Independence and Bayesian Networks (Part 1)

Wenhu Chen

Lecture 3

Outline

Learning Goals

- Given a probabilistic model,
  determine if two variables are unconditionally independent, or conditionally independent given a third variable.
- Give examples of deriving a compact representation of a joint distribution by using independence assumptions.
- Describe components of a Bayesian network.
- Compute a joint probability given a Bayesian network.
- Explain the independence relationships in the three key structures.

Learning Goals

Unconditional and Conditional Independence

Examples of Bayesian Networks

Why Bayesian Networks

Representing the Joint Distribution

Independence in Three Key Structures

Revisiting Learning Goals

(Unconditional) Independence

Definition ((unconditional) independence)

X and Y are (unconditionally) independent iff

$$P(X|Y) = P(X)$$

$$P(Y|X) = P(Y)$$

$$P(X \wedge Y) = P(X)P(Y)$$

Learning Y does NOT influence your belief about X.

(Unconditional) Independence

Definition ((unconditional) independence)

X and Y are (unconditionally) independent iff

$$P(X|Y) = P(X)$$

$$P(Y|X) = P(Y)$$

$$P(X \wedge Y) = P(X)P(Y)$$

Learning Y does NOT influence your belief about X.

$\rightarrow$ Convert between the two equations.

To specify joint probability, it is sufficient to specify the individual probabilities.

(Unconditional) Independence

Definition ((unconditional) independence)

X and Y are (unconditionally) independent iff

$$P(X|Y) = P(X)$$

$$P(Y|X) = P(Y)$$

$$P(X \wedge Y) = P(X)P(Y)$$

Learning Y does NOT influence your belief about X.

(Unconditional) Independence

Definition ((unconditional) independence)

X and Y are (unconditionally) independent iff

$$P(X|Y) = P(X)$$

$$P(Y|X) = P(Y)$$

$$P(X \land Y) = P(X)P(Y)$$

Learning Y does NOT influence your belief about X.

$\rightarrow$ To justify that

$$P(X \land Y) = P(X)P(Y)$$

we need to make four comparisons.

Conditional Independence

Definition (conditional independence)

X and Y are conditionally independent given Z if

$$P(X|Y \wedge Z) = P(X|Z).$$

$$P(Y|X \wedge Z) = P(Y|Z).$$

$$P(Y \wedge X|Z) = P(Y|Z)P(X|Z).$$

Learning Y does NOT influence your belief about X
if you already know Z.

Conditional Independence

Definition (conditional independence)

X and Y are conditionally independent given Z if

$$P(X|Y \wedge Z) = P(X|Z).$$

$$P(Y|X \wedge Z) = P(Y|Z).$$

$$P(Y \wedge X|Z) = P(Y|Z)P(X|Z).$$

Learning Y does NOT influence your belief about X
if you already know Z.

$\rightarrow$ X is conditionally independent of Y given Z.

Independence does not imply conditional independence, and vice versa.

Conditional Independence

Definition (conditional independence)

X and Y are conditionally independent given Z if

$$P(X|Y \wedge Z) = P(X|Z).$$

$$P(Y|X \wedge Z) = P(Y|Z).$$

$$P(Y \wedge X|Z) = P(Y|Z)P(X|Z).$$

Learning Y does NOT influence your belief about X
if you already know Z.

Conditional Independence

Definition (conditional independence)

X and Y are conditionally independent given Z if

$$P(X|Y \wedge Z) = P(X|Z).$$

$$P(Y|X \wedge Z) = P(Y|Z).$$

$$P(Y \wedge X|Z) = P(Y|Z)P(X|Z).$$

Learning Y does NOT influence your belief about X if you already know Z.

$\rightarrow$ To justify that

$$P(X \wedge Y|Z) = P(X|Z)P(Y|Z)$$

we need to make eight comparisons.

Q #1: Deriving a compact representation

Q: Consider a model with three random variables, A, B, C. What is the minimum number of probabilities required to specify the joint distribution?

(A) 3

(B) 7

(C) 8

(D) 16

Q #1: Deriving a compact representation

Q: Consider a model with three random variables, A, B, C. What is the minimum number of probabilities required to specify the joint distribution?

(A) 3

(B) 7

(C) 8

(D) 16

$\rightarrow$ (C) $P(A), P(B|A), P(C|A \wedge B)$. $1 + 2 + 4 = 7$ probabilities

Draw a graph to prove it to yourself.

Q #2: Deriving a compact representation

Q: Consider a model with three random variables, A, B, C. Assume that A, B, and C are independent. What is the minimum number of probabilities required to specify the joint distribution?

(A) 3

(B) 7

(C) 8

(D) 16

Q #2: Deriving a compact representation

Q: Consider a model with three random variables, A, B, C. Assume that A, B, and C are independent. What is the minimum number of probabilities required to specify the joint distribution?

(A) 3

(B) 7

(C) 8

(D) 16

$\rightarrow$ (A) $P(A), P(B), P(C)$. $1 + 1 + 1 = 3$ probabilities

Draw a graph to prove it to yourself.

Q #3: Deriving a compact representation

Q: Consider a model with three boolean random variables, A, B, C. Assume that A and B are conditionally independent given C. What is the minimum number of probabilities required to specify the joint distribution?

(A) 4

(B) 5

(C) 7

(D) 11

Q #3: Deriving a compact representation

Q: Consider a model with three boolean random variables, A, B, C. Assume that A and B are conditionally independent given C. What is the minimum number of probabilities required to specify the joint distribution?

(A) 4

(B) 5

(C) 7

(D) 11

$\rightarrow$ (B) $P(C), P(A|C), P(B|C)$. $1 + 2 + 2 = 5$ probabilities

Draw a graph to prove it to yourself.

Q #3a: Deriving a compact representation

Q: Given the joint probability distribution about A, B, and C, how many comparisons do we need to make to justify A and B are independent given C?

(A) 1

(B) 4

(C) 8

(D) 10

Q #3a: Deriving a compact representation

Q: Given the joint probability distribution about A, B, and C, how many comparisons do we need to make to justify A and B are independent given C?

(A) 1

(B) 4

(C) 8

(D) 10

$\rightarrow$ (C)
$p(B=T, A=T|C=T) = p(B=T|A=T) * p(C=T|C=T)$
$p(B=T, A=F|C=T) = p(B=T|C=T) * p(B=F|C=T)$
... A total of 8 eqaulities!

Q #3b: Deriving a compact representation

Q: Read the table to understand whether B and C are independent given A.

A	B	C	Prob
T	T	T	0.16
T	T	F	0.16
T	F	T	0.24
T	F	F	0.24
F	T	T	0.012
F	T	F	0.008
F	F	T	0.108
F	F	F	0.072

(A) B and C are independent given A

(B) B and C are not independent given A

Q #3b: Deriving a compact representation

Read the table to understand whether B and C are independent given A.

A	B	C	Prob
T	T	T	0.16
T	T	F	0.16
T	F	T	0.24
T	F	F	0.24
F	T	T	0.012
F	T	F	0.008
F	F	T	0.108
F	F	F	0.072

- Compute $p(B, C|A)$
- Compute $p(B|A)$ and $p(C|A)$
- Verify $p(B, C|A) = p(B|A) * p(C|A)$

Q #3b: Step-by-Step Derivation $p(B, C|A)$

A	B	C	Prob
T	T	T	0.16
T	T	F	0.16
T	F	T	0.24
T	F	F	0.24
F	T	T	0.012
F	T	F	0.008
F	F	T	0.108
F	F	F	0.072

Table: Merging $p(A, B, C)$.

- $p(B, C|A) = p(B, C, A)/p(A)$
- $p(A) = (0.16 + 0.16 + 0.24 + 0.24, 0.012 + 0.008 + 0.108 + 0.072) = (0.8, 0.2)$

Q #3b: Step-by-Step Derivation $p(B, C|A)$

B	C	(A)	Prob
T	T	T	0.16 / 0.8 = 0.2
T	F	T	0.16 / 0.8 = 0.2
F	T	T	0.24 / 0.8 = 0.3
F	F	T	0.24 / 0.8 = 0.3
T	T	F	0.012 / 0.2 = 0.06
T	F	F	0.008 / 0.2 = 0.04
F	T	F	0.108 / 0.2 = 0.54
F	F	F	0.072 / 0.2 = 0.36

Table: Computing $p(B, C|A)$.

- $p(B, C|A) = p(B, C, A)/p(A)$
- $p(A) = (0.8, 0.2)$
- $p(B, C|A)$ is displayed in the table

Q #3b: Step-by-Step Derivation

A	B	C	Prob
T	T	T	0.16
T	T	F	0.16
T	F	T	0.24
T	F	F	0.24
F	T	T	0.012
F	T	F	0.008
F	F	T	0.108
F	F	F	0.072

Table: Merge $p(A, B, C)$

- Marginalizing over variable C

Q #3b: Step-by-Step Derivation $p(B|A)$

A	B	Prob
T	T	0.32
T	F	0.48
F	T	0.02
F	F	0.18

Table: Computing $p(A, B)$

- Marginalizing over variable C
- Joint $p(A, B)$ is displayed in the table

Q #3b: Step-by-Step Derivation $p(B|A)$

B	(A)	Prob
T	T	0.32 / 0.8 = 0.4
F	T	0.48 / 0.8 = 0.6
T	F	0.02 / 0.2 = 0.1
F	F	0.18 / 0.2 = 0.9

Table: Computing $p(B|A)$

- Marginalizing over variable C
- Conditional $p(B|A)$ is displayed in the table

Q #3b: Step-by-Step Derivation $p(C|A)$

A	B	C	Prob
T	T	T	0.16
T	F	T	0.24
T	T	F	0.16
T	F	F	0.24
F	T	T	0.012
F	F	T	0.108
F	T	F	0.008
F	F	F	0.072

Table: Merging $p(A, B, C)$

- Marginalizing over variable B

Q #3b: Step-by-Step Derivation $p(C|A)$

A	C	Prob
T	T	0.4
T	F	0.4
F	T	0.12
F	F	0.08

Table: Computing $p(A, C)$

- Marginalizing over variable B

Q #3b: Step-by-Step Derivation $p(C|A)$

C	(A)	Prob
T	T	$0.4 \ / \ 0.8 = 0.5$
F	T	$0.4 \ / \ 0.8 = 0.5$
T	F	$0.12 \ / \ 0.2 = 0.6$
F	F	$0.08 \ / \ 0.2 = 0.4$

Table: Computing $p(C|A)$

- Marginalizing over variable B
- Computing $p(C|A)$

Q #3b: Step-by-Step Derivation (Verification)

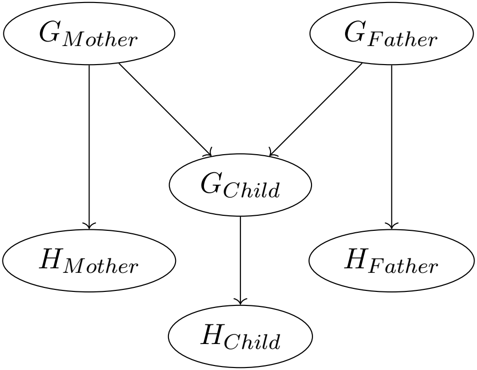
B	(A)	Prob
T	T	0.4
F	T	0.6
T	F	0.1
F	F	0.9

C	(A)	Prob
T	T	0.5
F	T	0.5
T	F	0.6
F	F	0.4

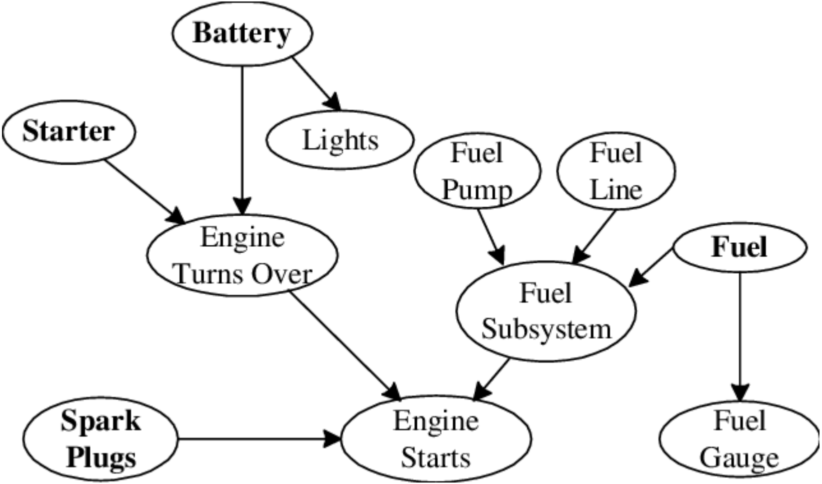
B	C	(A)	Prob
T	T	T	0.5 * 0.4 == 0.2
T	F	T	0.5 * 0.4 == 0.2
F	T	T	0.5 * 0.6 == 0.3
F	F	T	0.5 * 0.6 == 0.3
T	T	F	0.6 * 0.1 == 0.06
T	F	F	0.4 * 0.1 == 0.04
F	T	F	0.6 * 0.9 == 0.54
F	F	F	0.4 * 0.9 == 0.36

All of the probabilities are equal, therefore B and C are independent given A.

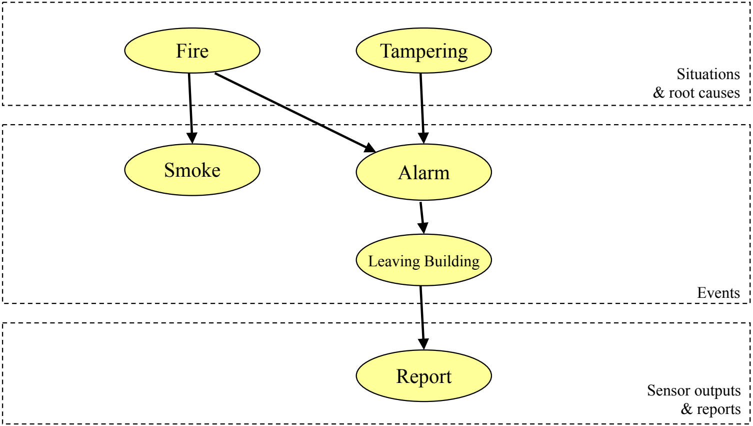
Inheritance of Handedness



Car Diagnostic Network

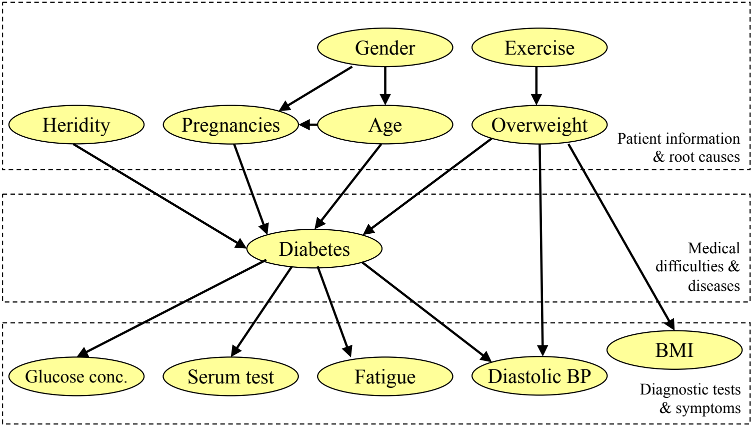


Example: Fire alarms



Report: “report of people leaving building because a fire alarm went off”

Example: Medical diagnosis of diabetes



Why Bayesian Networks?

A probabilistic model of the Holmes scenario:

- The random variables:
  Earthquake, Radio, Burglary, Alarm, Watson, and Gibbon.
- # of probabilities in the joint distribution: $2^6 = 64$.
- For example,

  $P(E \land R \land B \land A \land W \land G) =?$

  $P(E \land R \land B \land A \land W \land \neg G) =?$

  ... etc ...

We can compute any probability using the joint distribution, but

- Quickly become intractable as the number of variables grows.
- Unnatural and tedious to specify all the probabilities.

Why Bayesian Networks?

A Bayesian Network

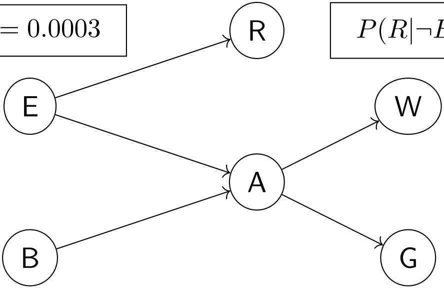
- is a compact version of the joint distribution
- takes advantage of the unconditional and conditional independence among the variables.

Reminder: Modelling the Holmes Scenario

→ The random variables:

- B: A Burglary is happening.
- A: The alarm is going.
- W: Dr. Watson is calling.
- G: Mrs. Gibbon is calling.
- E: Earthquake is happening.
- R: A report of earthquake is on the radio news.

A Bayesian Network for the Holmes Scenario



$P(E) = 0.0003$

$P(R|\neg E) = 0.0002,\ P(R|\ E) = 0.9$

$P(W|\neg A) = 0.4$
$P(W|\ A) = 0.8$

$P(G|\neg A) = 0.04$
$P(G|\ A) = 0.4$

$P(B) = 0.0001$

$P(A|\neg B \wedge \neg E) = 0.01,\ P(A|\neg B \wedge\ E) = 0.2$
$P(A|\ B \wedge \neg E) = 0.95,\ P(A|\ B \wedge\ E) = 0.96$

How many probabilities do we need to encode the Network?

Bayesian Network

A Bayesian Network is a *directed acyclic graph* (DAG).

- Each node corresponds to a random variable.
- X is a parent of Y if there is an arrow from node X to node Y.

  $\rightarrow$ Like a family tree, there are parents, children, ancestors, descendants.
- Each node X_i has a conditional probability distribution $P(X_i|Parents(X_i))$ that quantifies the effect of the parents on the node.

The Semantics of Bayesian Networks

Two ways to understand Bayesian Networks:

- A representation of the joint probability distribution
- An encoding of the conditional independence assumptions

Representing the joint distribution

The idea is that, given a random variable X, a small set of variables may exist that directly affect the variable's value in the sense that X is conditionally independent of other variables given values for the directly affecting variables.

- The set of locally affecting variables is called **Markov blanket**.
- Start with a set of random variables representing all the features of the model.
- Define the **parents** of random variable X_i, written as $parents(X_i)$.
- X_i is independent from others given the $parents(X_i)$.

Representing the joint distribution

Markov Blanket: a boundary of a random variable.

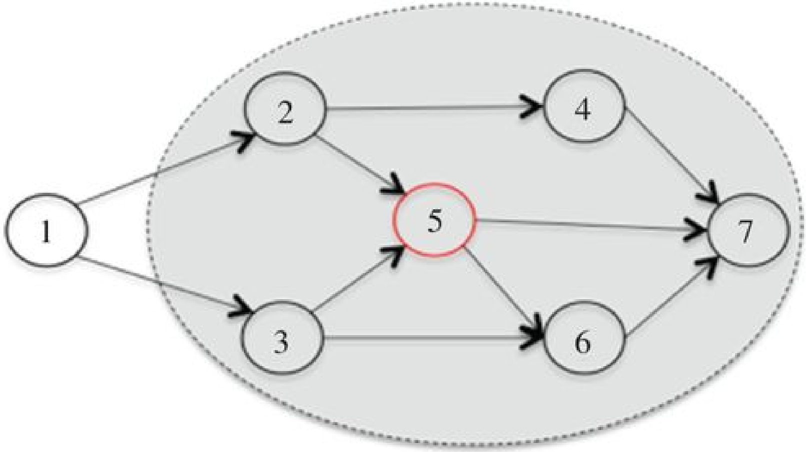


Figure: Markov Blanket for random variable 5.

Representing the joint distribution

We can compute the full joint probability using the following formula.

$$P(X_n \wedge \cdots \wedge X_1) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$

Representing the joint distribution

Example: What is the probability that all of the following occur?

- The alarm has sounded
- Neither a burglary nor an earthquake has occurred
- Both Watson and Gibbon call and say they hear the alarm
- There is no radio report of an earthquake

Representing the joint distribution

Example: What is the probability that all of the following occur?

- The alarm has sounded
- Neither a burglary nor an earthquake has occurred
- Both Watson and Gibbon call and say they hear the alarm
- There is no radio report of an earthquake

$\rightarrow$ Formulate as a joint probability:

$$
\begin{aligned}
&P(\neg B \wedge \neg E \wedge A \wedge \neg R \wedge G \wedge W) \\
&= P(\neg B)P(\neg E)P(A|\neg B \wedge \neg E)P(\neg R|\neg E)P(G|A)P(W|A) \\
&= (1-0.0001)(1-0.0003)(0.01)(1-0.0002)(0.4)(0.8) \\
&= 3.2 \times 10^{-3}
\end{aligned}
$$

Q #4: Calculating the joint probability

Q: What is the probability that all of the following occur?

- NEITHER a burglary NOR an earthquake has occurred,
- The alarm has NOT sounded,
- NEITHER of Watson and Gibbon is calling, and
- There is NO radio report of an earthquake?

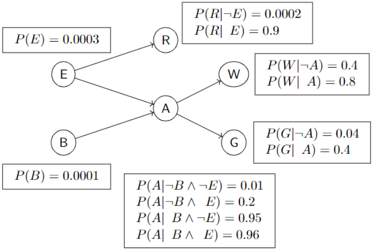
(A) 0.5699

(B) 0.6699

(C) 0.7699

(D) 0.8699

(E) 0.9699



Q #4: Calculating the joint probability

Q: What is the probability that all of the following occur?

- NEITHER a burglary NOR an earthquake has occurred,
- The alarm has NOT sounded,
- NEITHER of Watson and Gibbon is calling, and
- There is NO radio report of an earthquake?

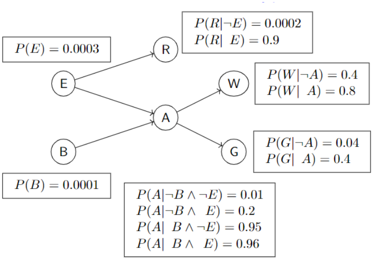
(A) 0.5699

(B) 0.6699

(C) 0.7699

(D) 0.8699

(E) 0.9699



$\rightarrow$ (A)

$$(1 - 0.0001)(1 - 0.0003)(1 - 0.01)(1 - 0.4)(1 - 0.04)(1 - 0.0002) = 0.5699$$

Burglary, Alarm and Watson

Q #5: Unconditional Independence

Q: Are Burglary and Watson independent?

(A) Yes

(B) No

(C) Can't tell.

Q #5: Unconditional Independence

Q: Are Burglary and Watson independent?

(A) Yes

(B) No

(C) Can't tell.

$\rightarrow$ Correct answer is *No*.

If you learned the value of B, would your belief of W change? If B is true, then Alarm is more likely to be true, and W is more likely to be true.

Q #6: Conditional Independence

Q: Are Burglary and Watson conditionally independent given Alarm?

(A) Yes

(B) No

(C) Can't tell

Q #6: Conditional Independence

Q: Are Burglary and Watson conditionally independent given Alarm?

(A) Yes

(B) No

(C) Can't tell

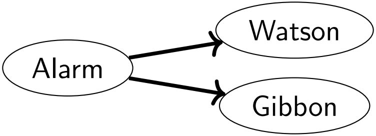
$\rightarrow$ Correct answer is *Yes*.

Assume that W does not observe B directly. W only observes A.

B and W could only influence each other through A.

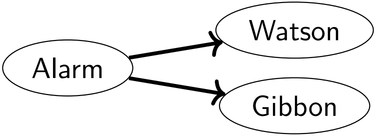
If A is known, then B and W do not affect each other.

Alarm, Watson and Gibbon



Q #7: Unconditional Independence

Q: Are Watson and Gibbon independent?



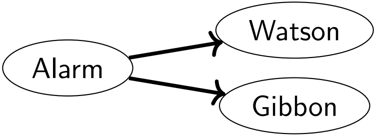
(A) Yes

(B) No

(C) Can't tell

Q #7: Unconditional Independence

Q: Are Watson and Gibbon independent?



(A) Yes

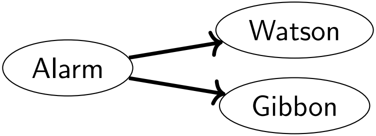
(B) No

(C) Can't tell

→ Correct answer is No. If Watson is more likely to call, then Alarm is more likely to go off, which means that Gibbon is more likely to call.

Q #8 Conditional Independence

Q: Are Watson and Gibbon conditionally independent given Alarm?



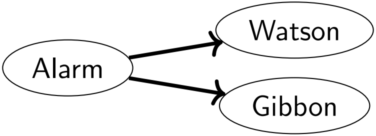
(A) Yes

(B) No

(C) Can't tell

Q #8 Conditional Independence

Q: Are Watson and Gibbon conditionally independent given Alarm?



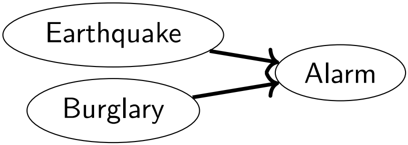
(A) Yes

(B) No

(C) Can't tell

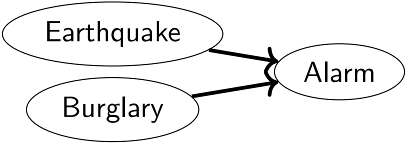
$\rightarrow$ Correct answer is Yes. Watson and Gibbon are both unreliable sensors for Alarm. If Alarm is known, then Watson and Gibbon do not affect each other.

Earthquake, Burglary, and Alarm



Q #9 Unconditional Independence

Q: Are Earthquake and Burglary independent?



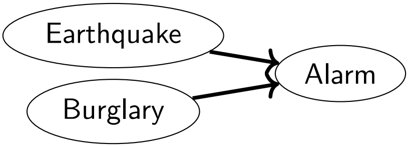
(A) Yes

(B) No

(C) Can't tell

Q #9 Unconditional Independence

Q: Are Earthquake and Burglary independent?



(A) Yes

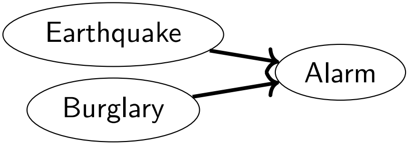
(B) No

(C) Can't tell

→ Correct answer is Yes.

Q #10: Conditional Independence

Q: Are Earthquake and Burglary conditionally independent given Alarm?



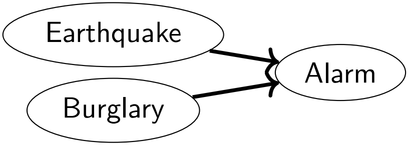
(A) Yes

(B) No

(C) Can't tell

Q #10: Conditional Independence

Q: Are Earthquake and Burglary conditionally independent given Alarm?



(A) Yes

(B) No

(C) Can't tell

$\rightarrow$ Correct answer is No. Suppose that the Alarm is going. If there is an Earthquake, then it is less likely that the Alarm is caused by Burglary. If there is a Burglary, it is less likely that the Alarm is caused by Earthquake.

Revisiting Learning Goals

- Given a probabilistic model,
  determine if two variables are unconditionally independent, or conditionally independent given a third variable.
- Give examples of deriving a compact representation of a joint distribution by using independence assumptions.
- Describe components of a Bayesian network.
- Compute a joint probability given a Bayesian network.
- Explain the independence relationships in the three key structures.