choose the ordering M, J, A, B, E



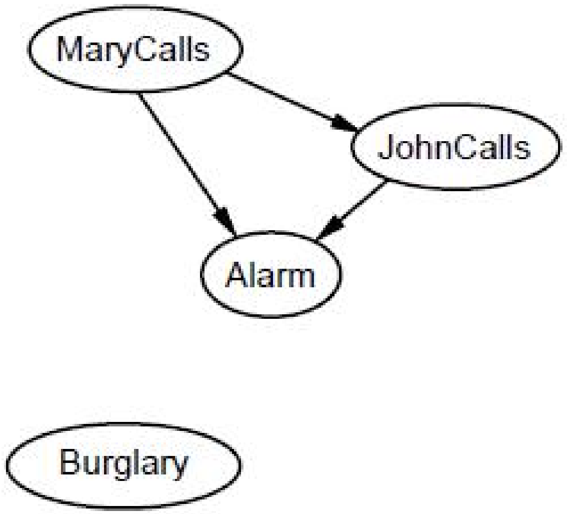
$P(J|M) = P(J)$?

Suppose we choose the ordering M, J, A, B, E



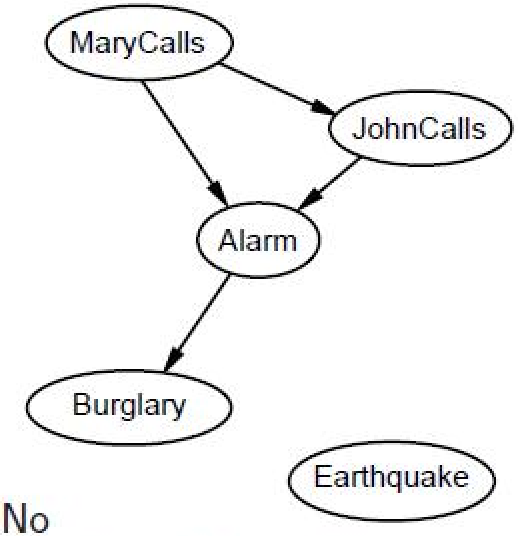
$P(J|M) = P(J)$? No
$P(A|J, M) = P(A|J)$? $P(A|J, M) = P(A)$?

Suppose we choose the ordering M, J, A, B, E



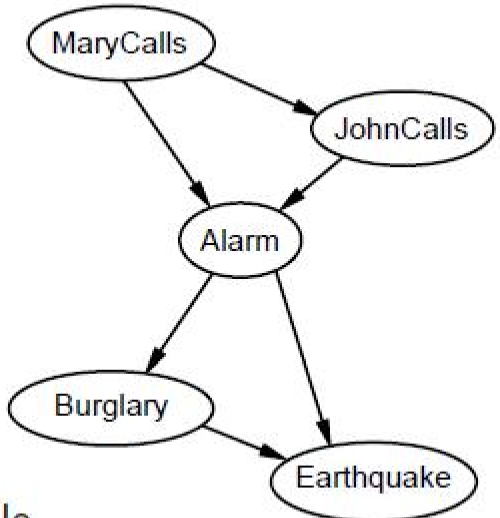
$P(J|M) = P(J)$? No
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No
$P(B|A,J,M) = P(B|A)$?
$P(B|A,J,M) = P(B)$?

Suppose we choose the ordering M, J, A, B, E

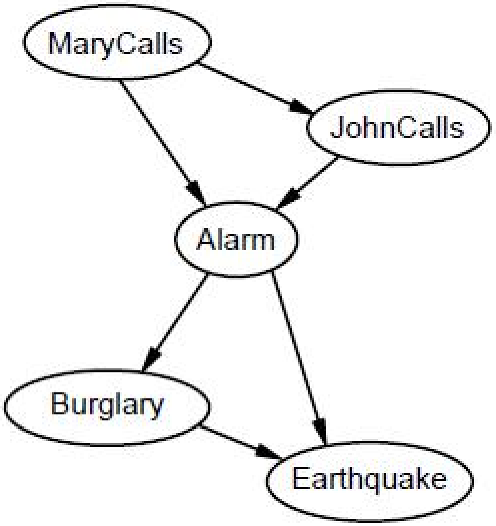


$P(J|M) = P(J)$? No
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No
$P(B|A,J,M) = P(B|A)$? Yes
$P(B|A,J,M) = P(B)$? No
$P(E|B,A,J,M) = P(E|A)$?
$P(E|B,A,J,M) = P(E|A,B)$?

Suppose we choose the ordering M, J, A, B, E



$P(J|M) = P(J)$? No
$P(A|J,M) = P(A|J)$? $P(A|J,M) = P(A)$? No
$P(B|A,J,M) = P(B|A)$? Yes
$P(B|A,J,M) = P(B)$? No
$P(E|B,A,J,M) = P(E|A)$? No
$P(E|B,A,J,M) = P(E|A,B)$? Yes



Deciding conditional independence is hard in noncausal directions

(Causal models and conditional independence seem hardwired for humans!)

Assessing conditional probabilities is hard in noncausal directions

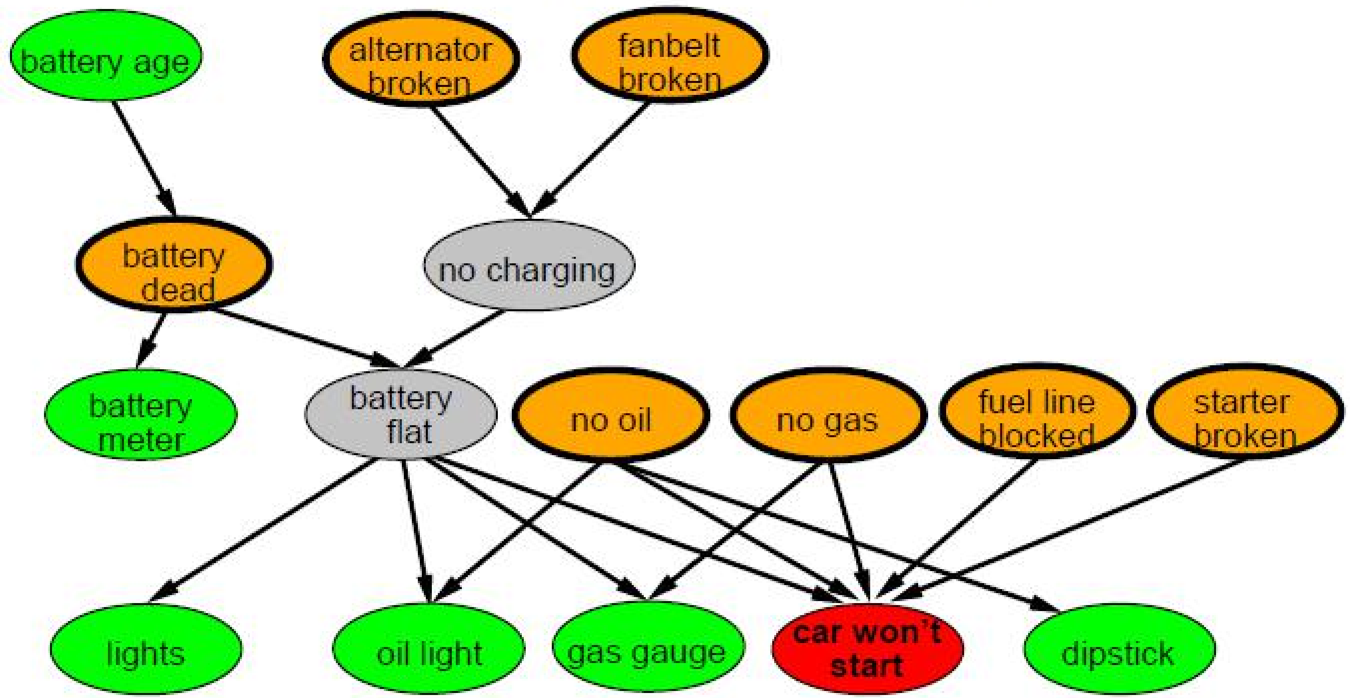
Network is less compact: $1+2+4+2+4=13$ numbers needed

Outro Exemplo: Conserto de Carro

Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



Inferência em Redes Bayesianas

- Dada uma rede, devemos ser capaz de inferir a partir dela isto é :

- Busca responder questões simples, $P(X| E=e)$
  - Ex. : $P(NoGas|Gauge = empty, Lights = on, Starts = false)$
- Ou questões conjuntivas: $P(X_i, X_j | E=e)$
  - Usando o fato:

$$\mathbf{P}(X_i, X_j|\mathbf{E}=\mathbf{e}) = \mathbf{P}(X_i|\mathbf{E}=\mathbf{e})\mathbf{P}(X_j|X_i, \mathbf{E}=\mathbf{e})$$

- A inferência pode ser feita a partir da distribuição conjunta total ou por **enumeração**

Inferência com Distribuição Conjunta Total: Exemplo

A	B	C	P(A,B,C)
F	F	F	P(A=F,B=F,C=F)
F	F	T	P(A=F,B=F,C=T)
F	T	F	..
F	T	T	..
T	F	F	...
T	F	T	....
T	T	F	.....
T	T	T	P(A=T,B=T,C=T)

- Por exemplo para saber
- P(A|b) temos
- P(A|b)= P(A, b)/P(b)=
- <P(a, b)/P(b);P(⌐a , b)/**P(b)** > =
- = α < P(a, b);P(⌐a , b)>
- = α [<P(a,b,c)+P(a,b,⌐c); P(⌐a,b,c)+P(⌐a,b, ⌐c)>]

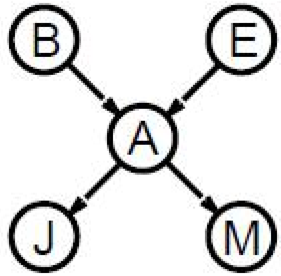
Observe que α pode ser visto como um fator de normalização para o vetor resultante da distribuição de probabilidade, pedida P(A|b). Assim pode-se evitar seu cálculo, Simplesmente normalizando <P(a,b); P(⌐a , b) >

Inferência por Enumeração

Slightly intelligent way to sum out variables from the joint without actually constructing its explicit representation

Simple query on the burglary network:

$\mathbf{P}(B|j,m)$
$= \mathbf{P}(B,j,m)/P(j,m)$
$= \alpha\mathbf{P}(B,j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B,e,a,j,m)$



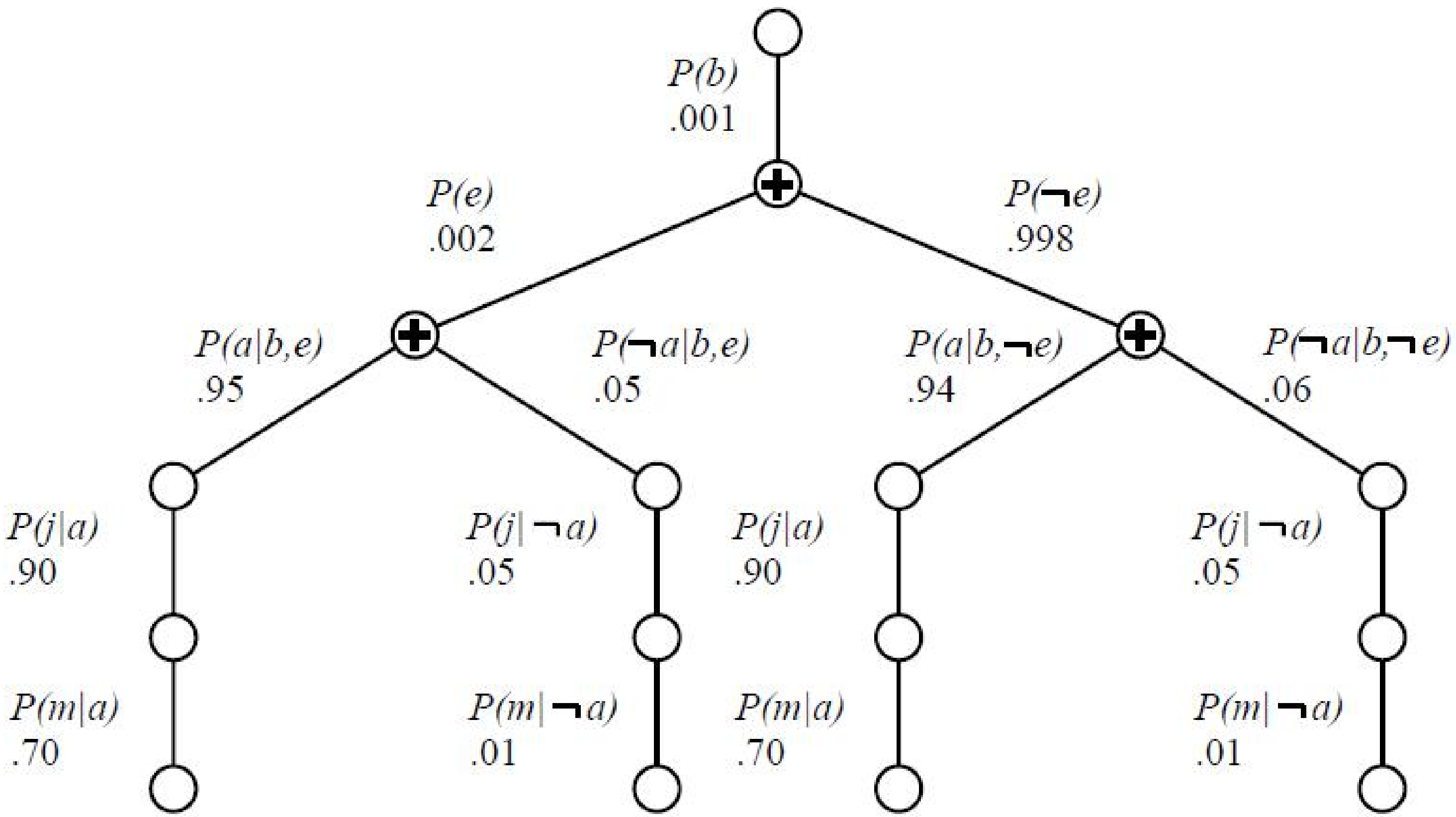
Rewrite full joint entries using product of CPT entries:

$\mathbf{P}(B|j,m)$
$= \alpha\Sigma_e\Sigma_a\mathbf{P}(B)P(e)\mathbf{P}(a|B,e)P(j|a)P(m|a)$
$= \alpha\mathbf{P}(B)\Sigma_e P(e)\Sigma_a\mathbf{P}(a|B,e)P(j|a)P(m|a)$

Recursive depth-first enumeration: $O(n)$ space, $O(d^n)$ time

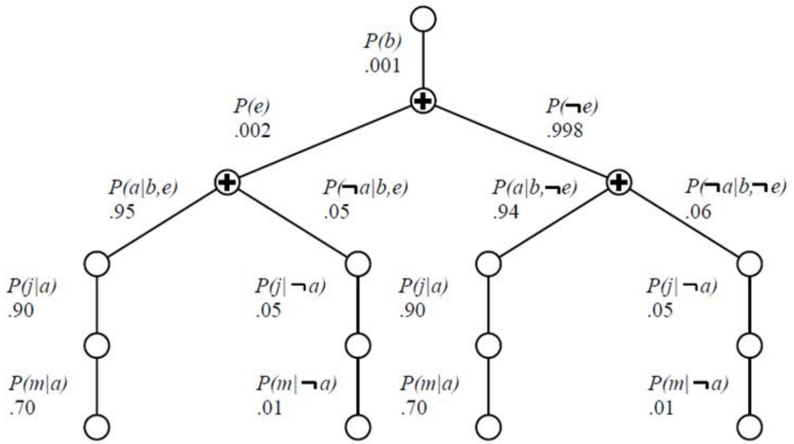
Inferência por Enumeração - 2

Enumeration is inefficient: repeated computation
e.g., computes $P(j|a)P(m|a)$ for each value of e



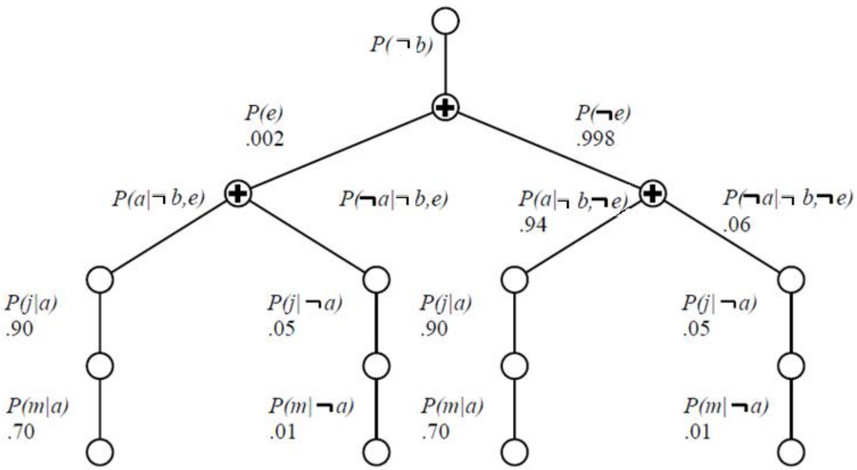
- Pode ser melhorada através do armazenamento dos valores já calculados (Programação Dinâmica)

Calculando P(b) não normalizado



				"P(b) nao normalizado"		
				0,0005922		
				0,001		
	+			0,5922426		
		0,001197			0,591046	
*		0,002	*		0,998	
+		0,598525	+		0,59223	
	0,5985	0,000025		0,5922	0,00003	Produtorio
	0,95	0,05		0,94	0,06	
	0,9	0,01		0,9	0,01	
	0,7	0,05		0,7	0,05	

Calculando P(não b) não normalizado



		"P(nao b) nao normalizado"	
		0,001492	
		0,999	
+		0,001493	
	0,000366		0,001127
	0,002 *		0,998
	0,183055 +		0,00113
0,1827	0,000355	0,00063	0,0005
0,29	0,71	0,001	0,999
0,9	0,01	0,9	0,01
0,7	0,05	0,7	0,05

Algoritmo de Enumeração

function ENUMERATION-ASK(X, $\mathbf{e}$, bn) **returns** a distribution over X
  inputs: X, the query variable
    $\mathbf{e}$, observed values for variables $\mathbf{E}$
    bn, a Bayes net with variables $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ /* $\mathbf{Y}$ = *hidden variables* */

  $\mathbf{Q}(X) \leftarrow$ a distribution over X, initially empty
  for each value x_i of X **do**
    $\mathbf{Q}(x_i) \leftarrow$ ENUMERATE-ALL(bn.VARS, $\mathbf{e}_{x_i}$)
      where $\mathbf{e}_{x_i}$ is $\mathbf{e}$ extended with $X = x_i$
  return NORMALIZE($\mathbf{Q}(X)$)

---

function ENUMERATE-ALL($vars$, $\mathbf{e}$) **returns** a real number
  if EMPTY?($vars$) **then return** 1.0
  $Y \leftarrow$ FIRST($vars$)
  if Y has value y in $\mathbf{e}$
    then return $P(y \mid parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}$)
    else return $\sum_y P(y \mid parents(Y)) \times$ ENUMERATE-ALL(REST($vars$), $\mathbf{e}_y$)
      where $\mathbf{e}_y$ is $\mathbf{e}$ extended with $Y = y$

Sumário

- Revisão de probabilidade
- Redes Bayesianas ou Redes de crença
- Inferência probabilística
- Aprendizado em método probabilísticos
- Métodos simplificados: Bayes ingênuo
- Decisão baseado em probabilidades: De probabilidades para ação

Aprendizado em modelos probabilísticos

- Aprender em redes bayesianas é o processo de determinar a topologia da rede (isto é, seu grafo direcionado) e as tabelas de probabilidade condicional

- Problemas?
  - Como determinar a topologia?
  - Como estimar as probabilidades ?
  - Quão complexas são essas tarefas?
    - Isto é quantas topologias e quantas probabilidades precisariam ser determinadas….

Tamanho das Tabelas de Probabilidade Condicional e Distribuição Conjunta Total

A	B	C	P(A,B,C)
F	F	F	P(A=F,B=F,C=F)
F	F	T	P(A=F,B=F,C=T)
F	T	F	..
F	T	T	..
T	F	F	...
T	F	T	....
T	T	F	.....
T	T	T	P(A=T,B=T,C=T)

- Vamos supor que cada variável é influenciada por no máximo k outras variáveis (Naturalmente, k<n=total de variáveis).
- Supondo variáveis booleanas, cada tabela de probabilidade condicional (CPT) terá no máximo 2^k entradas (ou probabilidades). Logo ao total haverá no máximo $n * 2^k$ entradas
- Enquanto, na distribuição conjunta Total haverá 2^n entradas. Por exemplo, para n=30 com no máximo cinco pais (k=5) isto significa 960 ao invés de mais um bilhão (2^{30})

Número de "entradas" da Distribuição Conjunta e na Rede Bayesiana – 2

- Em domínios onde **cada variável** pode ser diretemante **influenciada** por **todas** as outras, tem-se a rede totalmente conectada e assim exige-se a **quantidade de entradas da mesma ordem** da distribuição conjunta total

- Porém se essa dependência for tênue, **pode não valer** a pena a complexidade adicional na rede em relação ao pequeno ganho em exatidão

- Via de regra, se nos fixarmos em um **modelo causal** acabaremos tendo de especificar uma quantidade menor de números, e os números frequentemente serão mais fáceis de calcular. (Russel,Norvig, 2013, pg. 453)

- **Modelos causais** são aqueles onde se especifica no sentido causa efeito, isto é P(efeito|causa) ao invés de P(causa|efeito), oque geralmente é necessário para diagnóstico

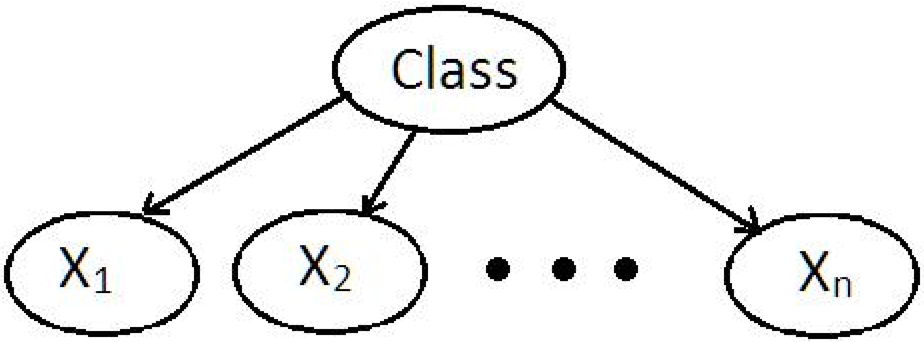
Simplificando a representação tabelas de probabilidade condicional (CPT)

- Vimos que que o número de entradas de uma CPT cresce exponencialmente
  - Para o caso binário e K pais, a CPT de um nó terá 2^k probabilidades a serem calculadas

- Vejamos uma abordagens para simplificar a rede através da adoção de hipóteses simplificadoras
  - Bayes Ingênuo

Naïve Bayes (Bayes Ingênuo)

- Uma classe particular e simples de redes bayesianas é chamada de Bayes Ingênuo (Naïve Bayes)
- Ela é simples por supor independência condicional entre todas as variáveis X dada a variável Class
- As vezes, chamado também de classificador Bayes, por ser frequen
abordagem inicial para



Naïve Bayes (Bayes Ingênuo) - 2

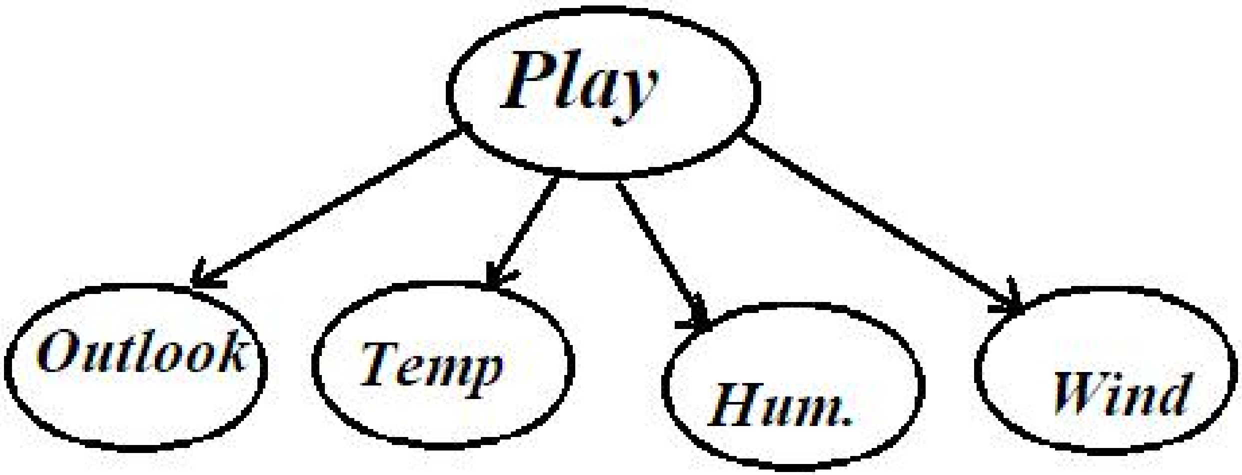
- A topologia simples traz a vantagem da representação concisa da Distribuição Conjunta Total.
- Como todo os nós tem no máximo um pai, cada CPT de no X tem apenas duas entradas e uma entrada no nó classe. Logo, (2n-1) entradas para toda a rede. Naïve Bayes é **linear** em relação ao número de nós (n) !!!!
- "Na prática, sistemas de Bayes ingênuos podem funcionar surpreendentemente bem….". pg. 438

Exemplo de Naïve Bayes

- Vamos retomar o exemplo do jogo de tênis

Ex	Céu	Temperatura	Umidade	Vento	JogarTênis
X1	Ensolarado	Quente	Alta	Fraco	NÃO
X2	Ensolarado	Quente	Alta	Forte	NÃO
X3	Nublado	Quente	Alta	Fraco	SIM
X4	Chuvoso	Boa	Alta	Fraco	SIM
X5	Chuvoso	Fria	Normal	Fraco	SIM
X6	Chuvoso	Fria	Normal	Forte	NÃO
X7	Nublado	Fria	Normal	Forte	SIM
X8	Ensolarado	Boa	Alta	Fraco	NÃO
X9	Ensolarado	Fria	Normal	Fraco	SIM
X10	Chuvoso	Boa	Normal	Fraco	SIM
X11	Ensolarado	Boa	Normal	Forte	SIM
X12	Nublado	Boa	Alta	Forte	SIM
X13	Nublado	Quente	Normal	Fraco	SIM
X14	Chuvoso	Boa	Alta	Forte	NÃO

Usando a abordagem Bayes ingênuo



- Problema a resolver:

Outlook	Temperature	Humidity	Windy	Play
sunny	cool	high	true	?

Solução:

- P(Play|Outlook, Temp, Hum, Wind)=
- P(Outlook, Temp, Hum, Wind|Play)P(Play)/P(Outlook, Temp, Hum, Wind)=
- Regra da cadeia e indepêndencia:
- P(Outlook|Play)P(Temp|Play)P(Hum|Play)P(Wind|Play)P(Play)/ P(Outlook, Temp, Hum, Wind)
- O método de inferência por enumeração já visto é aplicável!!!
- Estima-se as probabilidades pelo conjunto de treinamento

Contagens e probabilides estimadas pelo conjunto de treinamento

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

- P(Play=s|Outlook=sunny, Temp=cool, Hum=high, Wind= true)=
- P(sunny|play)P(cool|play)P(high|play)P(true|play)P(Play)/P(e) = 2/9*3/9*3/9*3/9*9/14/P(e) =0.0053/P(e)

Solução 3 - continuação

- Da mesma forma,
- P(sunny|play)P(cool|play)P(high|play)P(true|play)P(Play)/P(e) = 3/5*1/5*4/5*3/5*5/14/P(e) =0.0206/P(e)
- Mas P(H, e) e P(not H, e) tem que somar 1, assim:

$$\text{Probability of } yes = \frac{0.0053}{0.0053 + 0.0206} = 20.5\%,$$

$$\text{Probability of } no = \frac{0.0206}{0.0053 + 0.0206} = 79.5\%.$$

Estimativas de Probabilides

- Qual a estimativa da probabilidade P(Outlook=overcast|Play=no)?

Outlook			Temperature			Humidity			Windy			Play	
	yes	*no*		*yes*	*no*		*yes*	*no*		*yes*	*no*	*yes*	*no*
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

- Zero! Isto é razoável? Como resolver?
- Uma Solução: estimador de Laplace (Laplace smoothing). Seja V o número de valores possíveis para A, estima-se P(A|B) :
- P(A=a|B=b) = [N(A=a,B=b)+1]/[N(B=b)+V]

Decidindo baseado em probabilidades...

- A teoria da decisão sob incerteza é bastante ampla e tem suas bases em Teoria da Utilidade Esperada e na hipótese de agentes racionais oriundas da área de economia

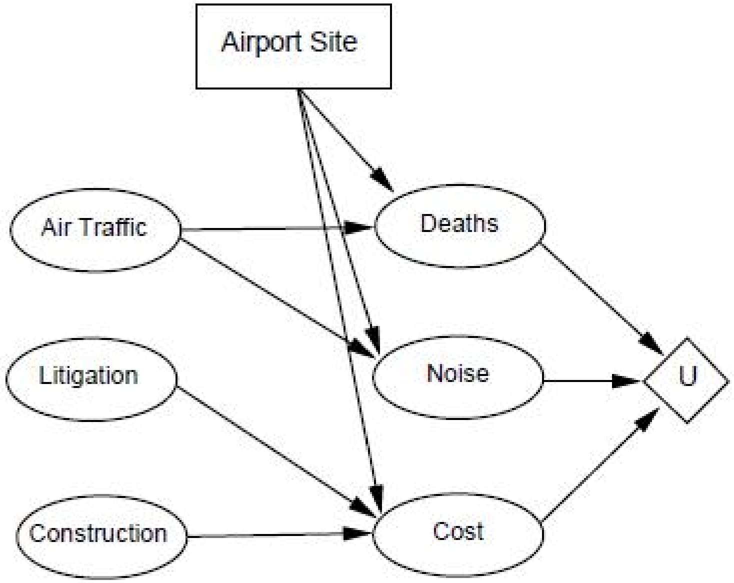
- Vamos falar um pouco sobre isso sobre o ponto de vista de quem deseja automatizar a decisão (fazer um agente que decida de acordo com as preferências de seu proprietário…)

Exemplo: Escolha da localização de um aeroporto

- Dependendo da posição pode-se alterar:
  - O risco de acidentes (logo, o número esperado de mortes.. **Deaths**
  - O incômodo causado pelo barulho dos aviões, quanto mais próximo de uma cidade pior….**Noise**
  - É fácil perceber que **Deaths** e **Noise** serão diretamente afetados pelo volume de **tráfego aéreo** no aeroporto.
  - Naturalmente, o custo também é alterado pela localização do aeroporto (Cost)
    - a desapropriação de um determinado terreno pode ser mais ou menos litigioso e os custos de ligação de transportes do aeroporto a cidade podem ser maiores ou menores afetando a **construção**

Redes de Decisão (Decision Networks)

Add action nodes and utility nodes to belief networks
to enable rational decision making



Algorithm:

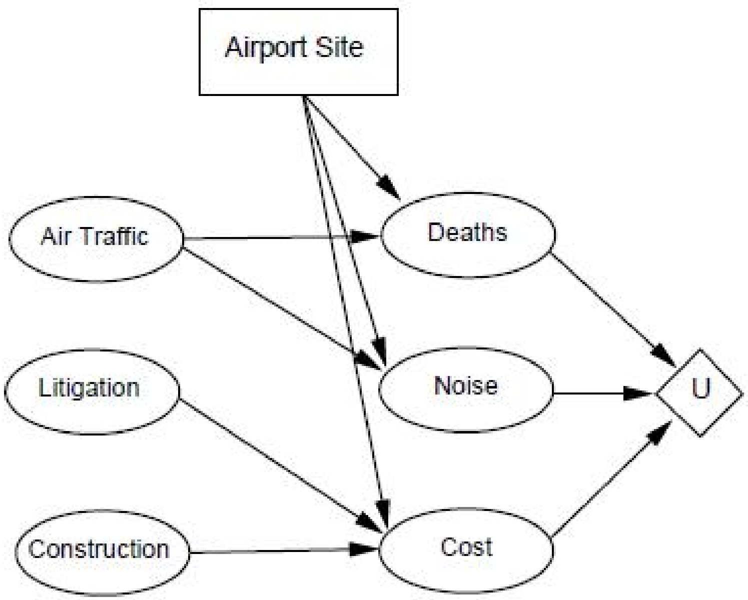
For each value of action node

compute expected value of utility node given action, evidence

Return MEU action

Podemos determinar distribuições de variáveis, mas como decidir?

- Nós de ação e nós de utilidade na rede;

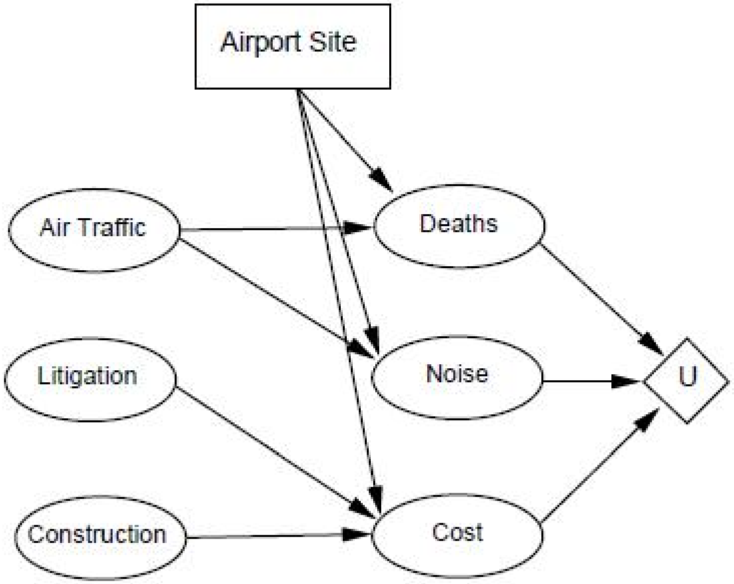


<u>Optimal decisions</u>: decision networks include utility information;
probabilistic inference required for $P(outcome|act\ on, ev\ dence)$

Processo de Decisão...

- pi = P(deaths=i| ASite=s, Noise=n) Ou
- P(outcome| action, evidence)

- Utilidade Esperada (action=a) = $\sum_i$ U(outcome$_i$)*P(outcome $_i$|action=a, evidence)
- Escolher ação que maximiza a utilidade esperada



Resumo

- Regra de Bayes permite que probabilidades desconhecidas sejam calculadas a partir de probabilidades condicionais conhecidas (no sentido causal geralmente)

- Independência condicional pode permitir a fatoração da distribuição conjunta total em distribuições condicionais menores. O modelo Bayes ingênuo pressupõe a independência do efeito dada uma única variável causal

- Eliminação de variáveis pode ser bastante eficiente em redes unicamente conectadas (apenas um caminho entre dois nós), mas são intratáveis em redes multiplamente conectadas

- Voltaremos a estudar construção de redes Redes Bayesianas e cálculo de probabilidades quando estivermos falando de Aprendizado

- Há outras alternativas a abordagem probabilística para tratar incerteza, Teoria de Dempster-Shafer, Lógica Nebulosa (difusa ou fuzzy)

- Lógica Fuzzy será discutida no final do próximo bimestre