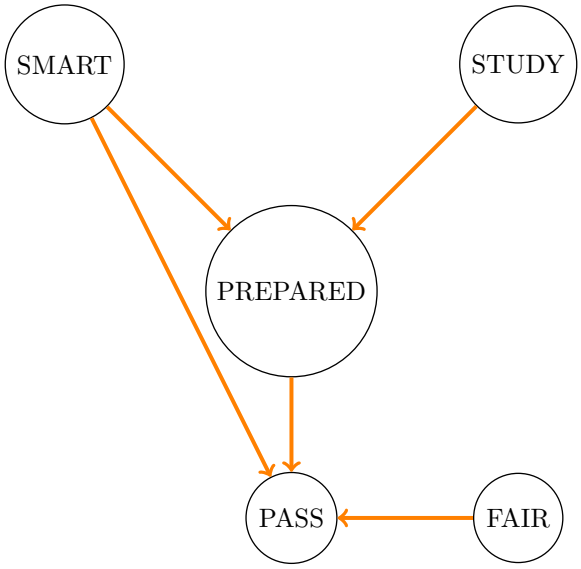
1 Esercitazione 4

1.1 Esercizio 1



P(SM)	
0.8	0.2

P(ST)	
0.6	0.4

P(FA)	
0.69	0.1

SM	ST	P(PR | SM, ST)	
T	T	0.9	0.1
T	F	0.5	0.5
F	T	0.7	0.3
F	F	0.1	0.9

FA	PR	SM	P(PA | PR, SM, ST)	
T	T	T	0.9	0.1
T	T	F	0.7	0.3
T	F	T	0.7	0.3
T	F	F	0.2	0.8
F	T	T	0.1	0.9
F	T	F	0.1	0.9
F	F	T	0.1	0.9
F	F	F	0.1	0.9

Dobbiamo generare 1000 sample con direct sampling.
Consideriamo come ordine topologico SM, ST, PR, FA, PA
Genera un campione vuol dire generare un valore per ciascuna della variabile che lo compone.
Supponiamo di poter creare un campione S_1 [0.35; 0.76; 0.51; 0.44; 0.08]
Iniziamo a creare il campione di rete Bayesiana.
Andiamo a vedere dove cade il valore 0.35 sulla distribuzione, e notiamo che essendo <= di 0.8 avremo che $SM = T$;
Analogamente per ST avremo 0.76, che essendo > di 0.6 avremo che $ST = F$;
Andiamo ora a calcolare PR: nel nostro caso dobbiamo vedere la riga della CPT quando $SM = T$; e $ST = F$;
Essendo il valore estratto $0.51 > 0.5$ concludiamo che $PR = F$;
Per FA come valore abbiamo 0.44 che essendo < 0.69 ci porta a dedurre che $FA = T$;
Essendo $0.08 < 0.7$ concludiamo che $PA = T$
Il vettore finale sarà quindi:

$$S1 \ = \ [SM = T;\ ST = F;\ PR = F;\ FA = T;\ PA = T]$$

La condizione per essere true è di essere <= al valore nella CPT, quindi se la distribuzione è 0.5 e 0.5 se si estrae 0.5 conta come true

Immaginiamo di avere fatto questo procedimento per mille campioni.
Avremo che:

- 790 $SM = T$
- 210 $SM = F$

Quindi che la distribuzione della variabile è <0.79, 0.21>
Questo metodo di campionamento (campionamento diretto) è utile quando non conoscono nessun valore di nessuna variabile; nel momento in cui conosco il valore di qualche variabile questo metodo risulta poco efficiente perchè potrei ottenere dei valori impossibili rispetto all'evidenza.
Utilizzando sempre la rete bayesiana dell'esercizio precedente, supponiamo di dover trovare

$$P(PA = T; \mid ST = F)$$

Questa richiesta con il direct sampling non sarebbe efficente perchè andremmo a generare dei valori per la variabili ST a true, che non sarebbero utili perchè in contrapposizione con l'evidenza.
Possiamo utilizzare il rejection sampling, utilizzando come ordinamento topologico SM, ST, PR, FA, PA
Utilizziamo ancora una volta i valori utilizzati in precedenza, ovvero S_1 [0.35; 0.76; 0.51; 0.44; 0.08]
Nel campione S1 avevamo che

$$S1 \; = \; [SM = T; \; ST = F; \; PR = F; \; FA = T; \; PA = T]$$

Essendo $ST = F$ coerente con l'evidenza vuol dire che accettiamo il campione.
Supponiamo ora di avere: S_2 [0.28; 0.03; 0.92; 0.92; 0.42]
Avremo quindi:

$$S2 \; = \; [SM = T; \; ST = T, \; PR = F \; FA = F \; PA = F]$$

Dobbiamo però rigettare il campione perchè il valore non è coerente con l'evidenza.
Supponiamo a questo punto di avere generato 1000 campioni:

- 730 $ST = T$ -> reject
- 270 $ST = F$ -> accept

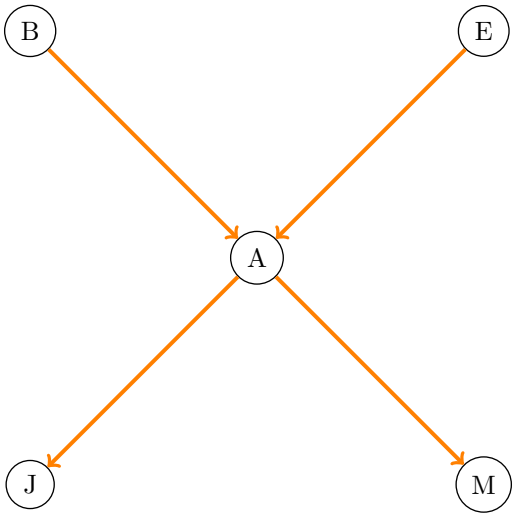
Dobbiamo considerare solo i 270 accettati, supponiamo che siano distribuiti in questa maniera:

- 130 $PA = T$
- 140 $PA = F$

Avremo quindi che

$$P(PA|ST = F) = \alpha \cdot < 130; \; 140 >=< 0.48; 0.52 >$$

1.2 Esercizio 2



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A | B, E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(M | A)

A	T	F
T	0.7	0.3
F	0.01	0.99

P(J | A)

A	T	F
T	0.9	0.1
F	0.05	0.95

- Trovare la probabilità $P(B = T|J = T, M = F)$ con Likelihood weighting

Anche per likelihood weighting dobbiamo utilizzare un ordine topologico, utilizziamoB, E, A, J, M
I campioni saranno:

$$S1 = [0.9994;\ 0.89;\ 0.3;\ /;\ /]$$

$$S2 = [0.7;\ 0.9;\ 0.6;\ /;\ /]$$

Rispetto al campione S1 avremo che:

$$S1 = [T; F; T; T; F]$$

A questo punto dobbiamo calcolare il peso per S1, ovvero la produttoria delle variabili per le quali conosciamo l'evidenza condizionate ai genitori:

$$w_{S_1} = P(J = T|A = T) \cdot P(M = F|A = T) = 0.9 \cdot 0.3 = 0.27$$

Analogamente avremo per S2:

$$S2 = [F; F; F; T; F]$$

Andiamo a stimare il peso:

$$w_{S_2} = P(J = T|A = F) \cdot P(M = F|A = F) = 0.05 \cdot 0.99 = 0.0495$$

Andiamo infine a stimare il valore richiesto che risulta essere:

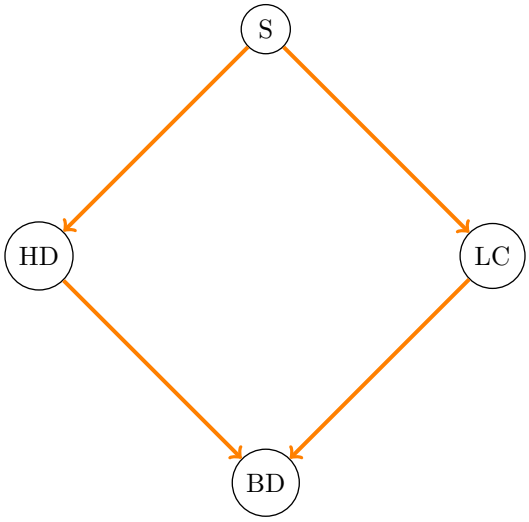
$$P(B = T|J = T, M = F) = \frac{\sum_{S_i:B=T} w_{s_i}}{\sum_{S_i} w_{s_i}}$$

Stiamo mettendo a rapporto i casi in cui la nostra variabile query assume valore uguale a vero rispetto a tutti i casi.

avremo quindi:

$$\frac{\sum_{S_i:B=T} w_{s_i}}{\sum_{S_i} w_{s_i}} = \frac{0.27}{0.27 + 0.0495} = 0.8450$$

1.3 Esercizio 3



P(S=T)	P(S = F)
0.4	0.6

S	P(HD = T)	P(HD = F)
T	0.4	0.6
F	0.3	0.7

S	P(LC = T)	P(LC = F)
T	0.7	0.3
F	0.2	0.8

HD	LC	P(BD = T)	P(BD =
T	T	0.4	0.6
F	T	0.7	0.3
T	F	0.5	0.5
F	F	0.2	0.8

La richiesta è di stimare con il metodo markov chain montecarlo la distribuzione di probabilità di $P(HD|S = T, BD = T)$.

Con MCMC non è necessario mantenere l'ordine topologico.

Il MCMC è diviso in tre passi:

- Fisso le evidenze (S e BD nel nostro caso)
- Inizializzo tutte le variabili senza evidenza in modo casuale, ad esempio nel nostro caso prendiamo LD = T e HD = F Avremo così lo stato $S_0 = [S = T, HD = F, LC = T, BD = T]$; sarebbe stato equivalente avere come stato iniziale $[S = T, LC = T, HD = F, BD = T]$ o altre permutazioni.
- Campionare le variabili non osservate, sia quelle nascoste che la variabili query, data la relativa markov blanket. Per ciascuna delle variabili non osservate dovrò determinare ad ogni iterazione:

$$P(X|MB(X)) \ = \alpha \cdot P(X|Pa(X)) \cdot \prod_{y \in children(x)} P(y|Pa(y))$$

  Dove MB è la markov blanket della variabile X

Iniziamo a a campionare LC data la sua MB, dobbiamo stimare sia la parte vera che la parte falsa:

- $P(LC = T|S = T, HD = F, BD = T)$
  I valori sono quelli presi dallo stato S_0.

$$= \alpha \cdot P(LC = T|S = T) \cdot P(BD = T|LC = T, HD = F) =$$

$$= \alpha \cdot 0.7 \cdot 0.5 = 0.35$$

- $P(LC = F|S = T, HD = F, BD = T)$

$$= \alpha \cdot P(LC|F|S = T) \cdot P(BD = T|LC = F, HD = F) =$$

$$= 0.3 \cdot 0.2 = 0.06$$

$$P(LC|MB(LC)) = \alpha < 0.35; 0.06 >=< 0.853; 0.147 >$$

Supponiamo di generare un numero casuale e di ottenere un valore che ci porta a dire che $LC = F$.
Avremo a questo punto un nuovo stato S_1:
$S_1 = [S = T; HD = F; LC = F; BD = T;]$
A questo punto dobbiamo calcolare $P(HD|MB(HD))$

- $P(HD = T|S = T, LC = F, BD = T)$

$$= \alpha \cdot P(HD = T|S = T) \cdot P(BD = T|HD = T, LC = F) =$$

$$= \alpha \cdot 0.4 \cdot 0.7 = 0.28$$

- $P(HD = F|S = T, LC = F, BD = T)$

$$= \alpha \cdot P(HD = F|S = T) \cdot P(BD = T|HD = F, LC = F) =$$

$$= \alpha \cdot 0.6 \cdot 0.2 = 0.12$$

Avremo quindi che $P(HD|MB(HD)) = \alpha < 0.28; 0.12 > = < 0.7; 0.3 >$
Supponiamo di estrarre un numero casuale che ci porta a determinare il valore $HD = T$
Avremo quindi un nuovo stato

$$S_2 = [S = T; HD = T; LC = F; BD = T;]$$

Sucessivamente si procede ricampionando LC, poi ancora HD (partendo dai valori in S_2)e così via.
Supponiamo di avere generato 1000 sample, di cui

- 800 HD = T
- 200 HD = F

Avremo quindi che $P(HD|S = T, BD = T) =< 0.8; 0.2 >$
Inizialmente i valori della variabile campionata saranno molto divergenti, dopo di che tenderanno a uniformarsi con l'aumentare delle iterazioni.
La fase iniziale è detta di burn in, perchè i valori della variabile campionata oscillano molto e vengono eliminati.
Si puo scegliere il valore finale o una finestra delle ultime stime.