CSCI567 Machine Learning (Spring 2018)

Michael Shindler

Lecture 21: April 9

Outline

Outline

Exam Viewing

- Go to the discussion you are enrolled in this week.
- We will have your exam there if
  - You circled the time on the cover
- If you circled a different one:
  - We probably have it.
  - If too many circled a time, we reduced to enrolled.

Outline

Markov chain

Definition

Given a sequentially ordered random variables $X_1, X_2, \cdots, X_t, \cdots, X_T$, called *states*,

- **Transition probability** for describing how the state at time $t-1$ changes to the state at time t,

$$P(X_t = \text{value}' | X_{t-1} = \text{value})$$

- **Initial probability** for describing the initial state at time $t = 1$.

$$P(X_1 = \text{value})$$

value represents possible values $\{X_t\}$ can take. Note that we will assume that all the random variables (at different times) can take value from the same set and assume that the transition probability does not change with respect to time t, i.e., a stationary Markov chain.

Special case and our focus for the rest of the course

When X_t are discrete, taking values from $\{1, 2, 3, \cdots, N\}$

- Transition probability becomes a table/matrix $\boldsymbol{A}$ whose elements are

$$a_{ij} = P(X_t = j | X_{t-1} = i)$$

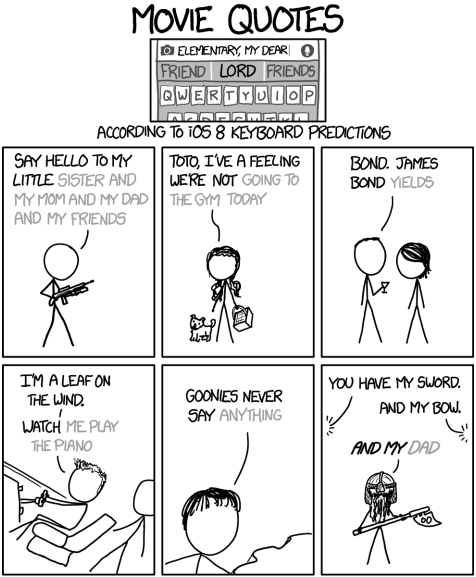
- Initial probability becomes a vector $\boldsymbol{\pi}$ whose elements are

$$\pi_i = P(X_1 = i)$$

where i or j index over from 1 to N. We have the following constraints

$$\sum_j a_{ij} = 1 \quad \sum_i \pi_i = 1$$

Additionally, all those numbers should be non-negative.


MOVIE QUOTES
ELEMENTARY, MY DEAR
FRIEND
LORD
FRIENDS
Q W E R T Y U I O P
ACCORDING TO iOS 8 KEYBOARD PREDICTIONS
SAY HELLO TO MY LITTLE SISTER AND MY MOM AND MY DAD AND MY FRIENDS
TOTO, I'VE A FEELING WE'RE NOT GOING TO THE GYM TODAY
BOND. JAMES BOND YIELDS
I'M A LEAF ON THE WIND.
WATCH ME PLAY THE PIANO
GOONIES NEVER SAY ANYTHING
YOU HAVE MY SWORD.
AND MY BOW.
AND MY DAD

High-order Markov

We have assumed the following Markov property

$$P(X_t|X_1, X_2, \cdots, X_{t-1}) = P(X_t|X_{t-1})$$

that is why we are only concerning with ourselves the *immediate* history.

We can extend to use more histories, thus high-order Markov

$$P(X_t|X_1, X_2, \cdots, X_{t-1}) = P(X_t|X_{t-1}, X_{t-2}, \cdots, X_{t-H})$$

For instance, the language model previously is an order-one HMM. Obviously, languages have long-range dependency so the past history (not just a single word) matters.

How to compute the probability of a sequence?

We need to compute

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_T = x_T)$$

We use the Markov property to factorize

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_T = x_T) = \tag{1}$$

$$P(X_1 = x_1) \prod_{t=2}^{T} P(X_t = x_t | X_{t-1} = x_{t-1}) \tag{2}$$

How to derive this? Details as an exercise but you should leverage the property in the following way:

$$\begin{aligned} P(X_1, X_2, X_3) &= P(X_3|X_1, X_2)P(X_1, X_2) \\ &= P(X_3|X_2)P(X_1, X_2) = P(X_3|X_2)P(X_2|X_1)P(X_1) \end{aligned}$$

Example

Suppose we have two possible states $X_t \in \{0, 1\}$, and we have observed the following 3 sequences

$$1\ 0\ 0\ 1$$

$$0\ 1\ 1\ 1$$

$$1\ 1\ 1\ 1$$

Thus

$$\pi_0 = \frac{1}{3}, \quad \pi_1 = \frac{2}{3}$$

and

$$a_{00} = \frac{1}{3}, \quad a_{01} = \frac{2}{3}$$
$$a_{10} = \frac{1}{6}, \quad a_{11} = \frac{5}{6}$$

Now with typo fixed!

Motivation example

Underlying process is Markov chain

Say, the temperature fluctuation in each month:
cold, cold, hot, hot, cold, hot, ...

But we observe only indirectly, through a related quantity

Say, we can measure how many scoops of ice creams that have been consumed
1, 3, 3, 2, 1, 1, ...

Question

How do we infer the trace of the temperatures from how much we have eaten the ice creams?

Formal definition of Hidden Markov Models (HMMs)

What are the variables?

- Underlying Markov chain, i.e., a set of random variables
  1. $Z_1, Z_2, \cdots, Z_t, \cdots Z_T$
  2. $Z_t \in \{s_1, s_2, s_3, \cdots, s_S\}$, a discrete set of S values
- Observed variable, i.e, a set of random variables
  1. $X_1, X_2, \cdots, X_t, \cdots X_T$
  2. $X_t \in \{o_1, o_2, o_3, \cdots, o_N\}$, a discrete set of N values

Key difference: Zs are never observed. However, their values can be inferred from the observed values Xs.

HMM defines a joint probability

$$
\begin{aligned}
&P(X_1, X_2, \cdots, X_T, Z_1, Z_2, \cdots, Z_T) \\
&= P(Z_1, Z_2, \cdots, Z_T) P(X_1, X_2, \cdots, X_T | Z_1, Z_2, \cdots, Z_T)
\end{aligned}
$$

- Markov assumption simplifies the first term

$$P(Z_1, Z_2, \cdots, Z_T) = P(Z_1) \prod_{t=2}^{T} P(Z_t | Z_{t-1})$$

- The *independence* assumption simplifies the second term

$$P(X_1, X_2, \cdots, X_T | Z_1, Z_2, \cdots, Z_T) = \prod_{t=1}^{T} P(X_t | Z_t)$$

Namely, each X_t is conditionally independent of anything else, if conditioned on Z_t.

In HMMs, we are often interested in the following problems

- Total probability of observing a whole sequence

$$P(x_1, x_2, \cdots, x_T)$$

- The most likely path of the Markov chain's states

$$(z_1^*, z_2^*, \cdots, z_T^*) = \arg\max P(z_1, z_2, \cdots, z_T | x_1, x_2, \cdots, x_T)$$

- The likelihood of a state at a given time

$$P(z_t | x_1, x_2, \cdots, x_T)$$

- The likelihood of two consecutive states at a given time

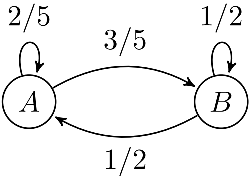
$$P(z_{t-1}, z_t | x_1, x_2, \cdots, x_T)$$

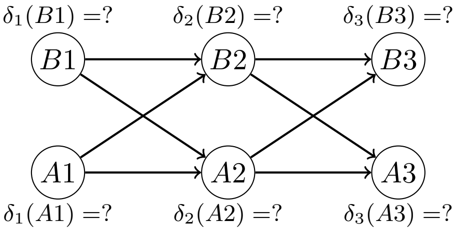
They are all related to how HMMs is to be used, as well as how to estimate parameters of HMMs from data.

Example

E	$p(E\|X=A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E\|X=B)$
"four lights"	$2/5$
"five lights"	$3/5$

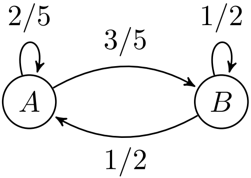


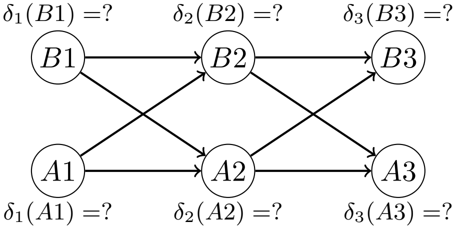


Example

E	$p(E\|X = A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E\|X = B)$
"four lights"	$2/5$
"five lights"	$3/5$



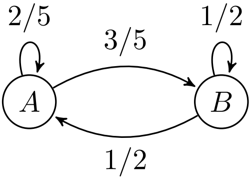


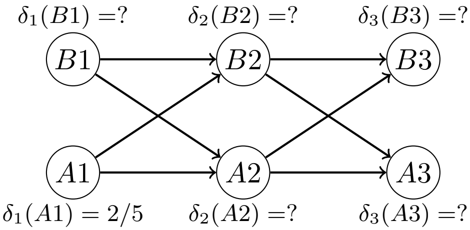
$$
\begin{aligned}
\delta_1(A) &=? \\
&= P(\mathsf{St} \to A) \cdot \delta_0(\mathsf{St}) \cdot P(4|A) \\
&= 1/2 \times 1 \times 4/5 = 2/5
\end{aligned}
$$

Example

E	$p(E\|X=A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E\|X=B)$
"four lights"	$2/5$
"five lights"	$3/5$



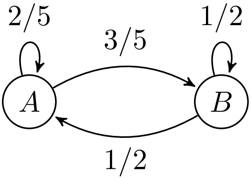


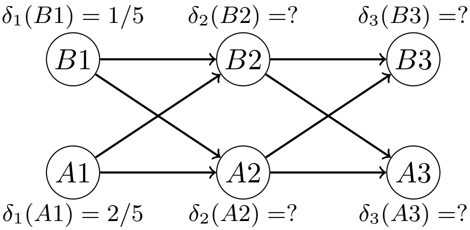
$$
\begin{aligned}
\delta_1(B) &= ? \\
&= P(\mathsf{St} \to B) \cdot \delta_0(\mathsf{St}) \cdot P(4|B) \\
&= 1/2 \times 1 \times 2/5 = 1/5
\end{aligned}
$$

Example

E	$p(E \mid X = A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E \mid X = B)$
"four lights"	$2/5$
"five lights"	$3/5$





$$\delta_2(A)? = P(A \to A) \cdot \delta_1(A) \cdot P(4|A)$$

$$? = 2/5 \cdot 2/5 \cdot 4/5 = 16/125$$

OR

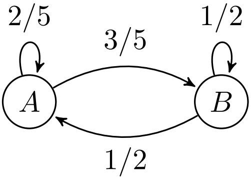
$$? = P(B \to A) \cdot \delta_1(B) \cdot P(4|A)$$

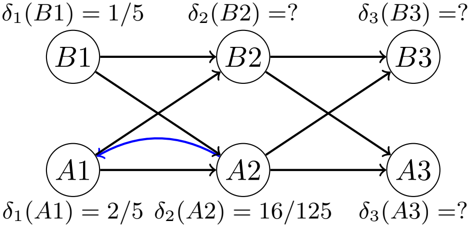
$$? = 1/2 \cdot 1/5 \cdot 4/5 = 4/50$$

Example

E	$p(E \mid X = A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E \mid X = B)$
"four lights"	$2/5$
"five lights"	$3/5$





$$\delta_2(B)? = P(A \to B) \cdot \delta_1(A) \cdot P(4|B)$$
$$? = 3/5 \times 2/5 \times 2/5 = 12/125$$

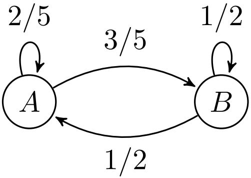
OR

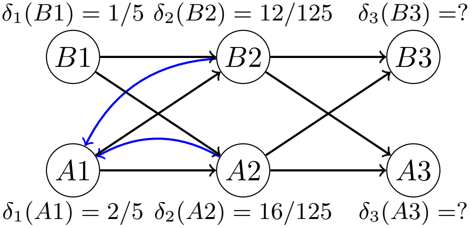
$$? = P(B \to B) \cdot \delta_1(B) \cdot P(4|B)$$
$$? = 1/2 \times 1/5 \times 2/5 = 1/25$$

Example

E	$p(E \mid X = A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E \mid X = B)$
"four lights"	$2/5$
"five lights"	$3/5$





$$\delta_3(A)? = P(A \to A) \cdot \delta_2(A) \cdot P(5|A)$$

$$? = 2/5 \times 16/125 \times 1/5$$

OR

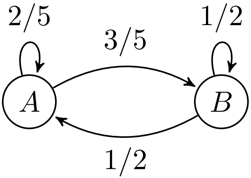
$$? = P(B \to A) \cdot \delta_2(B) \cdot P(5|A)$$

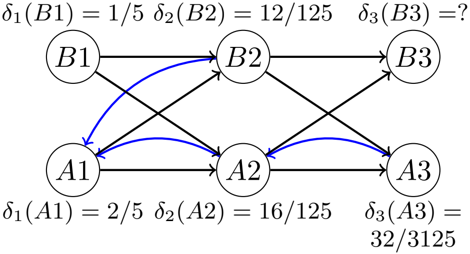
$$? = 1/2 \times 12/125 \times 1/5$$

Example

E	$p(E\|X=A)$
"four lights"	$4/5$
"five lights"	$1/5$

E	$p(E\|X=B)$
"four lights"	$2/5$
"five lights"	$3/5$





$$\delta_3(B)? = P(A \to B) \cdot \delta_2(A) \cdot P(5|B)$$

$$? = 3/5 \times 16/125 \times 3/5$$

OR

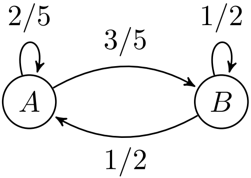
$$? = P(B \to B) \cdot \delta_2(B) \cdot P(5|B)$$

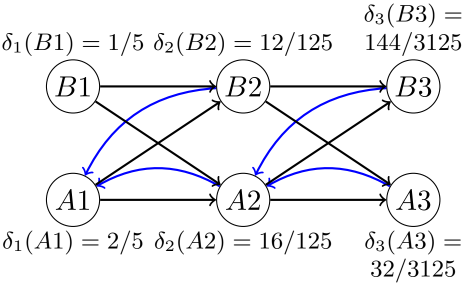
$$? = 1/2 \times 12/125 \times 3/5$$

Example

E	$p(E \mid X = A)$
“four lights”	$4/5$
“five lights”	$1/5$

E	$p(E \mid X = B)$
“four lights”	$2/5$
“five lights”	$3/5$





Most likely path?

- $\delta_3(A) = 32/3125$
- $\delta_3(B) = 144/3125$

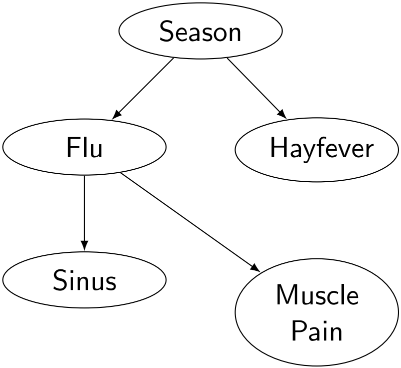
Outline

Graphical Models

- Bayes Nets
  - Probabilistic distribution represented with directed acyclic graphs (DAGs)
- Markov Networks
  - Probabilistic distribution represented with undirected graphs.

Exploring structures

- Draw links between variables
  - indicate dependencies and encode independence
  - Ex: flu and hayfever are independent in any given season
  - They independently occur *conditioned* on season
- This is example of Bayes Networks
  - Directed acyclic graphs
  - Compact representation of joint distribution

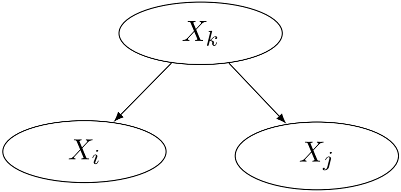


The key concept

- Conditional independence

$$X_i \perp\!\!\!\perp X_j | X_k$$

Graphical representation:



- This allows us to write:

$$\begin{aligned} p(X_i, X_j, X_k) &= p(X_i | X_j, X_k) p(X_j, X_k) \\ &= p(X_i | X_k) p(X_j | X_k) p(X_k) \end{aligned}$$

So factorizing

- An N-term joint distribution

$$P(X_1, X_2, \dots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)\dots$$
$$\dots P(X_N|X_1, X_2, \dots X_{N-1})$$

- We need only a subset of terms:

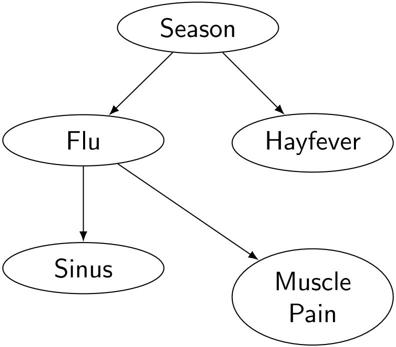
$$P(X_1, X_2, \dots X_n) = \prod_{i=1}^{N} P(X_i|S_i)$$

Where S_i is a subset of the $(N-1)$ other variables.

How is this going to help us?

Fractional and Conditional Independence

$$
\begin{aligned}
P(&\text{season=Fall, Flu=true,}\\
&\quad\text{Muscle pain=true,}\\
&\quad\text{Sinus=false,}\\
&\quad\text{Hayfever = false}) = \\
P(&\text{season=Fall})\times\\
&P(\text{Flu=true}|\text{season=Fall})\times\\
&P(\text{Hayfever=false}|\text{Season=Fall})\times\\
&P(\text{Muscle pain=true}|\text{Flu=true})\times\\
&P(\text{Sinus=true}|\text{Flu=true, Hayfever=false})
\end{aligned}
$$

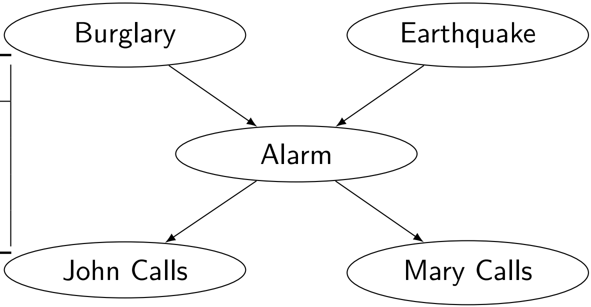


Total # parameters for 5 random variables is ?

Another (Classic) Example

$P(B)$
.001

$P(E)$
.002



B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001

J	$P(A\|J)$
T	.90
F	.05

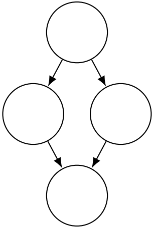
M	$P(A\|M)$
T	.70
F	.01

Formal definition of Bayesian Networks

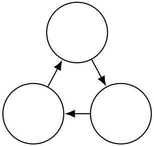
Structure (Graph G):

- Vertex are R.V.
- Edge: child depends on parent
- Is a DAG

This is okay:



This is not:



Conditional probability distributions (CPD): $P(X_i|Pa_{X_i})$ for every vertex.

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|Pa_{X_i})$$

Semantics of Bayesian Networks

The "syntax" view

Factorizing joint distribution with respect to graph structure.

What are the properties we can infer from the structure?

Semantics: local Markov property

$$X_i \perp \text{NonDescendants}_{X_i} \mid PA_{X_i}$$

The two views are equivalent

The following can be shown

- Factorization $\rightarrow$ local Markov properties

  If a distribution P factorizes according to the graph, then the distribution satisfies the local Markov properties (i.e., local conditional independences)

- Local Markov Properties

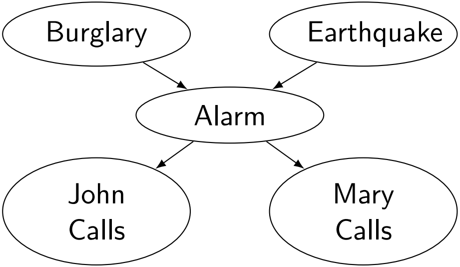
  If a distribution P satisfies local Markov properties implied in the graph, then the distribution factorizes according to the graph.

Examine the local Markov properties

$P(B)$
.001

$P(E)$
.002

B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001



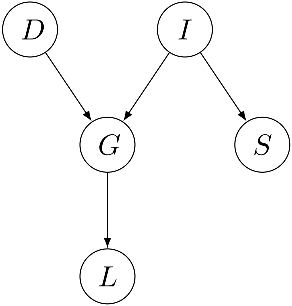
J	$P(A\|J)$
T	.90
F	.05

M	$P(A\|M)$
T	.70
F	.01

Examine the local Markov properties

$$X_i \perp \text{NonDescendants}_{X_i} \mid PA_{X_i}$$

- $L \perp I, D, S | G$
- $S \perp D, G, L | I$
- $G \perp S | D, I$
- $I \perp D$
- $D \perp I, S$



How to construct a Bayesian network

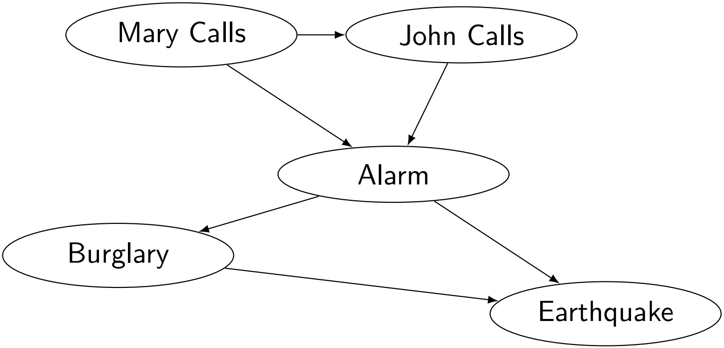
1. Choose an ordering of variables $X_1 \dots X_n$
2. For $i = 1$ to n
   - Add X_i to the network.
   - Select parent(s) from $X_1 \dots X_{i-1}$ such that

$$P(X_i|\text{Parents}(X_i)) = P(X_i|X_1 \dots, X_{i-1})$$

The choice of parents guarantees global semantics:

$$\begin{aligned} P(X_1, \dots, X_n) &= \prod_{i=1}^{n} P(X_i|X_i, \dots, X_{i-1}) && \text{(chain rule)} \\ &= \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) && \text{(by construction)} \end{aligned}$$

Different order gives a different network



How to use Bayesian Networks?

Once knowledge is encoded

We can query the network

- That is, ask questions
- That is, do (probabilistic) inference
- Let's see a few inference problems...

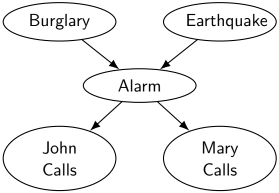
Causal Reasoning: How likely is it if John calls if there is a burglary?

Naive approach?
Better approach?

$P(B)$
.001

$P(E)$
.002

B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001



J	$P(A\|J)$
T	.90
F	.05

M	$P(A\|M)$
T	.70
F	.01

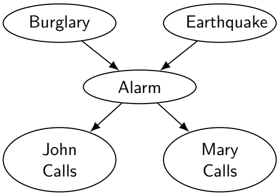
Diagnostic/Evidential Reasoning

John calls.
What is the probability there is a burglary?

$P(B)$
.001

$P(E)$
.002

B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001



J	$P(A\|J)$
T	.90
F	.05

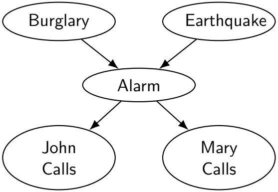
M	$P(A\|M)$
T	.70
F	.01

Explaining away

$P(B)$
.001

$P(E)$
.002

B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001



J	$P(A\|J)$
T	.90
F	.05

M	$P(A\|M)$
T	.70
F	.01

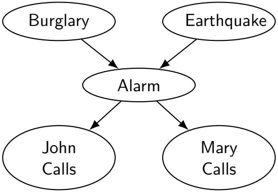
What is $P(\text{‘Burglary}' == \text{true} | \text{‘alarm}' == \text{true})$?

Explaining away

$P(B)$
.001

$P(E)$
.002

B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001



J	$P(A\|J)$
T	.90
F	.05

M	$P(A\|M)$
T	.70
F	.01

What is $P($‘Burglary$' ==$ true$|$‘alarm$' ==$ true$)$? 0.376

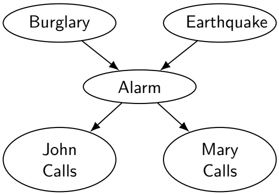
What is $P($‘Burglary$' ==$ true$|$‘alarm$' ==$ true&Earthquake $==$ ‘true$')$?

Explaining away

$P(B)$
.001

$P(E)$
.002

B	E	$P(A\|B,E)$
T	T	.95
T	F	.94
F	T	.29
F	F	.001



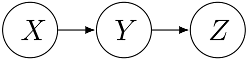
J	$P(A\|J)$
T	.90
F	.05

M	$P(A\|M)$
T	.70
F	.01

What is $P(\text{‘Burglary}' == \text{true}|\text{‘alarm}' == \text{true})$? 0.376
What is $P(\text{‘Burglary}' == \text{true}|\text{‘alarm}' == \text{true}\&\text{Earthquake} == \text{‘true}')$? 0.003

Maybe the graph can tell us more?

More independence



The local Markov property is $X \perp Z | Y$

Maybe the graph can tell us more?

More independence

$$X \to Y \to Z$$

The local Markov property is $X \perp Z | Y$

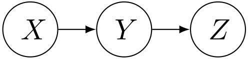
$$X \to Y \to W \to Z$$

The local Markov property is $X, Y \perp Z | W$

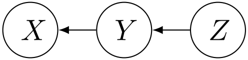
Is $X \perp Z | Y$?

Simple Cases

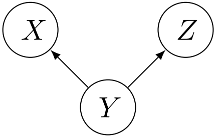
Indirect causal effect



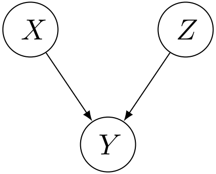
Indirect evidential effect



What are the independencies?



Common Cause

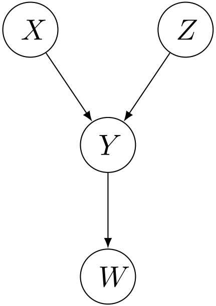


Common Effect

More v-structure

$X \perp Y$

How about? $X \perp Y | W$



But we have seen this structure before!