CS761 Artificial Intelligence

17. Inference with Bayesian Networks

Jiamou Liu
The University of Auckland

Recall: Probability and Uncertainty

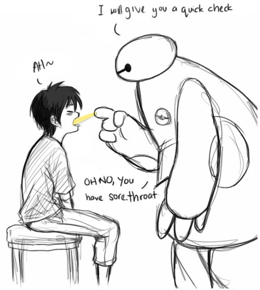
- We introduced probability theory as a language that extends from propositional logic to capture uncertainty.
- The probability of a proposition expresses the degree of belief (belief measure over a sample space) held about the proposition being true.
- When describing a world with uncertainty, we use a set of atomic propositions with a joint probability distribution.
  - Causal factor relation
  - Independence
- Probabilistic inference:

  Posterior probability $\propto$ Likelihood $\times$ Prior Probability

Diagnosing a Complex Problem

Example. [coughing?]

- When a person gets influenza (INF), he may develop a sore throat (SOR), a fever (FEV), or maybe bronchitis (BRO).
- When a person is a smoker (SMO), he may also get bronchitis.
- When a person gets bronchitis, he may develop coughing (COU) and wheezing (WEE).



Prior knowledge:

- *INF* and *SMO* are independent.
- *SOR*, *FEV*, and *BRO* are directly affected by *INF*, and are conditionally independent given *INF*. (*INF* ⇢ *SOR*,*INF* ⇢ *FEV*,*INF* ⇢ *BRO*)
- *BRO* is also affected by *SMO*. (*SMO* ⇢ *BRO*)
- *COU* and *WEE* are directly affected by *BRO*, and are conditionally independent given *BRO*. (*BRO* ⇢ *COU*, *BRO* ⇢ *WEE*)

For each atom, there is a conditional probability table:

INF	$\mathbf{P}_{INF}$
0	0.95
1	0.05

SMO	$\mathbf{P}_{SMO}$
0	0.8
1	0.2

INF	SOR	$\mathbf{P}_{SOR}$
0	0	0.9999
0	1	0.0001
1	0	0.7
1	1	0.3

INF	FEV	$\mathbf{P}_{FEV}$
0	0	0.95
0	1	0.05
1	0	0.1
1	1	0.9

INF	SMO	BRO	$\mathbf{P}_{BRO}$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.3
0	1	1	0.7
1	0	0	0.1
1	0	1	0.9
1	1	0	0.01
1	1	1	0.99

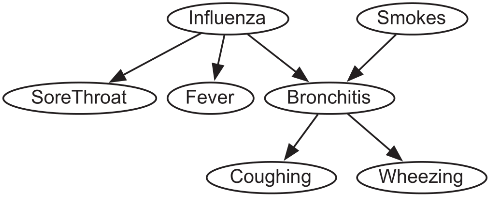
BRO	COU	$\mathbf{P}_{COU}$
0	0	0.93
0	1	0.07
1	0	0.2
1	1	0.8

BRO	WEE	$\mathbf{P}_{WEE}$
0	0	0.999
0	1	0.05
1	0	0.4
1	1	0.6

Question. Suppose the person is coughing. What is the probability that he has influenza?

Probabilistic Model

The problem can be represented by a directed graph of the problem domain:



Note.

- The graph is acyclic (i.e. does not contain any directed cycles).
- A node X connects to another node Y by a directed edge if the X is a direct causal factor of Y.

Definition [graph terminology]

Let $G = (V, E)$ be a directed graph.

- A root of G is any node with no parent;
- parents(X) $= \{Y \in V \mid (Y, X) \in E\}$ is the set of parents of X.
- a leaf is any node that is not a parent of any others; the other nodes are called the intermediate nodes;
- A node X is a descendent of Y if there is a directed path from Y to X.

Bayesian Networks

Definition [graph terminology]

Let $G = (V, E)$ be a directed graph.

- A root of G is any node with no parent;
- $\text{parents}(X) = \{Y \in V \mid (Y, X) \in E\}$ is the set of parents of X.
- a leaf is any node that is not a parent of any others; the other nodes are called the intermediate nodes;
- A node X is a descendent of Y if there is a directed path from Y to X.

Definition. [Bayesian network]

A Bayesian network is a tuple $(V, E, (\mathbf{P}_X)_{X \in V})$ where

- (V, E) forms a directed acyclic graph; each node $X \in V$ represents an atomic proposition.
- For every node X, $\mathbf{P}_X$ is a probability distribution of X conditioning on parents(X), i.e., $\mathbf{P}(X \mid \text{parents}(X))$.
- Local Markov property: Each node is independent from its non-descendants conditioned on its parents.

Example. [coughing?] A Bayesian network: $(V, E, (\mathbf{P}_X)_{X \in V})$ where

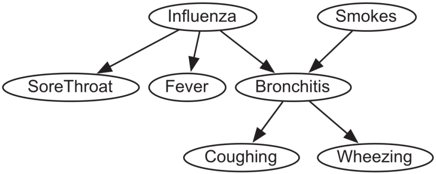
- V contains atoms $INF, SMO, SOR, FEV, BRO, COU, WEE$
- (Syntax) E represents all the dependencies between atoms:

$$INF \dashrightarrow SOR, INF \dashrightarrow FEV, INF \dashrightarrow BRO$$

$$SMO \dashrightarrow BRO, BRO \dashrightarrow COU, BRO \dashrightarrow WEE$$

- (Semantics) The CPT given earlier.

Note. Local Markov property ensures all observed conditional independence.



For any atom X, we use ℓ_X to denote a literal, i.e. X or $\neg X$.

Bayesian Network Inference Problem

INPUT Given a Bayesian network $(V, E, (\mathbf{P}_X)_{X \in V})$, two nodes $X, Y \in V$, and literal $\ell_Y \in \{Y, \neg Y\}$,

OUTPUT compute the probability $P(X \mid \ell_Y)$ and $P(\neg X \mid \ell_Y)$.

E.g. Given the earlier example Bayesian network, query for $P(INF \mid COU)$.

We now present methods to solve this problem.

Tools:

1. CPT operations
2. Chain rule

CPT Operations

CPT Operation 1: Restriction

For a conditional probability table $f(Y_1, \ldots, Y_k, X)$, for $\ell_X \in \{X, \neg X\}$, the ℓ_X-restriction operation produces a table $g(Y_1, \ldots, Y_k)$ such that for any $\ell_{Y_i} \in \{Y_i, \neg Y_i\}$,

$$g(\ell_{Y_1}, \ldots, \ell_{Y_k}) = f(\ell_{Y_1}, \ldots, \ell_{Y_k}, \ell_X).$$

E.g.

Y	X	$f(Y, X)$
0	0	0.99
0	1	0.01
1	0	0.05
1	1	0.95

$\Rightarrow$

Y	$g(Y)$
0	0.01
1	0.95

CPT Operation 2: Multiplication

For CPT $f(Y_1, \ldots, Y_k, X)$ and $g(X, Z_1, \ldots, Z_m)$, the multiplication operation produces a new CPT $h(Y_1, \ldots, Y_k, X, Z_1, \ldots, Z_m)$ such that for any literals $\ell_{Y_i} \in \{Y_i, \neg Y_i\}$, $\ell_X \in \{X, \neg X\}$, $\ell_{Z_j} \in \{Z_j, \neg Z_j\}$,

$$h(\ell_{Y_1}, \ldots, \ell_{Y_k}, \ell_X, \ell_{Z_1}, \ldots, \ell_{Z_m}) = f(\ell_{Y_1}, \ldots, \ell_{Y_k}, \ell_X) \times g(\ell_X, \ell_{Z_1}, \ldots, \ell_{Z_m}).$$

E.g.

A	B	f
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

$\times$

B	C	g
0	0	0.2
0	1	0.8
1	0	0.3
1	1	0.7

$=$

A	B	C	h
0	0	0	0.02
0	0	1	0.72
0	1	0	0.27
0	1	1	0.63
1	0	0	0.08
1	0	1	0.32
1	1	0	0.18
1	1	1	0.42

CPT Operation 3: Normalisation

For a conditional probability table $f(Y_1, \ldots, Y_k, X)$, the X-normalisation operation produces a new CPT $g(Y_1, \ldots, Y_k, X)$ such that for any $\ell_{Y_i} \in \{Y_i, \neg Y_i\}$, $\ell_X \in \{X, \neg X\}$,

$$g(\ell_{Y_1}, \ldots, \ell_{Y_k}, \ell_X) = \frac{f(\ell_{Y_1}, \ldots, \ell_{Y_k}, \ell_X)}{f(\ell_{Y_1}, \ldots, \ell_{Y_k}, X) + f(\ell_{Y_1}, \ldots, \ell_{Y_k}, \neg X)}.$$

E.g.

X	$f(X)$
0	0.0096
1	0.038

$\Rightarrow$

X	$g(X)$
0	0.2017
1	0.7983

CPT Operation 4: Summation

For a conditional probability table $f(Y_1, \ldots, Y_k, X)$, the X-sum operation produces a new CPT $g(Y_1, \ldots, Y_k)$ such that for any $\ell_{Y_i} \in \{Y_i, \neg Y_i\}$,

$$g(\ell_{Y_1}, \ldots, \ell_{Y_k}) = f(\ell_{Y_1}, \ldots, \ell_{Y_k}, X) + f(\ell_{Y_1}, \ldots, \ell_{Y_k}, \neg X).$$

E.g. C-sum of the following table:

A	B	C	f
0	0	0	0.02
0	0	1	0.72
0	1	0	0.27
0	1	1	0.63
1	0	0	0.08
1	0	1	0.32
1	1	0	0.18
1	1	1	0.42

$\Rightarrow$

A	B	g
0	0	0.74
0	1	0.9
1	0	0.4
1	1	0.6

Example. [simple diagnosis] To query $\mathbf{P}(Inf|Test)$:

Inf	$\mathbf{P}_{Inf}$
0	0.96
1	0.04

and

Inf	Test	$\mathbf{P}_{Test}(Inf)$
0	0	0.99
0	1	0.01
1	0	0.05
1	1	0.95

Note. $\mathbf{P}(Inf|Test) \propto \mathbf{P}(Inf) \times \mathbf{P}(Test \mid Inf)$.

Example. [simple diagnosis] To query $\mathbf{P}(Inf|Test)$:

Inf	$\mathbf{P}_{Inf}$
0	0.96
1	0.04

and

Inf	Test	$\mathbf{P}_{Test}(Inf)$
0	0	0.99
0	1	0.01
1	0	0.05
1	1	0.95

Note. $\mathbf{P}(Inf|Test) \propto \mathbf{P}(Inf) \times \mathbf{P}(Test \mid Inf)$.

Step 1 Restriction. In $\mathbf{P}_{Test}$ keep only those rows where *Test* is true.

Inf	$f(Inf)$
0	0.01
1	0.95

Step 2 Multiplication.

Inf	$\mathbf{P}_{Inf}$
0	0.96
1	0.04

$\times$

Inf	$f(Inf)$
0	0.01
1	0.95

$=$

Inf	$g(Inf)$
0	$0.01 \times 0.96 = 0.0096$
1	$0.95 \times 0.04 = 0.038$

Step 3. Normalisation.

Inf	$g(Inf)$
0	$0.0096/(0.0096 + 0.038) = 0.0096/0.0476 \approx 0.2017$
1	$0.038/(0.0096 + 0.038) = 0.038/0.0476 \approx 0.7983$

Therefore $P(Inf \mid Test) = 0.7983$ and $P(\neg Inf \mid Test) = 0.2017$.

Chain Rule

- For any two propositions a_1, a_2

$$P(a_1 \wedge a_2) = P(a_2 \mid a_1)P(a_1).$$

- For any three propositions a_1, a_2, a_3

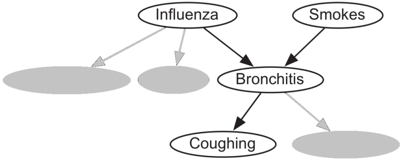
$$P(a_1 \wedge a_2 \wedge a_3) = P(a_1 \wedge a_2 \mid a_3)P(a_3) = P(a_1 \mid a_2 \wedge a_3)P(a_2 \mid a_3)P(a_3).$$

Chain Rule

For any n propositions $a_1, a_2, \ldots, a_n$

$$\begin{aligned} P\left(\bigwedge_{i=1}^{n} a_i\right) &= P(a_1 \mid a_2 \wedge \cdots \wedge a_n)P(a_2 \mid a_3 \wedge \cdots \wedge a_n) \ldots P(a_{n-1} \mid a_n)P(a_n) \\ &= \prod_{i=1}^{n} P\left(a_i \mid \bigwedge_{j=i+1}^{n} a_j\right) \end{aligned}$$

Example. [coughing?]



$$
\begin{aligned}
&\mathbf{P}(INF, SMO, BRO, COU)\\
=&\mathbf{P}(COU \mid BRO, SMO, INF)\mathbf{P}(BRO \mid SMO, INF)\mathbf{P}(SMO \mid INF)\mathbf{P}(INF)\\
=&\mathbf{P}(COU \mid BRO)\mathbf{P}(BRO \mid SMO, INF)\mathbf{P}(SMO)\mathbf{P}(INF) \quad \text{(by local Markov property)}
\end{aligned}
$$

Theorem [chain rule on a Bayesian network]

Let $(V, E, (\mathbf{P}_X)_{X\in V})$ be a Bayesian network. Then for any subset $S \subseteq V$

$$\mathbf{P}(S) = \prod_{X\in S} \mathbf{P}(X \mid \text{parents}(X)) = \prod_{X\in S} \mathbf{P}_X$$

Example. [coughing?] What is $\mathbf{P}(INF \mid COU)$?

Idea:

$$\begin{aligned}\mathbf{P}(INF \mid COU) = \mathbf{P}(COU \mid INF)\mathbf{P}(INF) &= \mathbf{P}(COU, INF)\\ &= \sum_{BRO}\sum_{SMO} \mathbf{P}(INF, COU, BRO, SMO)\\ &= \sum_{BRO}\sum_{SMO} \mathbf{P}_{COU}\mathbf{P}_{BRO}\mathbf{P}_{SMO}\mathbf{P}_{INF}\end{aligned}$$

Starting from the set of CPT: $\mathbf{P}_{COU}$, $\mathbf{P}_{BRO}$, $\mathbf{P}_{SMO}$, $\mathbf{P}_{INF}$

INF	SMO	BRO	$\mathbf{P}_{BRO}$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.3
0	1	1	0.7
1	0	0	0.1
1	0	1	0.9
1	1	0	0.01
1	1	1	0.99

INF	$\mathbf{P}_{INF}$
0	0.95
1	0.05

SMO	$\mathbf{P}_{SMO}$
0	0.8
1	0.2

BRO	COU	$\mathbf{P}_{COU}$
0	0	0.93
0	1	0.07
1	0	0.2
1	1	0.8

Question. How to compute $\sum_{BRO} \sum_{SMO} \mathbf{P}_{COU}\mathbf{P}_{BRO}\mathbf{P}_{SMO}\mathbf{P}_{INF}$?

Naive method.

1. **Step 1: Restriction.** In $\mathbf{P}_{COU}$, keep only those rows where COU is true.

   Obtain a new CPT $f(BRO)$.

2. **Step 2: Multiplication.** Compute $f(BRO) \times \mathbf{P}_{BRO} \times \mathbf{P}_{SMO} \times \mathbf{P}_{INF}$

   Obtain a new CPT $g(BRO, SMO, INF)$

3. **Step 3: Sum (to eliminate variables).** Perform SMO-sum and then BRO-sum, on $g(BRO, SMO, INF)$.

   Obtain a new CPT $h(INF)$

4. **Step 3: Normalisation.** Extract the conditional probabilities $P(INF \mid COU)$ and $P(\neg INF \mid COU)$.

Complexity. Suppose a CPT has r rows. The algorithm runs in $O(r^n)$.

We now present how to improve the running time of the algorithm.

Variable Elimination

Idea: Instead of multiply all tables and eliminate variables all at once, eliminate variables one-by-one.

1. **Step 1: Restriction.** In $\mathbf{P}_{COU}$, keep only those rows where COU is true. Obtain a new CPT $f(BRO)$.
2. **Step 2: Eliminate SMO.**
    - **Step 2.1:** Multiply $\mathbf{P}_{BRO} \times \mathbf{P}_{SMO}$ to get a new table $g(INF, SMO, BRO)$
    - **Step 2.2:** SMO-sum $g(INF, SMO, BRO)$ to get a new table $h(INF, BRO)$.
3. **Step 3: Eliminate BRO.**
    - **Step 3.1:** Multiply $f(BRO) \times h(INF, BRO)$ to get a new table $\ell(INF, BRO)$
    - **Step 3.2:** BRO-sum $\ell(INF, BRO)$ to get a new table $r(INF)$.
4. **Step 4: Multiplication.** $r(INF) \times \mathbf{P}(INF)$ to a table $s(INF)$.
5. **Step 5: Normalisation.** Retrieve $\mathbf{P}(INF \mid COU)$.

Illustration.

Step 1: Restriction. In $\mathbf{P}_{COU}$, keep only those rows where COU is true.

BRO	COU	$\mathbf{P}_{COU}$
0	0	0.93
0	1	0.07
1	0	0.2
1	1	0.8

$\Rightarrow$

BRO	$f(BRO)$
0	0.07
1	0.8

Step 2: Eliminate SMO.

2.1

INF	SMO	BRO	$\mathbf{P}_{BRO}$
0	0	0	0.9999
0	0	1	0.0001
0	1	0	0.3
0	1	1	0.7
1	0	0	0.1
1	0	1	0.9
1	1	0	0.01
1	1	1	0.99

$\times$

SMO	$\mathbf{P}_{SMO}$
0	0.8
1	0.2

=

INF	SMO	BRO	$g(INF, SMO, BRO)$
0	0	0	0.79992
0	0	1	0.00008
0	1	0	0.06
0	1	1	0.14
1	0	0	0.08
1	0	1	0.72
1	1	0	0.002
1	1	1	0.198

2.2 SMO-Sum

INF	SMO	BRO	$g(INF, SMO, BRO)$
0	0	0	0.79992
0	0	1	0.00008
0	1	0	0.06
0	1	1	0.14
1	0	0	0.08
1	0	1	0.72
1	1	0	0.002
1	1	1	0.198

$\Rightarrow$

INF	BRO	$h(INF, BRO)$
0	0	0.85992
0	1	0.14008
1	0	0.082
1	1	0.918

Step 3: Eliminate *BRO*.

3.1

BRO	$\mathbf{P}_{COU}$
0	0.07
1	0.8

$\times$

INF	BRO	$h(INF, BRO)$
0	0	0.85992
0	1	0.14008
1	0	0.082
1	1	0.918

$=$

INF	BRO	$\ell(INF, BRO)$
0	0	0.0601944
0	1	0.112064
1	0	0.00574
1	1	0.7344

3.2 *BRO*-sum

INF	BRO	$\ell(INF, BRO)$
0	0	0.0601944
0	1	0.112064
1	0	0.00574
1	1	0.7344

$\Rightarrow$

INF	$r(INF)$
0	0.1722584
1	0.74014

Step 4: Multiplication.

INF	$r(INF)$
0	0.1722584
1	0.74014

$\times$

INF	$\mathbf{P}_{INF}$
0	0.95
1	0.05

$=$

	$s(INF)$
0	0.16364548
1	0.037007

Step 5: Normalisation.

	s(INF)
0	0.16364548
1	0.037007

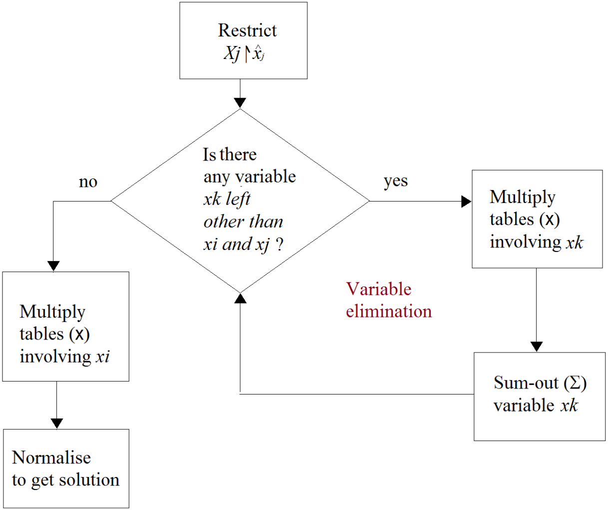
$\Rightarrow$

	result
$P(\neg INF \land COU)$	0.8156
$P(INF \land COU)$	0.1844

The probability of a person getting influenza when (only) coughing is observed is 18.44%!

Variable Elimination Algorithm

Solving the Bayesian network inference problem using the variable elimination algorithm (VE)



Variable Elimination (Rough steps)

INPUT:

- CPT $\mathbf{P}_X$ for each $X \in V$.
- Query variables X_i, X_j; Observed value for X_j.

OUTPUT: $\mathbf{P}(X_i \mid X_j$

1. Restrict tables that contain X_j to rows where X_j has the observed value.
2. Eliminate each of the non-query variables X_k (in certain order):
   - Multiply all tables containing X_k.
   - Sum-out X_k from the product table.
3. Multiply the remaining tables to get a table over a single variable X_i.
4. Normalise and return the result.

Complexity: If every atom appears in at most 2 CPT and one of them has at most 4 rows, then the algorithm runs in $O(n)$.
In the general case, the running time depends on the ordering in which we eliminate the variables, but finding the optimal ordering is NP-complete.

Summary of The Topic

The following are the main knowledge points covered:

- **Bayesian network:** $(V, E, (\mathbf{P}_X)_{X \in V})$ a graph structure denoting causal relations between atoms, CPTs
- **Local markov property**
- **Bayesian network inference:** Given a Bayesian network $(V, E, (\mathbf{P}_X)_{X \in V})$, nodes X, Y and observation $\ell_Y \in \{Y, \neg Y\}$, compute $\mathbf{P}(X \mid \ell_Y)$.
- **CPT Operations:**
  - Restriction
  - Multiplication
  - Normalisation
  - Summation
- **Chain rule:** $\pi(S) = \prod_{X \in S} \pi_X$
- **Naive method** for Bayesian network inference
- **Variable elimination**