Lecture 5: Bayesian Network

Topics of this lecture

- What is a Bayesian network?
- A simple example
- Formal definition of BN
- A slightly difficult example
- Learning of BN
- An example of learning
- Important topics in using BN

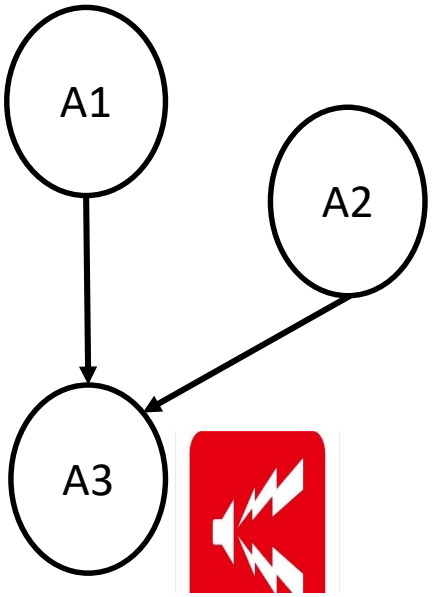
What is a Bayesian Network?

- A Bayesian network (BN) is a *graphical model* that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).
- In the literature, BN is also called Bayes network, belief network, etc. It is a special case of *causal network*.
- Compared with the methods we have studied so far, BN is more *visible and interpretable* in the sense that
  - Given some result, we can understand the most important factor(s), or
  - Given some factors or evidences, we can understand the most possible result(s).

A Simple Example (1)

Reference: Mathematics in statistic inference of Bayesian network," K. Tanaka, 2009

- Shown in the right is a simple BN.
- There are two factors here to activate the alarm.
  - Each of them has a different probability, P(A1) or P(A2), to occur, and
  - Each of them can activate the alarm with a conditional probability, P(A3|A1) or P(A3|A2).
- Based on the probabilities, we can understand, say, the probability that A1 occurred when the alarm is activated.
- In a BN, each event is denoted by a node, and the causal relations between the events are denoted by the edges.



A1: Burglary
A2: Earthquake
A3: Alarm

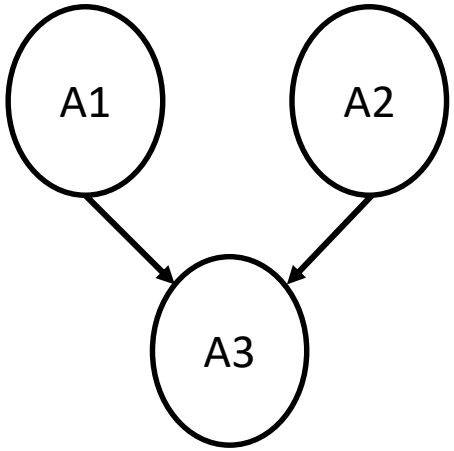
A Simple Example (1)

To define the BN, we need to find (learn) the following probabilities using data collected so far, or based on experiences of a human expert.

Probability of A1	
A1=1	0.1
A1=0	0.9

Probability of A2	
A2=1	0.1
A2=0	0.9

Conditional probabilities				
	A1=1		A1=0	
	A2=1	A2=0	A2=1	A2=0
A3=1	0.607	0.368	0.135	0.001
A3=0	0.393	0.632	0.865	0.999



A1: Burglary
A2: Earthquake
A3: Alarm

A Simple Example (3)

- To make a decision, the simplest (but not efficient) way is to find the joint probability of all events, and then marginalize it based on the evidences.
- For the simple example given here, we have

$$P(A1, A2, A3) = P(A3|A1, A2)P(A1, A2) \quad (1)$$
$$= P(A3|A1, A2)P(A1)P(A2)$$

- The second lines uses the fact (we believe) that A1 and A2 are independent.
- From the joint probability, we can find P(A1,A3), P(A2,A3), and P(A3) via marginalization.
- Using Bayes theorem, we can find
  - P(A1|A3)=P(A1,A3)/P(A3) ← probability of A1 when A3 is true
  - P(A2|A3)=P(A2,A3)/P(A3) ← probability of A2 when A3 is true

A Simple Example (4)

- Using the probabilities defined for the given example, we have
  - P(A1|A3)=P(A1,A3)/P(A3)=0.751486
  - P(A2|A3)=P(A2,A3)/P(A3)=0.349377
- Thus, for this example, when the alarm is activated, the most probable factor is burglary.
- Note that A1, A2, and A3 can represent other events, and the decisions can be made in the same way.
- For example, a simple BN can be defined for car diagnosis:
  - A3: Car engine does not work
  - A1: Run out of battery
  - A2: Engine starter problem

Formal Definition of Bayesian Network (1)

Reference: Bayesian Network, Maomi Ueno, 2013.

- Formally, a BN is defined by a 2-tuple or pair <G, Θ>, where G is a directed acyclic graph (DAG), and Θ is a set of probabilities (or weights of the edges).
- Each node (vertex) of G represents an event or a random variable Xi, and the edge (i,j) exists if the j-th node depends on the i-th node.
- The joint probability of the whole BN can be found by

$$P(x) = \prod_{i=1}^{N} P(x_i|\pi_i) \qquad (2)$$

- where πi is the set of parent nodes of xi.

Formal Definition of Bayesian Network (2)

- In fact, to find the joint probability using Eq. (2), we need some assumptions.
- Given a graph G. Let X, Y, and Z be three disjoint sets of nodes. If all paths between nodes of X and those of Y contain at least one node of Z, we say that **X and Y are separated by Z.** This fact is denoted by $I(X,Y|Z)_G$.
- On the other hand, we use I(X,Y|Z) to denote the fact that **X and Y are conditional independent given Z**.
- If $I(X,Y|Z)_G$ is equivalent to I(X,Y|Z) for all disjoint sets of nodes, G is said a **perfect map** (p-map).

Formal Definition of Bayesian Network (3)

- In fact, DAG is not powerful enough to represent all kinds of probability models.
- In the study of BN, we are more interested in graphs that can represent true conditional independences. These kind of graphs are called **independent map** (I-map).
- For a I-map, we have $I(X, Y|Z)_G \rightarrow I(X, Y|Z)$.
- To avoid trivial I-maps (complete graphs), we are more interested in **minimum I-maps**.
- A minimum I-map is an I-map with the minimum number of edges (if we delete any edge from G, G will not be I-map any more).

Formal Definition of Bayesian Network (4)

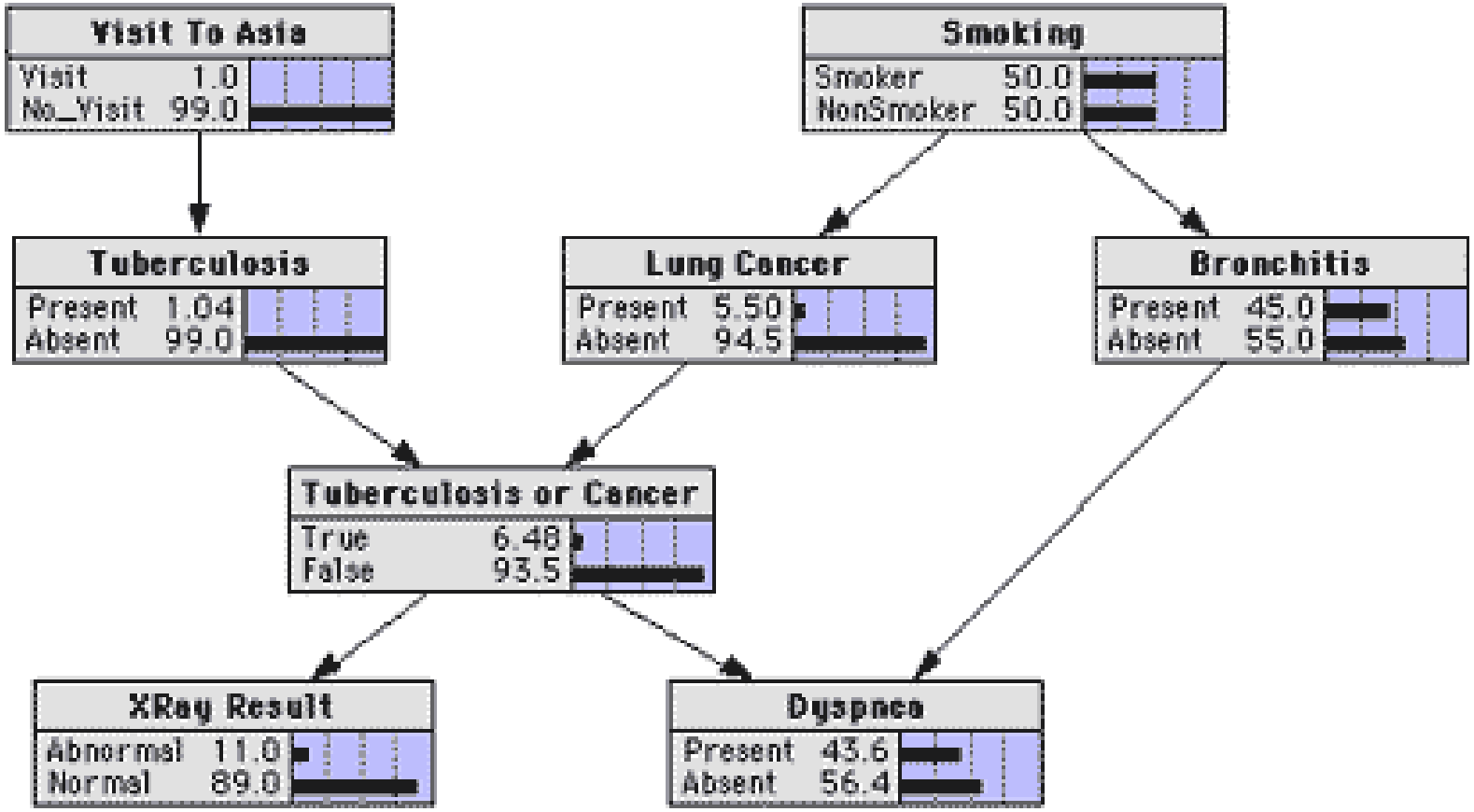
- D-separation is an important concept in BN.
- This concept is equivalent to conditional independence. That is, if X and Y are d-separated by Z, X and Y are conditional independent given Z, and vice versa.

$$I(X,Y|Z)_d \leftrightarrow I(X,Y|Z)_G \tag{3}$$

- In fact, Eq. (2) can be proved based on d-separation and the fact that there is no cycles in a BN.
- For more detail, referred to the reference book written by Maomi Ueno ("Bayesian Network", Corona-Sha).

A slightly difficult example (1)

http://www.norsys.com/networklibrary.html#



A slightly difficult example (2)

- ASIA is a BN for a fictitious medical example about
  - whether a patient has tuberculosis, lung cancer or bronchitis,
  - related to their X-ray, dyspnea, visit-to-Asia and smoking status.
- For convenience of discussion, we denote the above events using A1～A8 as follows:
  - A1: Visit to Asia; A2: Smoking; A3: Tuberculosis;
  - A4: Lung cancer; A5: Bronchitis;
  - A6: Tuberculosis or Cancer; A7: X-ray result; A8: Dyspnea.

Reference: “Mathematics in statistic inference of Bayesian network,” K. Tanaka, 2009

A slightly difficult example (3)

P(A1)	
A1=1	0.01
A1=0	0.99

P(A2)	
A2=1	0.5
A2=0	0.5

P(A3|A1)		
	A1=1	A1=0
A3=1	0.05	0.01
A3=0	0.95	0.99

P(A4|A2)		
	A2=1	A2=0
A4=1	0.1	0.01
A4=0	0.9	0.99

P(A5|A2)		
	A2=1	A2=0
A5=1	0.6	0.3
A5=0	0.4	0.7

P(A7|A6)		
	A1=6	A6=0
A7=1	0.98	0.05
A7=0	0.02	0.95

A slightly difficult example (4)

P(A6|A3, A4)				
	A3=1		A3=0	
	A4=1	A4=0	A4=1	A4=0
A6=1	1	1	1	0
A6=0	0	0	0	1

P(A8|A5, A6)				
	A5=1		A5=0	
	A6=1	A6=0	A6=1	A6=0
A8=1	0.9	0.8	0.7	0.1
A8=0	0.1	0.2	0.3	0.9

A slightly difficult example (5)

- Using Eq. (2), we have
- $P(A_1, \cdots, A_8) = P(A_8|A_5, A_6)P(A_7|A_6)P(A_6|A_3, A_4) \times P(A_5|A_2)P(A_4|A_2)P(A_3|A_1)P(A_1)P(A_2)$
- Marginalize this joint probability, we can find, for example

$$P(A_8) = \sum_{A_1}\sum_{A_2}\sum_{A_3}\sum_{A_4}\sum_{A_5}\sum_{A_6}\sum_{A_7} P(A_1, \cdots, A_8)$$

$$P(A_3, A_8) = \sum_{A_1}\sum_{A_2}\sum_{A_4}\sum_{A_5}\sum_{A_6}\sum_{A_7} P(A_1, \cdots, A_8)$$

- Using Bayesian theorem, we can find P(A3|A8) from P(A8) and P(A3,A8), and see A3 is the factor for causing A8.

A3: Tuberculosis A8: Dyspnea

A slightly difficult example (6)

- For the given BN, we have the results given below.
- From these results, we can see that probably A3 (Tuberculosis) is not the main factor for causing A8 (Dyspnea).

P(A8)	
A8=1	0.43597
A8=0	0.56403

P(A3|A8)		
	A8=1	A8=0
A3=1	**0.01883**	0.00388
A3=0	**0.98117**	0.99612

		P(A3,A8)
A3=1	A8=1	0.00821
A3=1	A8=0	0.02189
A3=0	A8=1	0.42776
A3=0	A8=0	0.56184

Learning of BN (1)

Reference: Bayesian Network, Maomi Ueno, 2013.

- Consider the learning problem of a BN containing N nodes representing the probability model of $x = \{x_1, x_2, \ldots, x_N\}$, where each variable can take a limited number of discrete values.
- From Eq. (2) we can see that to define a BN, we need to determine the conditional probabilities $P(x_i|\pi(x_i))$, where $\pi(x_i)$ is the set of parent nodes of x_i.
- If x_i can take r_i values, and $\pi(x_i)$ can take q_i patterns (=number of combinations of all parent nodes), altogether we have $N \cdot \prod_{i=1}^{N} r_i \cdot q_i$ parameters to determine in the learning process.

Learning of BN (2)

- Denoting the parameters by $\Theta = \{\theta_{ijk}\}, i = 1,2,\dots,N; j = 1,2,\dots,q_i; k = 1,2,\dots,r_i$, we can estimate Θ based on a training set.
- Suppose that we have already defined the structure of the BN. That is, G is given.
- Given an set of observations **x**, the likelihood of Θ is given by

$$P(\boldsymbol{x}|\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{q_i}\Delta_{ij}\prod_{k=1}^{r_i}\theta_{ijk}^{n_{ijk}} \propto \prod_{i=1}^{N}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{n_{ijk}} \qquad (4)$$

where

$$\Delta_{ij} = \frac{(\sum_{k=1}^{r_i} n_{ijk})}{\prod_{k=1}^{r_i} n_{ijk}!} \qquad (5)$$

is a normalization factor to make the sum of all probabilities 1.

- Here we have assumed that the likelihood follows the multinomial distribution.

Learning of BN (3)

- Where in Eq. (4), n_{ijk} is the number of observations in which xi takes the k-th value when the parent set takes the j-th pattern.
- To find the MAP (Maximum a posteriori) estimation of Θ, we assume that Θ follows Dirichlet distribution given by

$$P(\Theta) = \prod_{i=1}^{N} \prod_{j=1}^{q_i} \delta_{ij} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} \tag{6}$$

Where

$$\delta_{ij} = \frac{\Gamma(\sum_{k=1}^{r_i} \alpha_{ijk})}{\prod_{k=1}^{r_i} \Gamma(\alpha_{ijk})} \tag{7}$$

Γ is gamma function satisfying $\Gamma(x+1) = x\Gamma(x)$, and α_{ijk} is the hyper-parameter.

- The *a priori* probability (6) takes the same form as the likelihood given by (4).

Learning of BN (4)

- From Eqs. (4) and (6), we can find the joint probability P(**x**,Θ) as follows:

$$P(x,\Theta) \propto \prod_{i=1}^{N}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{\alpha_{ijk}+n_{ijk}-1} \tag{8}$$

- Maximize the above probability with respect to Θ, we get the following MAP estimation:

$$\widehat{\theta_{ijk}} = \frac{\alpha_{ijk}+n_{ijk}-1}{\alpha_{ij}+n_{ij}-r_i} \tag{9}$$

- Where $\alpha_{ij} = \sum_{k=1}^{r_i}\alpha_{ijk}$, and $n_{ij} = \sum_{k=1}^{r_i} n_{ijk}$.

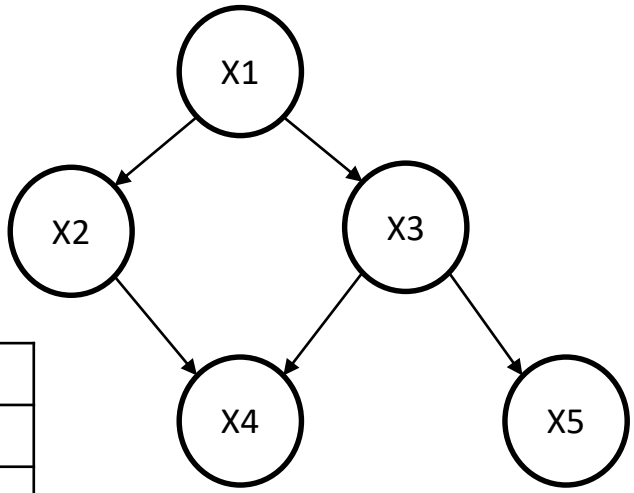
Evaluation of the learning algorithm

- To verify the learning algorithm, we can use a known BN, and re-estimate its parameters using the algorithm.
- The first step is to build a "*joint probability distribution table*" (JPDT), which is similar to the truth table for a logic formula.
- A JPDT has two columns. The left column contains all possible patterns of the variables and the right column contains the probability values. If there are 5 variables, there are 32 rows.
- The JPDT is built based on Eq. (2), which in turn is determined by using the structure of the BN and the given probabilities.
- Based on the JPDT, we can generate as many data as we want, and then re-estimate the parameters using these data.

An example of learning (1)

Reference: Bayesian Network, Maomi Ueno, 2013.

- A BN is given in the right figure.



P(X2|X1)

X1=1	X2=1	0.8
X1=1	X2=0	0.2
X1=0	X2=1	0.2
X1=0	X2=0	0.8

P(X1)

X1=1	0.8
X1=0	0.2

P(X3|X1)

X1=1	X3=1	0.8
X1=1	X3=0	0.2
X1=0	X3=1	0.2
X1=0	X3=0	0.8

P(X4|X2,X3)

X2, X3		X4	
1	1	1	0.8
1	1	0	0.2
1	0	1	0.6
1	0	0	0.4
0	1	1	0.6
0	1	0	0.4
0	0	1	0.2
0	0	0	0.8

P(X5|X3)

X3=1	X5=1	0.8
X3=1	X5=0	0.2
X3=0	X5=1	0.2
X3=0	X5=0	0.8

An example of learning (2)

- Based on the given probabilities, we can generate as many data as we can.
- Suppose now we have 20 data given in the right table, we can estimate the probabilities from the data, using the MAP estimation.
- The results is given in the next page.

#	X1,X2,X3,X4,X5
1	10111
2	00010
3	11111
4	11110
5	11110
6	11000
7	10000
8	00000
9	00000
10	01111
11	00000
12	11111
13	00111
14	11111
15	10111
16	11100
17	11010
18	11100
19	11010
20	11101

An example of learning (3)

- In the MAP estimation, we have assumed that all hyper-parameters are ½.

	True value	MAP estimation ($\alpha_{ijk} = 0.5$)
P(X2=1 | X1=1)	0.8	0.78
P(X2=1 | X1=0)	0.2	0.08
P(X3=1 | X1=1)	0.8	0.78
P(X3=1 | X1=0)	0.2	0.08
P(X4=1 | X2=1,X3=1)	0.8	0.65
P(X4=1 | X2=1, X3=0)	0.6	0.5
P(X4=1 | X2=0, X3=1)	0.6	0.5
P(X4=1 | X2=0, X3=0)	0.2	0.41
P(X5=1 | X3=1)	0.8	0.42
P(X5=1 | X3=0)	0.2	0.34
Mean squared error		0.028

Important topics in using BN

- In this lecture, we have studied the fundamental knowledge related to BN.
- To use BN more efficiently we should use algorithms that can calculate the various probabilities quickly, because marginalization can be very expensive when the number of nodes is large.
- To learn a BN with a proper structure, we need to adopt other algorithms (e.g. meta-heuristics like genetic algorithm), based on some optimization criterion (e.g. minimum description length, MDL).
- In any case, to obtain a good BN, we need a large amount of data, to guarantee the quality of the probabilities.

Homework

- For the example given in p. 23, write the equation for finding the joint probability based on Eq. (2).
- Based on the equation, write a program to build a “joint probability distribution table” (JPDT) based on your equation.
- Based on the JPDT, write a program to generate 100 data.