DEIS - Departamento de Engenharia Informática e de Sistemas
ISEC - Instituto Superior de Engenharia de Coimbra

Conhecimento e Raciocínio 2015/2016

Prática 6 - Redes Bayesianas

---

1. Suponha que num dado país

- 1% da população tem tuberculose
- Uma radiografia é positiva (indica a presença da doença) em 95% dos casos em que a doença foi efectivamente contraída
- Uma radiografia é interpretada como positiva em 0.5% dos casos que afinal se verificou não serem efectivamente de tuberculose

a) Desenhe uma Rede Bayesiana destinada a inferir a probabilidade de presença ou ausência de tuberculose dado um resultado positivo ou negativo da radiografia.
b) Calcule a probabilidade de ter tuberculose quando a radiografia é positiva
c) Calcule a probabilidade de não ter tuberculose quando a radiografia é positiva
d) Calcule a probabilidade de ter tuberculose quando a radiografia é negativa
e) Calcule a probabilidade de não ter tuberculose quando a radiografia é negativa
f) Qual o diagnóstico mais fiável ?

1. a) Podemos considerar os seguintes acontecimentos:

Acontecimento B = Indivíduo tem tuberculose
$P(B)=0.01$

Acontecimento ¬B=Indivíduo não tem tuberculose
$P(\neg B)=1-P(B)=1-0.01=0.99$

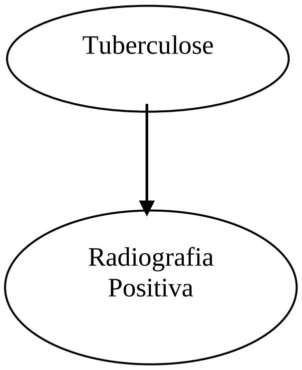
Acontecimento A = Radiografia é positiva
A probabilidade P(A) deste acontecimento não é conhecida porque:

1. Sabe-se que "suposto que o indivíduo tem tuberculose, a radiografia é positiva 95% das vezes". Portanto é $P(A|B)$ e não $P(A)$.
2. Sabe-se que "suposto que o indivíduo não tem tuberculose, a radiografia é positiva 0.5% das vezes". Portanto é $P(A|\neg B)$ e não $P(A)$.

Em resumo:

$P(A|B) = 0.95 \Rightarrow P(\neg A|B) = 0.05$
$P(A|\neg B) = 0.005 \Rightarrow P(\neg A|\neg B) = 0.995$

Acontecimento ¬A = Radiografia é negativa
Sobre este acontecimento nada se sabe. Contudo, os dados anteriores já permitem construir a rede, como se segue:



	True	False
Tuberculose	0.01	0.99

Radiog / Tub.	**Tub.True**	**Tub.False**
Rad.Positiva	0.95	0.005
Rad.Negativa	0.05	0.995

b) A Fórmula de Bayes é $P(B \mid A) = \dfrac{P(A \mid B).P(B)}{P(A)}$

e portanto neste caso permite calcular a probabilidade P(tuberculose | Rad.positiva) a partir de

P(A|B) = P(rad.positiva | tuberculose) = 0.95
P(B) = P(tuberculose) = 0.01
P(A) = P(rad.positiva) = desconhecida

Portanto, é preciso manipular o denominador para se exprimir P(A) de outra forma. Ora, dada uma população, um indivíduo está tuberculoso ou não está. Portanto, B e $\neg$B são acontecimentos que cobrem exaustivamente o espaço amostral e são independentes, porque a probabilidade de um não afecta a do outro. Então podemos escrever:

A= A $\cap$ (B $\cup$ $\neg$B)
A= (A $\cap$ B) $\cup$ (A $\cap$ $\neg$B)

Ou, em termos de probabilidades P(A)=[P(A) $\cap$ (P(B)] + [P(A) $\cap$ P($\neg$B)]

Ora, como (ver slides e bibliografia) $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ temos

$$P(A \cap B) = P(A \mid B).(P(B)$$

Substituindo acima, obtemos:

$$P(A) = P(A \mid B).(P(B) + P(A \mid \neg B).(P(\neg B)$$

Substituindo no denominador da Fórmula de Bayes, obtemos:

$$P(B \mid A) = \frac{P(A \mid B).P(B)}{P(A \mid B).P(B) + P(A \mid \neg B).(P(\neg B)}$$

Como sabemos:

P(B)=0.01
P(A|B) = 0.95
P(A|$\neg$B) = 0.005

P($\neg$B)=1-P(B)=1-0.01=0.99
P($\neg$A|B) = 0.05
P($\neg$A|$\neg$B) = 0.995

$$
\begin{aligned}
P(tuberculose \mid positiva) &= \\
&= P(B \mid A) \\
&= \frac{P(A \mid B).P(B)}{P(A \mid B).P(B) + P(A \mid \neg B).(P(\neg B)} \\
&= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.005 \times 0.99} \\
&= 0.657 \\
&= 65.7\%
\end{aligned}
$$

c) Neste caso pretendemos calcular P(¬B|A). Para isso basta substituir B por ¬B na fórmula acima, obtendo-se:

$$
\begin{aligned}
P(\neg B \mid A) &= \frac{P(A \mid \neg B).P(\neg B)}{P(A \mid \neg B).P(\neg B) + P(A \mid B).(P(B)} = \\
&= \frac{0.005 \times 0.99}{0.005 \times 0.99 + 0.95 \times 0.01} \\
&= 0.343
\end{aligned}
$$

d) Neste caso pretendemos calcular P(B|¬A). Para isso basta substituir A por ¬A na fórmula acima, obtendo-se:

$$
\begin{aligned}
P(B \mid \neg A) &= \frac{P(\neg A \mid B).P(B)}{P(\neg A \mid B).P(B) + P(\neg A \mid \neg B).(P(\neg B)} \\
&= \frac{0.05 \times 0.01}{0.05 \times 0.01 + 0.995 \times 0.99} \\
&= 0.000507
\end{aligned}
$$

e) Neste caso pretendemos calcular P(¬B|¬A). Para isso basta substituir A por ¬A e B por ¬B na fórmula acima, obtendo-se:

$$
\begin{aligned}
P(\neg B \mid \neg A) &= \frac{P(\neg A \mid \neg B).P(\neg B)}{P(\neg A \mid \neg B).P(\neg B) + P(\neg A \mid B).(P(B)} \\
&= \frac{0.995 \times 0.99}{0.995 \times 0.99 + 0.05 \times 0.01} \\
&= 0.99949
\end{aligned}
$$

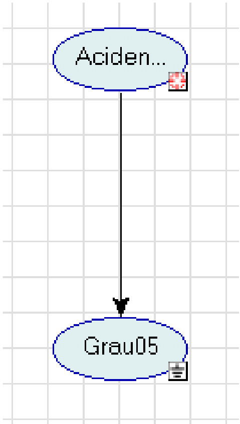
f) O diagnóstico mais fiável, e praticamente certo, é o de ausência de tuberculose quando a radiografia for negativa (0.99949) dado que para um diagnóstico positivo baseado numa radiografia positiva, a probabilidade de tuberculose efectiva é de apenas 0.657.

2. Considere os seguintes dados:

- Numa população, 1 em cada 1000 indivíduos tem um acidente de automóvel.
- Dos que já tiveram um acidente, 70% tinha taxa de alcolémia > 0.5 g/l
- Dos que nunca tiveram um acidente, 1% tem taxa de alcolémia > 0.5g/l

a) Desenhe uma rede Bayesiana destinada a modelar este problema, mostrando em cada nó a tabela de probabilidades a ele associadas.
b) Calcule a probabilidade de acidente para taxa > 0.5 g/l
c) Calcule a probabilidade acidente para taxa < 0.5 g/l
d) Que conclusão pode tirar de a) b) e c) ?

2. a) São necessários 2 nós: Um para acidente e outro para alcolémia. As probabilidades que são dadas são as condicionais de (alcolémia | acidente) e portanto a seta deve apontar de acidente para alcolémia. De acordo com o enunciado as tabelas de probabilidade são as assinaladas na figura.



	P(Acidente)
T	0.001
F	0.999

Acidente	P(A>0.5 | Acidente)	P(A<0.5 | Acidente)
T	0.70	0.30
F	0.01	0.99

b) Como a rede só tem 2 nós, os cálculos podem realizar-se recorrendo à seguinte versão da regra de Bayes (ver trabalho prático sobre Redes Bayesianas e bibliografia):

$$P(A|B) = \frac{P(B|A).P(A)}{P(B|A).P(A) + P(B|\neg A).P(\neg A)}$$

Neste caso, P(A) é a probabilidade de ocorrer um acidente, que representaremos por P(D) (o "D" vem de "desastre", para não se confundir com "A" de alcolémia); e B representa a existência de alcolémia, que representaremos simplesmente por A. Assim:

$$P(D|A) = \frac{P(A|D).P(D)}{P(A|D).P(D) + P(A|\neg D).P(\neg D)}$$

De acordo com o enunciado temos:

P(A|D)	=0.7
P(D)	=0.001
P(A|¬D)	=0.01
P(¬D)	=0.999

$$P_1(D|A) = \frac{0.7x0.001}{0.7x0.001 + 0.01x0.999} = \frac{0.0007}{0.0007 + 0.00999} = 0.06548$$

c) As tabelas de probabilidade são as mesmas:

	P(Acidente)
T	0.001
F	0.999

Acidente	P(A>0.5 | Acidente)	P(A<0.5 | Acidente)
T	0.70	0.30
F	0.01	0.99

Porém, o acontecimento a considerar é agora ¬A (não alcolémia). Portanto, na fórmula:

$$P(D|A) = \frac{P(A|D).P(D)}{P(A|D).P(D) + P(A|\neg D).P(\neg D)}$$

substituimos A por ¬A. Obtemos:

$$P(D|\neg A) = \frac{P(\neg A|D).P(D)}{P(\neg A|D).P(D) + P(\neg A|\neg D).P(\neg D)}$$

P(¬A | D) =0.3
P(D) =0.001
P(¬A | ¬D) = 0.99
P(¬D) =0.999

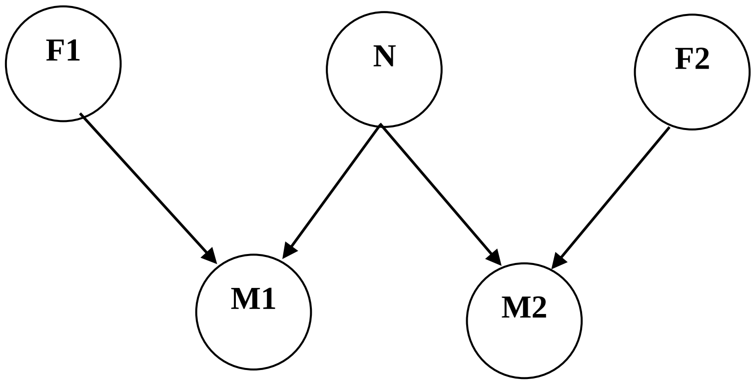
$$P_2(D|\neg A) = \frac{0.3x0.001}{0.3x0.001 + 0.99x0.999} = 0.0030$$

d) Em resumo:

P(acidente | alcolémia > 0.5) = 0.065
P(acidente | alcolémia < 0.5) = 0.003

Portanto, é evidente que para a população considerada, uma taxa de 0,5 g/l agrava a probabilidade de ocorrer um acidente.

3. Dois astrónomos, em diferentes partes do mundo, fazem medidas M_1 e M_2 do número de estrelas N que observam numa dada região do espaço, usando telescópios. Contudo, devido a interferências de vária natureza, existe uma pequena probabilidade de erro na contagem. Além disso, cada um dos telescópios tem uma certa probabilidade de se encontrar ligeiramente desfocado. Sejam estes eventos F_1 e F_2. Desenhe uma rede Bayesiana destinada a modelar este problema.



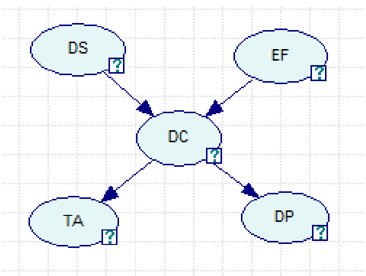
REVISÃO - Redes Bayesianas, Regras e CBR - Exame 1ª Chamada 2008/2009

1. Desenhe uma Rede Bayesiana capaz de inferir a probabilidade de uma dada pessoa ter uma doença do coração (DC) sabendo que:

- O exercício físico (EF) diminui a probabilidade de DC
- Uma dieta adequada (DA) também diminui a probabilidade de DC
- A doença do coração pode originar dois sintomas:
  - Tensão alta (TA)
  - Dor no peito (DP).

Junto de cada nó coloque a respectiva tabela de probabilidade preenchendo-a com as classificações que achar adequadas, com base na seguinte definição de termos linguísticos para o valor da probabilidade, *p*:

- B = probabilidade Baixa
- M = probabilidade Média
- A = probabilidade Alta
- MA = probabilidade Muito Alta



DS
B

EF
B

DC	
DS=F, EF=F	MA
DS=F, EF=T	A
DS=T, EF=F	M
DS=T, EF=T	B

TA	
DC=F	B
DC=T	A

DP	
DC=F	B
DC=T	A

2. Para o problema descrito em 1. implemente um sistema baseado em regras que permita inferir se um determinado indivíduo tem uma doença do coração.

No final de cada regra coloque entre parêntesis o respectivo Factor de Certeza que considerar mais adequado, utilizando a seguinte simbologia:

- B = certeza Baixa
- M = certeza Média
- A = certeza Alta

If DS then not DC (B)
If EF then not DC (B)
If TA then DC (A)
If DP then DC (M)

Outros FC, desde que lógicos, também seriam aceites

Notar que p.e. a regra If not DS then DC não está completamente correcta: o enunciado diz "uma dieta saudável (DS) diminui a probabilidade de DC", mas NÃO DIZ que "uma dieta não saudável aumenta a probabilidade de DC" (Slides teóricas, aquisição do conhecimento, secção 2.3.5)

3. Para o problema descrito em 1. implemente um sistema baseado no paradigma CBR, através:

- Da definição de 2 Known-Cases adequados ao início de funcionamento do sistema
- Da definição das tabelas de semelhança que considerar necessárias
- Para cada atributo considere 3 graduações possíveis designadas por X, Y, Z
- Para cada entrada das tabelas os valores de semelhança 0, 0.5 e 1, que deve usar atendendo à ordenação alfabética de X, Y e Z.

Os casos adequados ao início de funcionamento do sistema devem cobrir pelo menos 80% das soluções possíveis (Slides teóricas, secção 6.4). Portanto tem de haver um caso para DC=T e outro para DC=F. Os atributos a considerar são DS, EF, TA e DP.

Caso 1: DS=F, EF=F, TA=T, DP=T DC=T
Caso 2: DS=T, EF=T, TA=F, DP=F DC=F

Cada atributo tem 3 níveis possíveis designados genericamente por X, Y e Z, pelo que a tabela pode ser única:

	X	**Y**	**Z**
X	1	0,5	0
Y	0,5	1	0,5
Z	0	0,5	1

(seriam aceites soluções com 4 tabelas iguais, uma para cada atributo)

REVISÃO - Redes Bayesianas, Regras e CBR - Exame Setembro 2008/2009

1. Considere uma rede Bayesiana destinada a inferir a probabilidade de um determinado cliente marcar ou não um quarto num determinado hotel. As variáveis a considerar são as seguintes:

Perfil_Hotel = A, B, C
Categoria_Hotel = S, NS (Semelhante ou Não Semelhante à pretendida pelo cliente)
Preço_Hotel = S, NS (Semelhante ou Não Semelhante ao pretendido pelo cliente)
Localização_Hotel = Z, NZ (na Zona ou afastado da Zona pretendida pelo cliente)
Vaga_Hotel = S, N (Existe ou não existe um quarto vago)

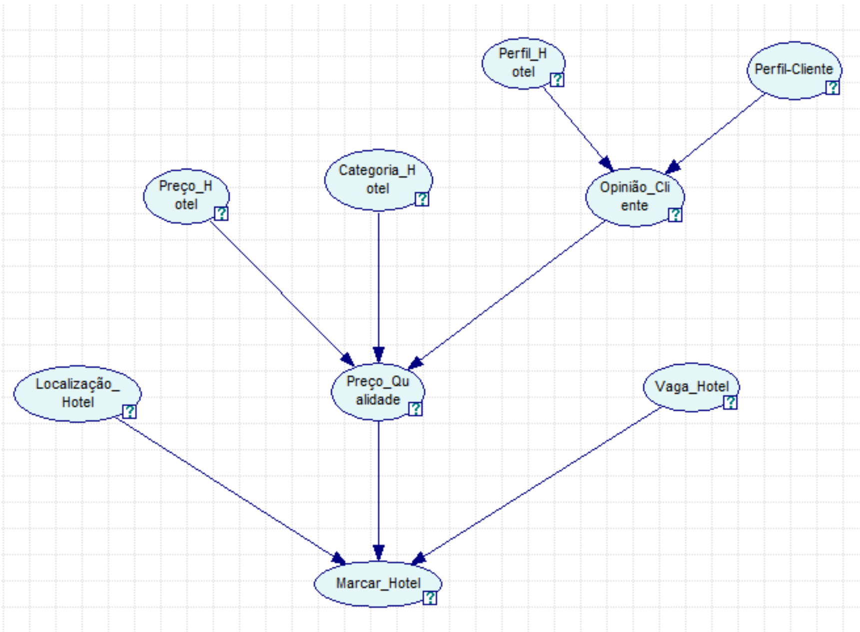
Perfil_Cliente = A, B, C
Opinião_Cliente = B, M (Boa ou Má)
Preço_Qualidade = A, M, B (Alta, Média, Baixa)

Marcar_Hotel = S, N é a variável alvo (*target*)

Se o perfil do cliente coincidir com o perfil do hotel a sua opinião será boa. As variáveis que influenciam directamente Marcar_Hotel são Localização_Hotel, Preço_Qualidade e Vaga_Hotel.

a) Desenhe a rede Bayesiana.

De acordo com o enunciado Perfil_Hotel e Perfil_Cliente determinam Opinião_Cliente. Também de acordo com o enunciado, Localização_Hotel, Preço_Qualidade e Vaga-Hotel determinam Marcar_Hotel. Uma vez englobadas estas variáveis na rede, as ligações entre as restantes são evidentes. A rede resultante é a seguinte:



b) Desenhe uma tabela de probabilidade associada ao nó Marcar_Hotel. Os valores indicados são à sua escolha mas devem ser razoáveis, isto é, de acordo com aquilo que é previsível que aconteça.

Nó Marcar_Hotel

Localização_...	Z						NZ					
Preço_Qualida...	A		M		B		A		M		B	
Vaga_Hotel	S	N	S	N	S	N	S	N	S	N	S	N
Marca	0.9	0	0.7	0	0.5	0	0.5	0	0.3	0	0.1	0
NaoMarca	0.1	1	0.3	1	0.5	1	0.5	1	0.7	1	0.9	1

A probabilidade 0 é determinada pelo facto de que se não houver vaga, o hotel não é marcado, de certeza. É fundamental que este valor figure na tabela acima (assim como o seu complementar, valor 1, para NãoMarca)

c) Desenhe a tabela de probabilidade associada ao nó Perfil-Hotel sabendo que há 50% de hotéis com o perfil A, 30% com o perfil B e 20% como perfil C.

Nó Perfil_Hotel

A	0.5
B	0.3
C	0.2

2. **a)** De acordo com a filosofia base do sistema de desenvolvimento utilizado nos laboratórios para o paradigma CBR, defina os 3 conceitos de natureza hierárquica, entre si, que o enunciado apresentado em 1. sugere. Assinale a relação de hierarquia por meio de setas entre os referidos conceitos.

Opinião_Cliente → Preço_Qualidade → Hotel_Marcado

b) Dê exemplo de 2 casos para cada um dos conceitos definidos em 2a)

Conceito Opinião_Cliente

Atributos		Solução
Perfil_Cliente	Perfil_Hotel	Opinião_Cliente
A	A	B
A	C	M

Conceito Preço_Qualidade

Atributos			Solução
Categoria_Hotel	Preço_Hotel	Opinião_Cliente	Preço_Qualidade
S	S	B	A
NS	NS	M	B

Conceito Marcar_Hotel

Atributos			Solução
Localização_Hotel	Preço_Qualidade	Vaga_Hotel	Marcar_Hotel
Z	A	S	S
NZ	B	N	N

NOTA: Em cada um dos conceitos optou-se pelos 2 casos extremos, i.e., todos os atributos favoráveis ou todos os atributos desfavoráveis, resultando para a solução, respectivamente, um valor máximo e ou valor mínimo. No conceito Marcar_Hotel outro caso típico que poderia ser considerado é

Z	A	**N**	N

que exemplifica o caso em que tudo é favorável à marcação do hotel excepto a vaga que não existe, pelo que neste caso a solução seria N, de certeza.

c) Algum atributo deve ser mandatório (*mandatory*)? Justifique.

Sim, o atributo Vaga-Hotel deveria ser mandatório. Trata-se de arranjar um sistema que de certeza exclua todos os hotéis que não têm vaga: tipicamente um sistema CBR apresenta várias soluções ordenadas por grau de compatibilidade com o query-case. Para garantir a exclusão dos hotéis que não têm vaga, o query-case poderia implicitamente (por meio de uma regra de conclusão, por exemplo) assumir sempre Vaga-Hotel=S. Nesta abordagem, e sendo o atributo Vaga-Hotel mandatório, este facto implicaria que todos os casos com Vaga-Hotel=N seriam excluídos. Apenas um sistema deste tipo garantiria a exclusão de hotéis sem vaga.

3. **a)** Escreva um conjunto de 7 regras (e não mais) indicando os respectivos factores de certeza (FC) e destinadas a calcular a certeza com que um hotel será marcado por um dado cliente. Na ausência de factores de certeza imperativos utilize FC=0,8.

1. IF Perfil_Hotel=Perfil_Cliente THEN Opiniao_Cliente=B FC=0,8
2. **IF Preço_Hotel=S THEN Preço_Qualidade=A FC=0,8**
3. **IF Categoria_Hotel=S THEN Preço_Qualidade=A FC=0,8**
4. **IF Opinião_Cliente=B THEN Preço_Qualidade=A FC=0,8**
5. **IF Preço-Qualidade=A THEN Marcar_Hotel=S FC=0,8**
6. **IF Localização_Hotel=Z THEN Marcar_Hotel=S FC=0,8**
7. IF Vaga_Hotel=S THEN Marcar_Hotel=S FC=0,8
8. IF Vaga_Hotel=N THEN Marcar_Hotel=N FC=1

A regra 8 não seria necessária dado o enunciado referir a conclusão Marcar_Hotel (i.e. positiva, marcação possível). A regra 8 poderia no entanto ser admissível, mas com FC=1 dado que se o hotel não tem vaga, NÃO pode ser marcado, de certeza.

b) Com base no sistema criado em 3 a) calcule o grau de certeza, para todos os desfechos possíveis, com que um dado hotel será marcado, admitindo que são conhecidos os seguintes factos todos com FC=1:

Preço_Hotel = S
Categoria_Hotel = S
Opinião_Cliente = B
Localização_Hotel = Z

$$CF_{combine}(CF_1, CF_2) = \begin{cases} CF_1 + CF_2 - CF_1.CF_2 & CF_1 \geq 0 \wedge CF_2 \geq 0 \\ \dfrac{CF_1 + CF_2}{1 - \min(|CF_1|, |CF_2|)} & CF_1 < 0 \textbf{ ou (exlusivo)} CF_2 < 0 \\ CF_1 + CF_2 + CF_1.CF_2 & CF_1 < 0 \wedge CF_2 < 0 \end{cases}$$

As regras que intervêm no cálculo estão marcadas a Bold na alínea anterior. Pelo disparo de cada uma delas resultam os seguintes factores de certeza para as conclusões:

R2: Preço_Qualidade=0.8*1=0,8
R3: Preço_Qualidade=0.8*1=0,8
R4: Preço_Qualidade=0.8*1=0,8

Combinação das mesmas conclusões (fórmula 1)
R2+R3: Preço_Qualidade=0,8+0,8-0,8*0,8=1,6-0,64=0,96
Anterior+R4 Preço_Qualidade=0,96+0,8-0,96*0,8=0,992

R5: Marcar_Hotel=0,8*0,992=0,7936
R6: Marcar_Hotel=0.8*1=0,8

Combinação das mesmas conclusões (fórmula 1)
R5+R6: Marcar_Hotel=0,7936+0,8-0,8*0,7936=0,956

Nada é dito sobre Vaga-Hotel pelo que há que considerar, para termos todos os desfechos possíveis conforme refere o enunciado, Vaga-Hotel=S e Vaga-Hotel=N, ou seja, aqui há que considerar também a regra R8:

IF Vaga_Hotel=N THEN Marcar_Hotel=N com **FC=1**
Ou, equivalentemente:
IF Vaga_Hotel=N THEN Marcar_Hotel=S mas com **FC=-1** (certeza da conclusão contrária)

Desfecho1:
Vaga_Hotel=S com FC=0.8, dispara R7
R7: Marcar_Hotel=1*0,8=0,8
Combinação de conclusões iguais (fórmula 1):
R7+ (R5+R6) Marcar_Hotel=0,8+0,956-0,8*0,956=0,991

Desfecho2:
Vaga_Hotel=N com FC=1, dispara R8
R8: Marcar_Hotel=1*-1=-1
Combinação de conclusões de sinal contrário (fórmula 2):

R8+ (R5+R6) $$CF(Marcar\_Hotel) = \frac{CF_1 + CF_2}{1 - \min(|CF_1|, |CF_2|)} = \frac{\mathbf{0{,}956 + (-1)}}{\mathbf{1 - \min(|0{,}956|, |-1|)}} = \frac{-0{,}044}{1 - 0{,}956} = \frac{-0{,}044}{0{,}044} = -1$$

c) Atente nos resultados obtidos em b). Parecem-lhe correctos? Prove que a filosofia MYCIN garante resultados correctos nas 3 situações limite do tipo da que este caso sugere.

Os resultados parecem correctos. Chama a atenção o facto de, apesar de todos os factores serem favoráveis, a filosofia MYCIN garantir a atribuição de um CF de -1 (certeza absoluta de que o hotel não é marcado) caso não exista vaga. Isto acontece sempre que uma regra dá como conclusão um facto de FC=1 ou FC=-1. Qualquer das 3 fórmulas de combinação de conclusões (dadas no formulário de a)) exibe esta propriedade, muito interessante e adequada à realidade, dado que se uma conclusão é certa de certeza absoluta…então nada deverá conseguir enfraquecê-la. O funcionamento das 3 fórmulas, neste aspecto, pode ser provado da seguinte forma:

Seja FC1=1 e FC2=X tal que X é positivo (e menor que 1)
Fórmula 1: FC=FC1+FC2-FC1.FC2= 1+X-1*X = 1+X-X= **1**

Seja FC1=-1 e FC2=X tal que X é negativo (e de módulo menor que 1)
Fórmula 3: FC=FC1+FC2+FC1.FC2= -1+X-1*X = -1+X-X= **-1**

Seja FC1=1 e FC2=X tal que X é negativo (e de módulo menor que 1)
Fórmula 2: $$CF = \frac{CF_1 + CF_2}{1 - \min(|CF_1|, |CF_2|)} = \frac{1 + X}{\mathbf{1 - min}(|\mathbf{1}|, |X|)} = \frac{1 - |X|}{1 - X} = \frac{1 - X}{1 - X} = 1$$

Seja FC1=-1 e FC2=X tal que X é positivo (e de módulo menor que 1)
Fórmula 2: $$CF = \frac{CF_1 + CF_2}{1 - \min(|CF_1|, |CF_2|)} = \frac{-1 + X}{\mathbf{1 - min}(|\mathbf{-1}|, |X|)} = \frac{-1 + X}{1 - X} = -1$$

Conclusão: quando existe um FC de valor 1 ou -1, o seu valor "absorve" qualquer outro, mantendo-se a conclusão da combinação das regras sempre com o valor desse FC (+1 ou -1.