Machine Intelligence

Lecture 6: Inference in Bayesian networks

Thomas Dyhre Nielsen

Aalborg University

Tentative course overview

Topics:

- Introduction
- Search-based methods
- Constrained satisfaction problems
- Logic-based knowledge representation
- Representing domains endowed with uncertainty.
- Bayesian networks
- **Inference in Bayesian networks**
- Machine learning
- Planning
- Reinforcement learning
- Multi-agent systems

Exact Inference

Notation: We use bold upper-case characters to denote sets of random variables, and bold lower-case characters to denotes sets of corresponding values. An expression $\mathbf{E} = \mathbf{e}$ abbreviates $E_1 = e_1, \ldots, E_l = e_l$.

Posterior Marginals

Inference Problem:

- given: a Bayesian network
- given: an assignment of values to some of the variables in the network: $E_i = e_i$ $(i = 1, \ldots, l)$ ("Instantiation of the nodes $\mathbf{E}$", "evidence $\mathbf{E} = \mathbf{e}$ entered").
- want: for variables $A \notin \mathbf{E}$ the *posterior marginal* $P(A \mid \mathbf{E} = \mathbf{e})$.

According to the definition of conditional probability:

$$P(A \mid \mathbf{E} = \mathbf{e}) = \frac{P(A, \mathbf{E} = \mathbf{e})}{P(\mathbf{E} = \mathbf{e})}$$

It is sufficient to compute for each $a \in D_A$ the value

$$P(A = a, \mathbf{E} = \mathbf{e}).$$

Together with

$$P(\mathbf{E} = \mathbf{e}) = \sum_{a \in D_A} P(A = a, \mathbf{E} = \mathbf{e})$$

this gives the desired posterior distribution.

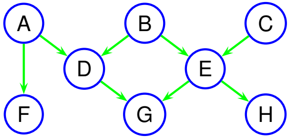
Inference as summation

Let $\mathbf{Y} = Y_1, \ldots, Y_l$ be the random variables of the network not belonging to $A \cup \mathbf{E}$. Then

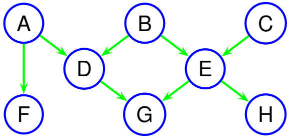
$$P(A = a, \mathbf{E} = \mathbf{e}) = \sum_{y_1 \in D_{Y_1}} \ldots \sum_{y_l \in D_{Y_l}} P(A = a, \mathbf{E} = \mathbf{e}, Y_1 = y_1, \ldots, Y_l = y_l).$$

For each $\mathbf{y}$ the probability $P(A = a, \mathbf{E} = \mathbf{e}, \mathbf{Y} = \mathbf{y})$ can be computed from the network (in time linear in the number of random variables).

Inference Problems I



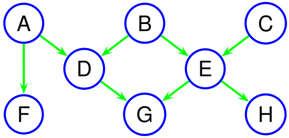
Find $P(B|a,f,g,h) = \frac{P(B,a,f,g,h)}{P(a,f,g,h)}$



Find $P(B|a, f, g, h) = \frac{P(B,a,f,g,h)}{P(a,f,g,h)}$

We can if we have access to $P(A, B, C, D, E, F, G, H)$:

$$P(A, B, C, D, E, F, G, H) = P(A)P(B)P(C)P(D|A, B) \cdot \ldots \cdot P(H|E)$$



Find $P(B|a,f,g,h) = \frac{P(B,a,f,g,h)}{P(a,f,g,h)}$

We can if we have access to $P(A,B,C,D,E,F,G,H)$:

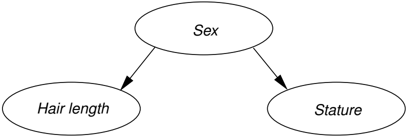
$$P(A,B,C,D,E,F,G,H) = P(A)P(B)P(C)P(D|A,B)\cdot\ldots\cdot P(H|E)$$

Inserting evidence we get:

$$P(B,a,f,g,h) = \sum_{C,D,E} P(a,B,C,D,E,f,g,h)$$

and

$$P(a,f,g,h) = \sum_{B} P(B,a,f,g,h)$$



Conditional probability tables:

Sex	
male	0.49
female	0.51

	Sex	
Hair length	*male*	*female*
long	0.05	0.6
short	0.95	0.4

	Sex	
Stature	*male*	*female*
≤ 1.68	0.08	0.47
> 1.68	0.92	0.53

Posterior probability inference: Given the value of some observed variables (the evidence) compute the conditional distribution of some other variable:

$$P(\textit{Stature} \mid \textit{Hair length} = \textit{long}) = ?$$
$$P(\textit{Sex} \mid \textit{Hair length} = \textit{short}, \textit{Stature} \leq 1.68) = ?$$

$P(\textit{Sex})$

Sex	
male	0.49
female	0.51

$P(\textit{Hair length} \mid \textit{Sex})$

	Sex	
Hair length	male	female
long	0.05	0.6
short	0.95	0.4

$P(\textit{Stature} \mid \textit{Sex})$

	Sex	
Stature	male	female
≤ 1.68	0.08	0.47
> 1.68	0.92	0.53

Query: $P(\textit{Stature} \mid \textit{Hair length} = \textit{long}) = ?$

Naive Solution Step 1

$P($*Sex*$)$

Sex	
male	0.49
female	0.51

$P($*Hair length* $\mid$ *Sex*$)$

	Sex	
Hair length	*male*	*female*
long	0.05	0.6
short	0.95	0.4

$P($*Stature* $\mid$ *Sex*$)$

	Sex	
Stature	*male*	*female*
≤ 1.68	0.08	0.47
> 1.68	0.92	0.53

Query: $P($*Stature* $\mid$ *Hair length = long*$) = ?$

Step 1: Construct joint distribution

$P($*Sex*, *Hair length*, *Stature*$)$

	Sex			
	male		*female*	
	Hair length		*Hair length*	
Stature	*long*	*short*	*long*	*short*
≤ 1.68				
> 1.68				

Naive Solution Step 1

$P(\textit{Sex})$

Sex	
male	0.49
female	0.51

$P(\textit{Hair length} \mid \textit{Sex})$

	Sex	
Hair length	*male*	*female*
long	0.05	0.6
short	0.95	0.4

$P(\textit{Stature} \mid \textit{Sex})$

	Sex	
Stature	*male*	*female*
≤ 1.68	0.08	0.47
> 1.68	0.92	0.53

Query: $P(\textit{Stature} \mid \textit{Hair length} = \textit{long}) = ?$

Step 1: Construct joint distribution

$P(\textit{Sex}, \textit{Hair length}, \textit{Stature})$

	Sex			
	male		*female*	
	Hair length		Hair length	
Stature	*long*	*short*	*long*	*short*
≤ 1.68	0.00196			
> 1.68				

$P(\textit{Sex})$

Sex	
male	0.49
female	0.51

$P(\textit{Hair length} \mid \textit{Sex})$

	Sex	
Hair length	*male*	*female*
long	0.05	0.6
short	0.95	0.4

$P(\textit{Stature} \mid \textit{Sex})$

	Sex	
Stature	*male*	*female*
≤ 1.68	0.08	0.47
> 1.68	0.92	0.53

Query: $P(\textit{Stature} \mid \textit{Hair length} = \textit{long}) = ?$

Step 1: Construct joint distribution

$P(\textit{Sex}, \textit{Hair length}, \textit{Stature})$

	Sex			
	male		*female*	
	Hair length		Hair length	
Stature	*long*	*short*	*long*	*short*
≤ 1.68	0.00196	0.03724		
> 1.68				

Naive Solution Step 1

$P(\textit{Sex})$

Sex	
male	0.49
female	0.51

$P(\textit{Hair length} \mid \textit{Sex})$

	Sex	
Hair length	male	female
long	0.05	0.6
short	0.95	0.4

$P(\textit{Stature} \mid \textit{Sex})$

	Sex	
Stature	male	female
≤ 1.68	0.08	0.47
> 1.68	0.92	0.53

Query: $P(\textit{Stature} \mid \textit{Hair length} = \textit{long}) = ?$

Step 1: Construct joint distribution

$P(\textit{Sex}, \textit{Hair length}, \textit{Stature})$

	Sex			
	male		female	
	Hair length		Hair length	
Stature	long	short	long	short
≤ 1.68	0.00196	0.03724	0.14382	0.09588
> 1.68	0.02254	0.42826	0.16218	0.10812

Naive Solution Step 2

Joint distribution $P($*Sex*, *Hair length*, *Stature*$)$

	Sex			
	male		*female*	
	Hair length		*Hair length*	
Stature	*long*	*short*	*long*	*short*
≤ 1.68	0.00196	0.03724	0.14382	0.09588
> 1.68	0.02254	0.42826	0.16218	0.10812

Step 2 "Enter evidence" :

$P($*Sex*, *Hair length*, *Stature*$)$

	Sex			
	male		*female*	
	Hair length		*Hair length*	
Stature	*long*	*short*	*long*	*short*
≤ 1.68	0.00196	~~0.03724~~	0.14382	~~0.09588~~
> 1.68	0.02254	~~0.42826~~	0.16218	~~0.10812~~

$\xrightarrow{\textit{Hair length = long}}$

$P($*Sex*, *Hair length=long*, *Stature*$)$

	Sex	
Stature	*male*	*female*
≤ 1.68	0.00196	0.14382
> 1.68	0.02254	0.16218

Note: the table on the right shows neither a joint nor a conditional distribution!

Step 3 Marginalize (sum out *Sex* variable):

$P($*Sex*, *Hair length=long*, *Stature*$)$

Stature	*Sex*: *male*	*Sex*: *female*
≤ 1.68	0.00196	0.14382
> 1.68	0.02254	0.16218

$$\xrightarrow{\sum_{Sex}}$$

$P($*Hair length=long*, *Stature*$)$

Stature	
≤ 1.68	0.14578
> 1.68	0.18472

Step 3 Marginalize (sum out *Sex* variable):

$P($*Sex*, *Hair length=long*, *Stature*$)$

Stature	*Sex*: *male*	*Sex*: *female*
≤ 1.68	0.00196	0.14382
> 1.68	0.02254	0.16218

$\xrightarrow{\sum_{Sex}}$

$P($*Hair length=long*, *Stature*$)$

Stature	
≤ 1.68	0.14578
> 1.68	0.18472

Step 4 Normalize

$P($*Hair length=long*, *Stature*$)$

Stature	
≤ 1.68	0.14578
> 1.68	0.18472

$\xrightarrow{\frac{1}{0.14578+0.18472}}$

$P($*Stature* $\mid$ *Hair length=long*$)$

Stature	
≤ 1.68	0.441
> 1.68	0.559

Naive Solution: Summary

Construct Joint: $P(\textit{Sex}, \textit{Hair length}, \textit{Stature}) =$
$P(\textit{Sex})P(\textit{Hair length} \mid \textit{Sex})P(\textit{Stature} \mid \textit{Sex})$

Naive Solution: Summary

Construct Joint: $P(\textit{Sex}, \textit{Hair length}, \textit{Stature}) =$
$P(\textit{Sex})P(\textit{Hair length} \mid \textit{Sex})P(\textit{Stature} \mid \textit{Sex})$

Insert Evidence: $P(\textit{Sex}, \textit{Hair length=long}, \textit{Stature})$

Naive Solution: Summary

Construct Joint: $P(\textit{Sex}, \textit{Hair length}, \textit{Stature}) = P(\textit{Sex})P(\textit{Hair length} \mid \textit{Sex})P(\textit{Stature} \mid \textit{Sex})$

Insert Evidence: $P(\textit{Sex}, \textit{Hair length=long}, \textit{Stature})$

Marginalize: $P(\textit{Hair length=long}, \textit{Stature})$=

$$\sum_{s \in \{\textit{male},\textit{female}\}} P(\textit{Sex=s}, \textit{Hair length=long}, \textit{Stature})$$

Naive Solution: Summary

Construct Joint: $P(\textit{Sex}, \textit{Hair length}, \textit{Stature}) = P(\textit{Sex})P(\textit{Hair length} \mid \textit{Sex})P(\textit{Stature} \mid \textit{Sex})$

Insert Evidence: $P(\textit{Sex}, \textit{Hair length=long}, \textit{Stature})$

Marginalize: $P(\textit{Hair length=long}, \textit{Stature})$=

$$\sum_{s \in \{\textit{male},\textit{female}\}} P(\textit{Sex=s}, \textit{Hair length=long}, \textit{Stature})$$

Condition: $P(\textit{Stature} \mid \textit{Hair length=long}) =$

$$\frac{P(\textit{Hair length=long}, \textit{Stature})}{P(\textit{Hair length=long}, \textit{Stature} \leq 1.68) + P(\textit{Hair length=long}, \textit{Stature} > 1.68)}$$

Construct Joint: $P(\textit{Sex}, \textit{Hair length}, \textit{Stature}) =$
$P(\textit{Sex})P(\textit{Hair length} \mid \textit{Sex})P(\textit{Stature} \mid \textit{Sex})$

Insert Evidence: $P(\textit{Sex}, \textit{Hair length=long}, \textit{Stature})$

Marginalize: $P(\textit{Hair length=long}, \textit{Stature})$=

$$\sum_{s \in \{\textit{male}, \textit{female}\}} P(\textit{Sex=s}, \textit{Hair length=long}, \textit{Stature})$$

Condition: $P(\textit{Stature} \mid \textit{Hair length=long}) =$

$$\frac{P(\textit{Hair length=long}, \textit{Stature})}{P(\textit{Hair length=long}, \textit{Stature} \leq 1.68) + P(\textit{Hair length=long}, \textit{Stature} > 1.68)}$$

Complexity

Complexity dominated by initial table $P(\textit{Sex}, \textit{Hair length}, \textit{Stature})$ (size 2^3).

For model with n binary variables:

$$O(2^n)$$

Problem

The joint probability distribution will contain exponentially many entries.

Idea

We can use

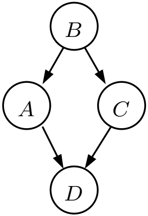
- the form of the joint distribution P
- the law of distributivity

to make the computation of the sum more efficient.

Variable Elimination

Thus, we can adapt our elimination procedure so that:

- we marginalize out variables sequentially
- when marginalizing out a particular variable X, we only need to consider the factors involving X.



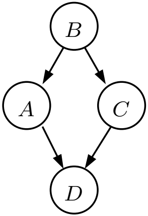
	B	
	t	*f*
	.5	.5

	A	
B	*t*	*f*
t	.7	.3
f	.1	.9

	C	
B	*t*	*f*
t	.7	.3
f	.2	.8

		D	
A	C	*t*	*f*
t	*t*	.9	.1
t	*f*	.7	.3
f	*t*	.8	.2
f	*f*	.4	.6

$$P(A, D = f) =$$



	B	
	t	f
	.5	.5

	A	
B	t	f
t	.7	.3
f	.1	.9

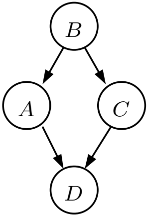
	C	
B	t	f
t	.7	.3
f	.2	.8

		D	
A	C	t	f
t	t	.9	.1
t	f	.7	.3
f	t	.8	.2
f	f	.4	.6

$$P(A, D = f) =$$

$$\sum_{b \in \{t,f\}} \sum_{c \in \{t,f\}} P(B = b, A, C = c, D = f) =$$

Example



	B: t	B: f
	.5	.5

B	A: t	A: f
t	.7	.3
f	.1	.9

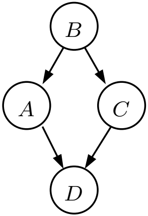
B	C: t	C: f
t	.7	.3
f	.2	.8

A	C	D: t	D: f
t	t	.9	.1
t	f	.7	.3
f	t	.8	.2
f	f	.4	.6

$$P(A, D = f) =$$

$$\sum_{b \in \{t,f\}} \sum_{c \in \{t,f\}} P(B = b, A, C = c, D = f) =$$

$$\sum_{b \in \{t,f\}} \sum_{c \in \{t,f\}} P(B = b) P(A \mid B = b) P(C = c \mid B = b) P(D = f \mid A, C = c) =$$



	B	
	t	f
	.5	.5

	A	
B	t	f
t	.7	.3
f	.1	.9

	C	
B	t	f
t	.7	.3
f	.2	.8

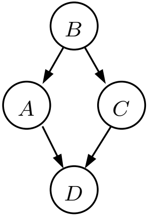
		D	
A	C	t	f
t	t	.9	.1
t	f	.7	.3
f	t	.8	.2
f	f	.4	.6

$$P(A, D = f) =$$

$$\sum_{b \in \{t,f\}} \sum_{c \in \{t,f\}} P(B = b, A, C = c, D = f) =$$

$$\sum_{b \in \{t,f\}} \sum_{c \in \{t,f\}} P(B = b)P(A \mid B = b)P(C = c \mid B = b)P(D = f \mid A, C = c) =$$

$$\sum_{b \in \{t,f\}} P(B = b)P(A \mid B = b) \sum_{c \in \{t,f\}} P(C = c \mid B = b)P(D = f \mid A, C = c)$$



	B	
	t	f
	.5	.5

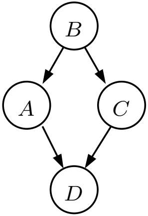
	A	
B	t	f
t	.7	.3
f	.1	.9

	C	
B	t	f
t	.7	.3
f	.2	.8

		D	
A	C	t	f
t	t	.9	.1
t	f	.7	.3
f	t	.8	.2
f	f	.4	.6

$$\sum_b P(B=b)P(A \mid B=b)\sum_c P(C=c \mid B=b)P(D=f \mid A, C=c) =$$
$$\sum_b P(B=b)P(A \mid B=b)F_1(B=b, A) = F_2(A)$$

Example continued



	B = t	B = f
	.5	.5

B	A = t	A = f
t	.7	.3
f	.1	.9

B	C = t	C = f
t	.7	.3
f	.2	.8

A	C	D = t	D = f
t	t	.9	.1
t	f	.7	.3
f	t	.8	.2
f	f	.4	.6

$$\sum_b P(B=b)P(A \mid B=b) \sum_c P(C=c \mid B=b) P(D=f \mid A, C=c) =$$

$$\sum_b P(B=b)P(A \mid B=b) F_1(B=b, A) = F_2(A)$$

where

B	C = t	C = f
t	.7	.3
f	.2	.8

A	C	D = t	D = f
t	t	.9	.1
t	f	.7	.3
f	t	.8	.2
f	f	.4	.6

$\mapsto$

b	a	$F_1(B, A)$
t	t	.7· .1 + .3· .3 = .16
t	f	.7·.2 + .3·.6 = .32
f	t	.2·.1 + .8·.3 = .26
f	f	.2·.2 + .8·.6 = .52

and

	B = t	B = f
	.5	.5

B	A = t	A = f
t	.7	.3
f	.1	.9

b	a	$F_1(B, A)$
t	t	.16
⋮	⋮	⋮

$\mapsto$

a	$F_2(A)$
t	...
f	...

Calculus of factors

- The procedure operates on **factors**: functions of subsets of variables
- Required operations on factors:
  - *multiplication*
  - *marginalization* (summing out selected variables)
  - *restriction* (setting selected variables to specific values)

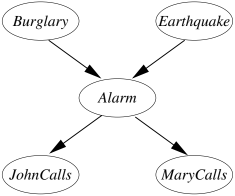
Complexity

Call subsets $\mathbf{U}$ of $\mathbf{V}$ that are the arguments of factors $P(\ldots \mid \ldots)$ resp. $F_j(\ldots)$ which appear in the elimination process *factor sets*.

The complexity of variable elimination is exponential in the size of the largest factor set.

The size of the largest factor set can depend strongly on the order in which variables are summed out!

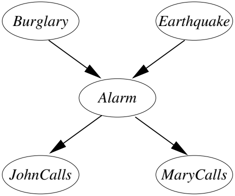
Example



Bad ordering for computing $P(MC, B = t)$:

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} \sum_{a \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$

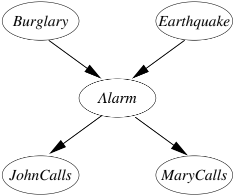
Example



Bad ordering for computing $P(MC, B = t)$:

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} \sum_{a \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$

Example

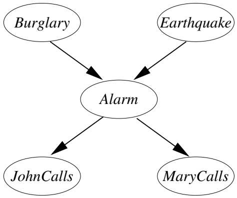


Bad ordering for computing $P(MC, B = t)$:

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} \sum_{a \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B = t) P(EQ = eq) F_1(eq, jc, MC) =$$

Example

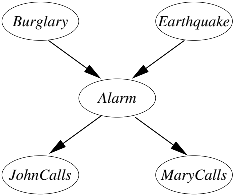


Bad ordering for computing $P(MC, B=t)$:

$$\sum_{eq\in\{t,f\}}\sum_{jc\in\{t,f\}}\sum_{a\in\{t,f\}} P(B=t)P(EQ=eq)P(A=a \mid B=t, EQ=eq)P(JC=jc \mid A=a)P(MC \mid A=a) =$$

$$\sum_{eq\in\{t,f\}}\sum_{jc\in\{t,f\}} P(B=t)P(EQ=eq)F_1(eq, jc, MC) =$$

Example



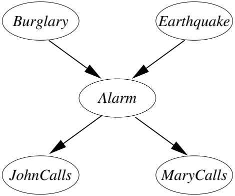
Bad ordering for computing $P(\mathit{MC}, \mathit{B} = t)$:

$$\sum_{\mathit{eq} \in \{t,f\}} \sum_{\mathit{jc} \in \{t,f\}} \sum_{\mathit{a} \in \{t,f\}} P(\mathit{B} = t)P(\mathit{EQ} = \mathit{eq})P(\mathit{A} = \mathit{a} \mid \mathit{B} = t, \mathit{EQ} = \mathit{eq})P(\mathit{JC} = \mathit{jc} \mid \mathit{A} = \mathit{a})P(\mathit{MC} \mid \mathit{A} = \mathit{a}) =$$

$$\sum_{\mathit{eq} \in \{t,f\}} \sum_{\mathit{jc} \in \{t,f\}} P(\mathit{B} = t)P(\mathit{EQ} = \mathit{eq})F_1(\mathit{eq}, \mathit{jc}, \mathit{MC}) =$$

$$\sum_{\mathit{eq} \in \{t,f\}} P(\mathit{B} = t)P(\mathit{EQ} = \mathit{eq})F_2(\mathit{eq}, \mathit{MC}) =$$

Example



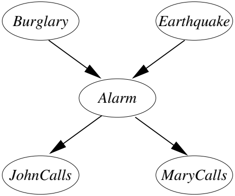
Bad ordering for computing $P(MC, B=t)$:

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} \sum_{a \in \{t,f\}} P(B=t)P(EQ=eq)P(A=a \mid B=t, EQ=eq)P(JC=jc \mid A=a)P(MC \mid A=a) =$$

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B=t)P(EQ=eq)F_1(eq, jc, MC) =$$

$$\sum_{eq \in \{t,f\}} P(B=t)P(EQ=eq)F_2(eq, MC) =$$

Example



Bad ordering for computing $P(MC, B = t)$:

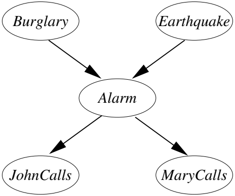
$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} \sum_{a \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B = t) P(EQ = eq) F_1(eq, jc, MC) =$$

$$\sum_{eq \in \{t,f\}} P(B = t) P(EQ = eq) F_2(eq, MC) =$$

$$P(B = t) F_3(MC)$$

Example



Bad ordering for computing $P(MC, B = t)$:

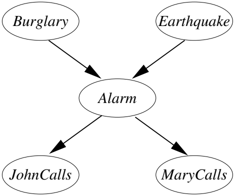
$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} \sum_{a \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$

$$\sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B = t) P(EQ = eq) F_1(eq, jc, MC) =$$

$$\sum_{eq \in \{t,f\}} P(B = t) P(EQ = eq) F_2(eq, MC) =$$

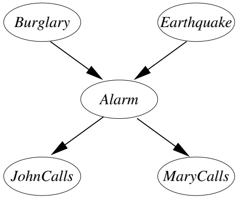
$$P(B = t) F_3(MC)$$

Largest factor (F_1) is function of 3 variables!



Good ordering for computing $P(MC, B = t)$:

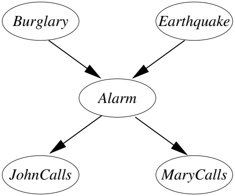
$$\sum_{a \in \{t,f\}} \sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$



Good ordering for computing $P(MC, B = t)$:

$$\sum_{a\in\{t,f\}} \sum_{eq\in\{t,f\}} \sum_{jc\in\{t,f\}} P(B = t)P(EQ = eq)P(A = a \mid B = t, EQ = eq)P(JC = jc \mid A = a)P(MC \mid A = a) =$$

$$\sum_{a\in\{t,f\}} \sum_{eq\in\{t,f\}} P(B = t)P(EQ = eq)P(A = a \mid B = t, EQ = eq)P(MC \mid A = a)F_1(a) =$$



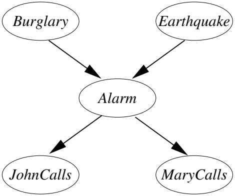
Good ordering for computing $P(MC, B = t)$:

$$\sum_{a \in \{t,f\}} \sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(JC = jc \mid A = a) P(MC \mid A = a) =$$

$$\sum_{a \in \{t,f\}} \sum_{eq \in \{t,f\}} P(B = t) P(EQ = eq) P(A = a \mid B = t, EQ = eq) P(MC \mid A = a) F_1(a) =$$

$$\sum_{a \in \{t,f\}} P(B = t) P(MC \mid A = a) F_1(a) F_2(a) =$$

Alarm Example continued



Good ordering for computing $P(MC, B = t)$:

$$\sum_{a \in \{t,f\}} \sum_{eq \in \{t,f\}} \sum_{jc \in \{t,f\}} P(B = t)P(EQ = eq)P(A = a \mid B = t, EQ = eq)P(JC = jc \mid A = a)P(MC \mid A = a) =$$

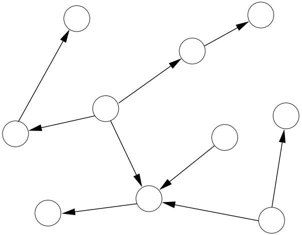
$$\sum_{a \in \{t,f\}} \sum_{eq \in \{t,f\}} P(B = t)P(EQ = eq)P(A = a \mid B = t, EQ = eq)P(MC \mid A = a)F_1(a) =$$

$$\sum_{a \in \{t,f\}} P(B = t)P(MC \mid A = a)F_1(a)F_2(a) =$$

$$P(B = t)F_3(MC)$$

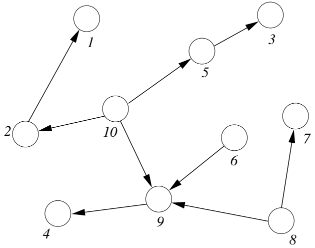
Largest factor ($P(A \mid B = t, EQ)$) is function of 2 variables!

A **singly connected network** is a network in which any two nodes are connected by at most one path of undirected edges:



Singly connected networks

A **singly connected network** is a network in which any two nodes are connected by at most one path of undirected edges:



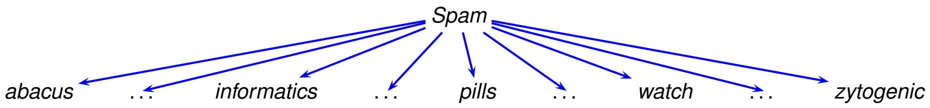
For singly connected network: any elimination order that “peels” variables from outside will only create factors of only one variable.

The complexity of inference is therefore linear in the total size of the network (= combined size of all conditional probability tables).

Example: Spam filter

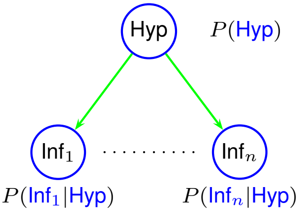
- A single *query variable*: *Spam*
- Many observable features (e.g. words appearing in the body of the message): *abacus*,...,*informatics*, *pills*, ..., *watch*,..., *zytogenic*

Network Structure:



- Inference with large number of variables possible
- Essentially how *Thunderbird* spam filter works

Naïve Bayes models

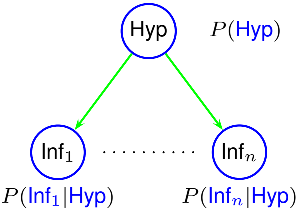


We want the posterior probability of the hypothesis variable Hyp given the observations $\{\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n\}$:

$$P(\text{Hyp}|\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n) = \frac{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n|\text{Hyp})P(\text{Hyp})}{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n)}$$

Note: The model assumes that the information variables are independent given the hypothesis variable.

Naïve Bayes models



We want the posterior probability of the hypothesis variable Hyp given the observations $\{\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n\}$:

$$P(\text{Hyp}|\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n) = \frac{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n|\text{Hyp})P(\text{Hyp})}{P(\text{Inf}_1 = e_1, \ldots, \text{Inf}_n = e_n)}$$

$$= \mu \cdot P(\text{Inf}_1 = e_1|\text{Hyp}) \cdot \ldots \cdot P(\text{Inf}_n = e_n|\text{Hyp})P(\text{Hyp})$$

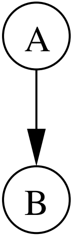
Note: The model assumes that the information variables are independent given the hypothesis variable.

Approximate Inference

Sample Generator

Observation: can use Bayesian network as random generator that produces states $\mathbf{X} = \mathbf{x}$ according to distribution P defined by the network.

Example:



	A: t	A: f
	.2	.8

A	B: t	B: f
t	.7	.3
f	.4	.6

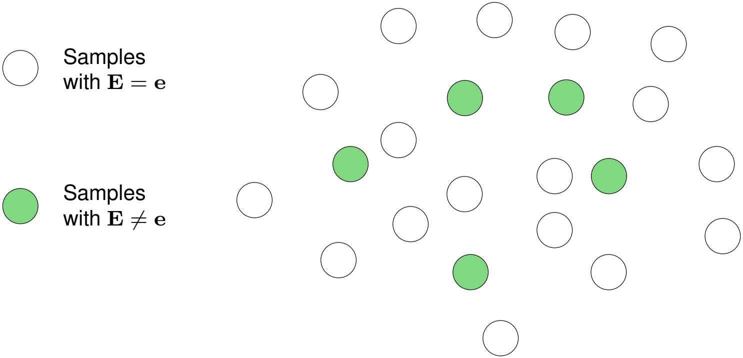
- Generate random numbers r_1, r_2 uniformly from [0,1].
- Set $A = t$ if $r_1 \leq .2$ and $A = f$ else.
- Depending on the value of A and r_2 set B to t or f.

Random generation of one state: linear in size of network.

Approximate Inference from Samples

To compute an approximation of $P(\mathbf{E} = \mathbf{e})$ ($\mathbf{E}$ a subset of the variables in the Bayesian network):

- generate a (large) number of random states
- count the frequency of states in which $\mathbf{E} = \mathbf{e}$.



Hoeffding Bound

- p: true probability $P(\mathbf{E} = \mathbf{e})$
- s: estimate for p from sample of size n
- ϵ: an error bound > 0.

Then

$$P(|s - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Hoeffding Bound

- p: true probability $P(\mathbf{E} = \mathbf{e})$
- s: estimate for p from sample of size n
- ϵ: an error bound > 0.

Then

$$P(|s - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Required Sample Size

To obtain an estimate that *with probability at most* δ has an *accuracy at least* ϵ, it is sufficient to take

$$n = -\mathit{ln}(\delta/2)/(2\epsilon^2) \text{ samples.}$$

Hoeffding Bound

- p: true probability $P(\mathbf{E} = \mathbf{e})$
- s: estimate for p from sample of size n
- ϵ: an error bound > 0.

Then

$$P(|s - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Required Sample Size

To obtain an estimate that *with probability at most* δ has an *accuracy at least* ϵ, it is sufficient to take

$$n = -\mathit{ln}(\delta/2)/(2\epsilon^2) \text{ samples.}$$

Example

To get an error ϵ of less than 0.1 in 95% of the cases ($\delta = 0.05$), we need:

$$n > -\mathit{ln}(0.05/2)/(2 \cdot 0.1^2) \approx 184 \text{ samples}$$

Hoeffding Bound

- p: true probability $P(\mathbf{E} = \mathbf{e})$
- s: estimate for p from sample of size n
- ϵ: an error bound > 0.

Then

$$P(|s - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Required Sample Size

To obtain an estimate that *with probability at most* δ has an *accuracy at least* ϵ, it is sufficient to take

$$n = -\mathit{ln}(\delta/2)/(2\epsilon^2) \text{ samples.}$$

Example

To get an error ϵ of less than 0.1 in 95% of the cases ($\delta = 0.05$), we need:

$$n > -\mathit{ln}(0.05/2)/(2 \cdot 0.1^2) \approx 184 \text{ samples}$$

How many samples do we need if the error should be less than 0.01?

Hoeffding Bound

- p: true probability $P(\mathbf{E} = \mathbf{e})$
- s: estimate for p from sample of size n
- ϵ: an error bound > 0.

Then

$$P(|s - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Required Sample Size

To obtain an estimate that *with probability at most* δ has an *accuracy at least* ϵ, it is sufficient to take

$$n = -\mathit{ln}(\delta/2)/(2\epsilon^2) \text{ samples.}$$

Example

To get an error ϵ of less than 0.1 in 95% of the cases ($\delta = 0.05$), we need:

$$n > -\mathit{ln}(0.05/2)/(2 \cdot 0.1^2) \approx 184 \text{ samples}$$

How many samples do we need if the error should be less than 0.01? 18444 samples

Conditional Probabilities

In practice, want to compute conditional probabilities $P(A = a \mid \mathbf{E} = \mathbf{e})$.

Remember exact inference:

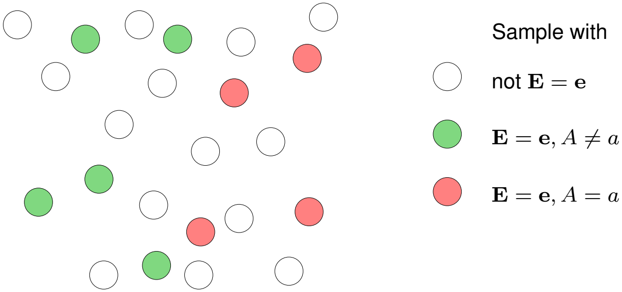
- compute $P(A = a, \mathbf{E} = \mathbf{e})$
- compute $P(\mathbf{E} = \mathbf{e})$
- divide

Approximate attempt:

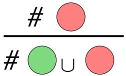
- approximate $P(A = a, \mathbf{E} = \mathbf{e})$ (use Hoeffding)
- approximate $P(\mathbf{E} = \mathbf{e})$ (use Hoeffding)
- divide

Problem: no error bound on the quotient of the two approximations as an approximation for the true quotient $P(A = a, \mathbf{E} = \mathbf{e})/P(\mathbf{E} = \mathbf{e})$!

The simplest approach: **Rejection Sampling**

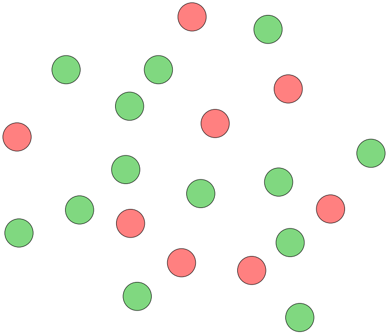


Approximation for $P(A = a \mid \mathbf{E} = \mathbf{e})$:



Problem with rejection sampling: samples with $\mathbf{E} \neq \mathbf{e}$ are useless!

Ideally: would draw samples directly from the conditional distribution $P(\mathbf{A} \mid \mathbf{E} = \mathbf{e})$.



Likelihood weighting I

First idea (*not* to be followed)

- Fix evidence variables to their observed states.
- Sample from non-evidence variables.

First idea (*not* to be followed)

- Fix evidence variables to their observed states.
- Sample from non-evidence variables.

Problem: This gives a sampling distribution

$$\prod_{X \in \mathbf{X} \setminus \mathbf{E}} P(X \mid \mathrm{pa}(X) \setminus \mathbf{E}, \mathrm{pa}(X) \cap \mathbf{E})$$

somewhere between $P(\mathbf{X})$ and $P(\mathbf{X} \mid \mathbf{e})$.

First idea (*not* to be followed)

- Fix evidence variables to their observed states.
- Sample from non-evidence variables.

Problem: This gives a sampling distribution

$$\prod_{X \in \mathbf{X} \setminus \mathbf{E}} P(X \mid \mathrm{pa}(X) \setminus \mathbf{E}, \mathrm{pa}(X) \cap \mathbf{E})$$

somewhere between $P(\mathbf{X})$ and $P(\mathbf{X} \mid \mathbf{e})$.

Likelihood weighting

We would like to sample from

$$P(\mathbf{X}, \mathbf{e}) = \underbrace{\prod_{X \in \mathbf{X} \setminus \mathbf{E}} P(X \mid \mathrm{pa}(X) \setminus \mathbf{E}, \mathrm{pa}(X) \cap \mathbf{E})}_{\text{Part 1}} \cdot \underbrace{\prod_{E \in \mathbf{E}} P(E = e \mid \mathrm{pa}(E) \setminus \mathbf{E}, \mathrm{pa}(E) \cap \mathbf{E})}_{\text{Part 2}}$$

So instead weigh each generated sample with a weight corresponding to Part 2.

Likelihood weighting

Estimate $P(X = e \mid \mathbf{e})$ as

$$\hat{P}(X = e \mid \mathbf{e}) = \frac{\sum_{\textit{sample}: X = x} w(\textit{sample})}{\sum_{\textit{sample}} w(\textit{sample})},$$

where

$$w(\textit{sample}) = \prod_{E \in \mathbf{E}} P(E = e \mid \mathrm{pa}(E) = \pi) \qquad \text{(Part 2 on the previous slide)}$$

and π is the values of $\mathrm{pa}(E)$ under *sample* and $\mathbf{e}$.

Likelihood weighting

Estimate $P(X = e \mid \mathbf{e})$ as

$$\hat{P}(X = e \mid \mathbf{e}) = \frac{\sum_{\textit{sample}: X=x} w(\textit{sample})}{\sum_{\textit{sample}} w(\textit{sample})},$$

where

$$w(\textit{sample}) = \prod_{E \in \mathbf{E}} P(E = e \mid \mathrm{pa}(E) = \pi) \qquad \text{(Part 2 on the previous slide)}$$

and π is the values of $\mathrm{pa}(E)$ under *sample* and $\mathbf{e}$.

Importance sampling

Likelihood weighting is an instance of importance sampling, where

- samples are weighted and can come from (almost) any proposal distribution.