Probability and Uncertainty: Bayesian Networks

Will Devanny

Russell and Norvig 14.1-14.2

Today's Lecture

- Why probability?
  - What is it good for?
- Quick probability review
  - Russel and Norvig Chapter 13 or discussion or office hours
- Problems with naive usage of probabilities
- Bayesian networks

Brief History of Probability in AI

Early AI (1950-1970)

- Probability used to solve AI problems
- Mixed success

Logical AI (1970-1990)

- Researchers realize full probability models are intractable
- Abandoned probability for logic
- New problem: logic has troubles in the real world

Probabilistic AI (1990-present)

- Judea Pearl invents Bayesian Networks! (1988)
- Approximate model of probability is tractable
- Developed algorithms to learn these new models
- Techniques now used in: vision, speech, video games, etc.

Problems with Logic

Logic deals with *true*, *false*, and *unknown*

- What about a value that is almost always true?
- Living in Irvine I can reasonably act like it won't snow

No loose implications

- "If I leave two hours ahead of time I will usually arrive at the airport in time for my flight."

Solution is to use probability

- We have some belief about how likely events are
  - "99.9% chance it won't snow tomorrow."

Reverend Thomas Bayes

Lived from 1701-1761

Developed Bayes's Rule while trying to prove existence of god

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayesians vs. Frequentists

Frequentist: old school of thought

- Probability is how often a coin comes up head when flipped many times

Bayesians vs. Frequentists

Frequentist: old school of thought

- Probability is how often a coin comes up head when flipped many times

Bayesians: newer school of thought

- Probability is a belief
- I am 15% sure Brazil will win the 2014 World Cup
- My belief will change when presented with new evidence

Bayesians vs. Frequentists

Frequentist: old school of thought

Probability is how often a coin comes up head when flipped many times

Bayesians: newer school of thought

Probability is a belief

I am 15% sure Brazil will win the 2014 World Cup

My belief will change when presented with new evidence

I remember 2014 World Cup is in Brazil!

$\Rightarrow$ I am now 25% sure Brazil will win

Bayesians vs. Frequentists

Frequentist: old school of thought

Probability is how often a coin comes up head when flipped many times

Bayesians: newer school of thought

Probability is a belief

I am 15% sure Brazil will win the 2014 World Cup

My belief will change when presented with new evidence

I remember 2014 World Cup is in Brazil!

$\Rightarrow$ I am now 25% sure Brazil will win

Illustrative example: "$P($ Life on other planets"$)$"

Probability Review

Space of events: Ω

Made up of atomic events

Rolling two dice: $\Omega = \{(1,1),(1,2),(1,3),\ldots(6,6)\}$

(5,4) means first die was five and second was four

Random variable - some real valued function of atomic events

Sum of the two dice, value of the first die, etc.

Probability Review

Space of events: Ω

Made up of atomic events

Rolling two dice: $\Omega = \{(1,1),(1,2),(1,3),\ldots(6,6)\}$

(5,4) means first die was five and second was four

Random variable - some real valued function of atomic events

Sum of the two dice, value of the first die, etc.

Axioms of Probability:

1. $\forall e \in \Omega \quad P(e) \geq 0$
2. $P(\Omega) = 1$
3. If A and B are mutually exclusive then

$$P(A \vee B) = P(A) + P(B)$$

Independence

A and B are said to be independent iff $P(A \wedge B) = P(A)P(B)$

This is a very strong statement about events

- Relatively uncommon in complicated systems
  - Height and reading ability?
- However very useful when it is applicable

Independence

A and B are said to be independent iff $P(A \wedge B) = P(A)P(B)$

This is a very strong statement about events

- Relatively uncommon in complicated systems
  - Height and reading ability? No!
- However very useful when it is applicable

Independence

A and B are said to be independent iff $P(A \wedge B) = P(A)P(B)$

This is a very strong statement about events

Relatively uncommon in complicated systems

Height and reading ability? No!

However very useful when it is applicable

How do we find out two things are independent?

If we have a table of probabilities,
run through computations and check

Sometimes we can deduce or assume independence
from our model of the world

Probability as Generalized Logic

Statements in logic are one of three values:

True, False, or Unknown

Real world not always simple implication

What if a statement is true in all but one possible model?

Uncertainty due to:

- Things we did/could not measure
- Imperfect knowledge
- Noisy measurements

Probability as Generalized Logic

Statements in logic are one of three values:

True, False, or Unknown

Real world not always simple implication

What if a statement is true in all but one possible model?

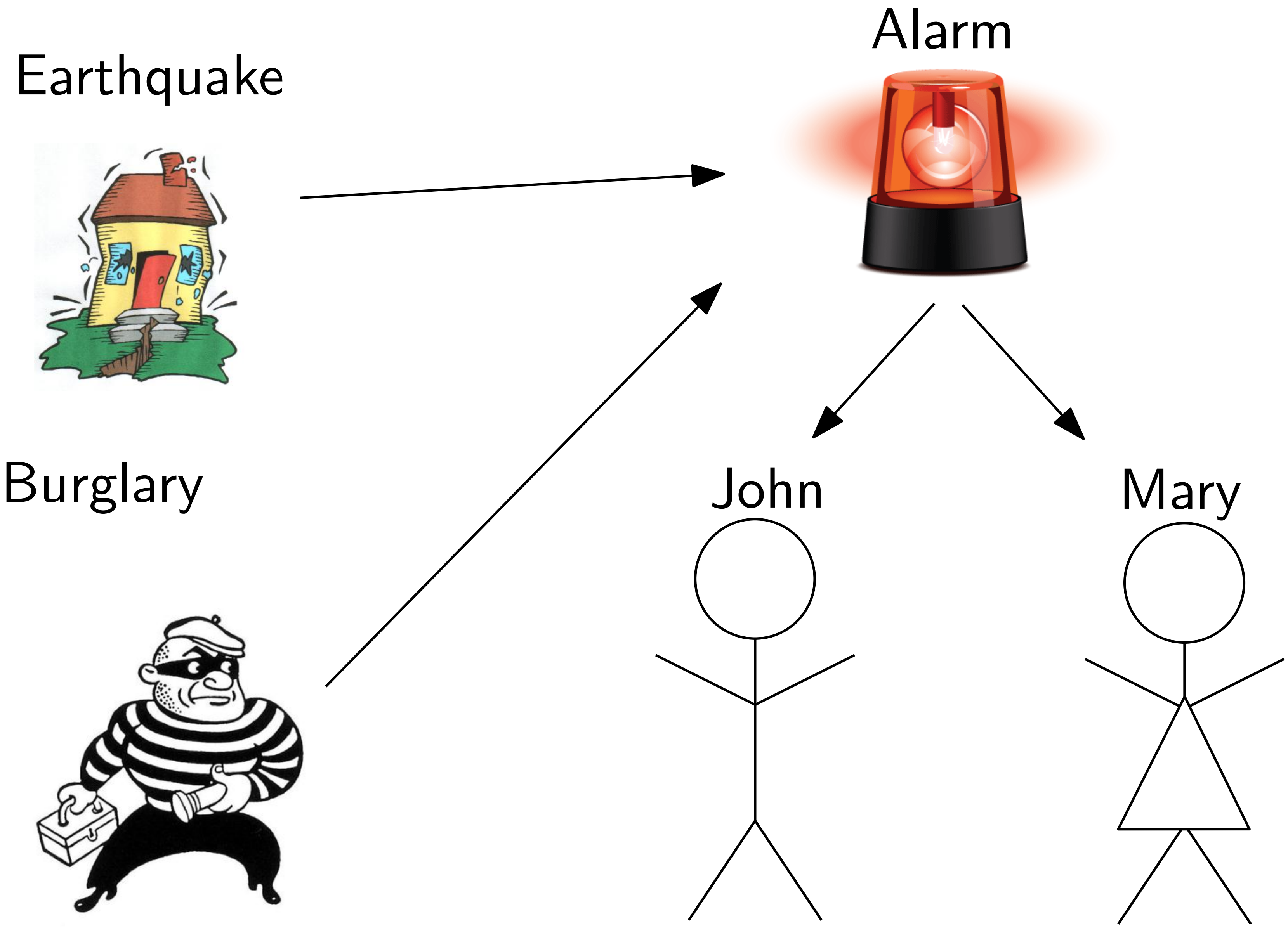
Uncertainty due to:

- Things we did/could not measure
- Imperfect knowledge
- Noisy measurements

Probability

- False $= 0$, True $= 1$, Unknown $\in [0, 1]$
- Represent uncertainty and partial knowledge in probabilities

An Example



The Random Variables

E - was there an earthquake?

B - was there a burglary?

A - did the alarm go off?

J - did John call me?

M - did Mary call me?

We will use uppercase when talking about a variable
and lowercase when assigning that variable

e.g. a means the alarm went off and
$\neg e$ means there was no earthquake

Law of Total Probability

If we have probability of all atomic events then
we can use sums to find probability of a random variable

$$P(a) = P(a \wedge b) + P(a \wedge \neg b)$$

If more variables:

$$P(a) = \sum_{x \in B} \sum_{y \in C} \sum_{z \in D} P(a \wedge x \wedge y \wedge z)$$

To compute this summation we use a joint distribution table

Joint Distribution

A giant table of probabilities

		a	$\neg a$
e	b	0.0001	0.00001
	$\neg b$	0.0009	0.00099
$\neg e$	b	0.0008	0.00009
	$\neg b$	0.00001	0.9971

$$P(a \wedge b \wedge \neg e) =$$

Joint Distribution

A giant table of probabilities

		a	$\neg a$
e	b	0.0001	0.00001
	$\neg b$	0.0009	0.00099
$\neg e$	b	$\boxed{0.0008}$	0.00009
	$\neg b$	0.00001	0.9971

$$P(a \wedge b \wedge \neg e) = \boxed{0.0008}$$

Joint Distribution

A giant table of probabilities

		a	$\neg a$
e	b	0.0001	0.00001
	$\neg b$	0.0009	0.00099
$\neg e$	b	0.0008	0.00009
	$\neg b$	0.00001	0.9971

$$P(a \wedge b \wedge \neg e) = \ 0.0008$$

$$P(a) = ?$$

Joint Distribution

A giant table of probabilities

		a	$\neg a$
e	b	0.0001	0.00001
	$\neg b$	0.0009	0.00099
$\neg e$	b	0.0008	0.00009
	$\neg b$	0.00001	0.9971

$$P(a \wedge b \wedge \neg e) = \ 0.0008$$

$$P(a) = \boxed{\begin{array}{c} 0.0001 \\ + \\ 0.0009 \\ + \\ 0.0008 \\ + \\ 0.00001 \end{array}}$$

Joint Distribution

A giant table of probabilities

		a	$\neg a$
e	b	0.0001	0.00001
	$\neg b$	0.0009	0.00099
$\neg e$	b	0.0008	0.00009
	$\neg b$	0.00001	0.9971

$$P(a \land b \land \neg e) = \ 0.0008$$

$$P(a) = \ 0.00181$$

Joint Distribution

A giant table of probabilities

		a	$\neg a$
e	b	0.0001	0.00001
	$\neg b$	0.0009	0.00099
$\neg e$	b	0.0008	0.00009
	$\neg b$	0.00001	0.9971

$$P(a \wedge b \wedge \neg e) = \ 0.0008$$

$$P(a) = \ 0.00181$$

Problem: requires $2^k - 1$ entries where k is the number of variables

Conditional Probability

Want to know probability of an event A
given we know another event B happened

$P(A|B)$ read "probability of A given B"

Basic fact: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$

Conditional Probability

Want to know probability of an event A
given we know another event B happened

$P(A|B)$ read "probability of A given B"

Basic fact: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$

$P(A)$ is often called a prior
$P(A|B)$ often called posterior

Conditional Probability

Want to know probability of an event A
given we know another event B happened

$P(A|B)$ read "probability of A given B"

Basic fact: $P(A|B) = \frac{P(A \wedge B)}{P(B)}$

$P(A)$ is often called a prior
$P(A|B)$ often called posterior

Example:

$P($ rain tomorrow | rain today$)$

$P($ earthquake | alarm went off $)$

Conditional Independence

A and B are conditionally independent given C iff

$$P(A \wedge B|C) = P(A|C)P(B|C)$$

Equivalent to saying $P(A|B \wedge C) = P(A|C)$

In English means if we know about C then
knowing about B does not help us predict A

B contains no information about A
that we didn't know from C

Conditional Independence

A and B are conditionally independent given C iff

$$P(A \wedge B|C) = P(A|C)P(B|C)$$

Equivalent to saying $P(A|B \wedge C) = P(A|C)$

In English means if we know about C then
knowing about B does not help us predict A

B contains no information about A
that we didn't know from C

NOT the same as independence!

Height and reading ability are not independent

Height and reading ability are conditionally independent given age

Bayes Rule

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Bayes Rule

$P(A|B) = \frac{P(A \wedge B)}{P(B)}$

$\Rightarrow \; P(A|B)P(B) = P(A \wedge B) = P(B|A)P(A)$

Bayes Rule

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

$$\Rightarrow \quad P(A|B)P(B) = P(A \wedge B) = P(B|A)P(A)$$

$$\Rightarrow \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Rule

$P(A|B) = \frac{P(A \wedge B)}{P(B)}$

$$\Rightarrow \;\; P(A|B)P(B) = P(A \wedge B) = P(B|A)P(A)$$

$$\Rightarrow \;\; P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

So what?

Often allows us to transform into probabilities we know

If A is a disease and B is the symptoms then
want to know $P(A|B)$, but only know $P(B|A)$

Factoring a Joint Distribution

$$
\begin{aligned}
P(A \wedge B \wedge C \wedge D) &= P(A|B \wedge C \wedge D)P(B \wedge C \wedge D) \\
&= P(A|B \wedge C \wedge D)P(B|C \wedge D)P(C|D)P(D)
\end{aligned}
$$

If we use joint distribution tables for all of these then
$2^{k-1} \ldots + 4 + 2 + 1 = 2^k - 1$ different values to store

$8 + 4 + 2 + 1 = 15$ in above example

Factoring a Joint Distribution

$$P(A \wedge B \wedge C \wedge D) = P(A|B \wedge C \wedge D)P(B \wedge C \wedge D)$$

$$= P(A|B \wedge C \wedge D)P(B|C \wedge D)P(C|D)P(D)$$

If we use joint distribution tables for all of these then

$2^{k-1} \ldots + 4 + 2 + 1 = 2^k - 1$ different values to store

$8 + 4 + 2 + 1 = 15$ in above example

Idea: If possible use conditional independence!

$$P(A \wedge B \wedge C \wedge D) = P(A|B \wedge C)P(B|D)P(C|D)P(D)$$

Now only $4 + 2 + 2 + 1 = 9$ values

Factoring a Joint Distribution

$$
\begin{aligned}
P(A \wedge B \wedge C \wedge D) &= P(A|B \wedge C \wedge D)P(B \wedge C \wedge D) \\
&= P(A|B \wedge C \wedge D)P(B|C \wedge D)P(C|D)P(D)
\end{aligned}
$$

If we use joint distribution tables for all of these then
$2^{k-1} \ldots + 4 + 2 + 1 = 2^k - 1$ different values to store

$8 + 4 + 2 + 1 = 15$ in above example

Idea: If possible use conditional independence!

$$P(A \wedge B \wedge C \wedge D) = P(A|B \wedge C)P(B|D)P(C|D)P(D)$$

Now only $4 + 2 + 2 + 1 = 9$ values

Factoring order matters!

Improving the Example

$$P(E \wedge B \wedge A \wedge J \wedge M) = \\ P(J|A)P(M|A)P(A|B \wedge E)P(E)P(B)$$

If we know about the alarm, then the phone calls are independent of each other, the earthquake, and the burglary

Instead of 31 values we only need 10

Improving the Example

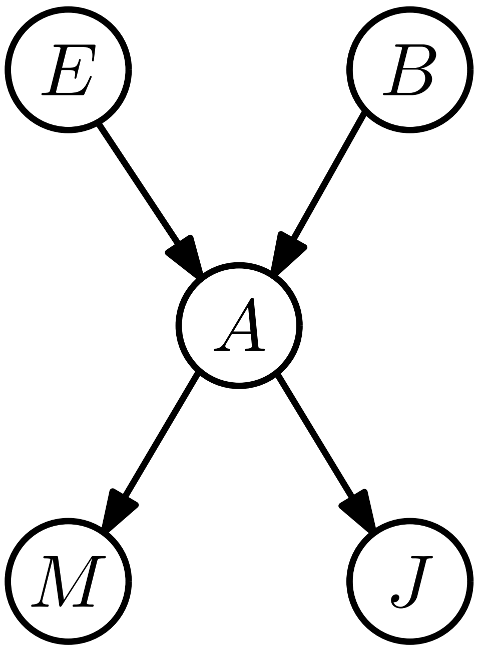
$$P(E \wedge B \wedge A \wedge J \wedge M) =$$
$$P(J|A)P(M|A)P(A|B \wedge E)P(E)P(B)$$

If we know about the alarm, then the phone calls are independent of each other, the earthquake, and the burglary

Instead of 31 values we only need 10

We just made our first Bayesian network!

Our First Bayesian Network



Node for each random variable

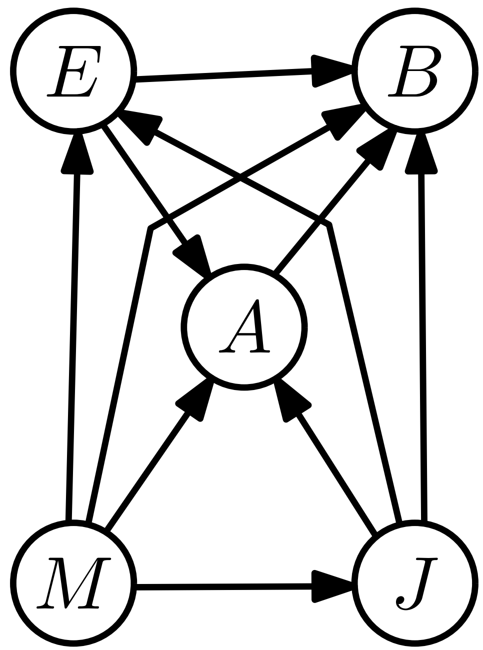
If X appears in the givens for $P(Y|\ldots)$
then draw arrow from X to Y

$$P(E \land B \land A \land J \land M) =$$
$$P(J|A)P(M|A)P(A|B \land E)P(E)P(B)$$

Can translate back and forth between graph and factorization

Our First Bayesian Network

What happens if we had factored differently?



$$P(A \wedge B \wedge E \wedge M \wedge J) =$$
$$P(B|A \wedge E \wedge M \wedge J)P(A|E \wedge M \wedge J)$$
$$P(E|J \wedge M)P(J|M)P(M)$$

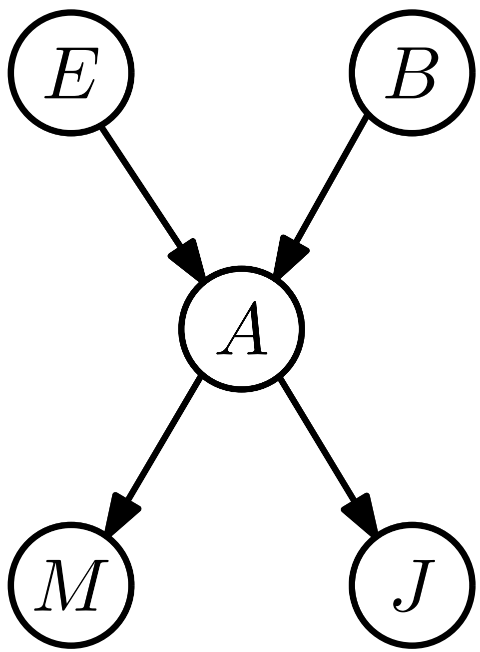
None of the conditional independence helps

Our First Bayesian Network

What do we need to store?

$P(E) = 0.002$

$P(B) = 0.001$



E	B	$P(A \mid B \wedge E)$
e	b	0.95
$\neg e$	b	0.94
e	$\neg b$	0.29
$\neg e$	$\neg b$	0.001

A	$P(M \mid A)$
a	0.9
$\neg a$	0.05

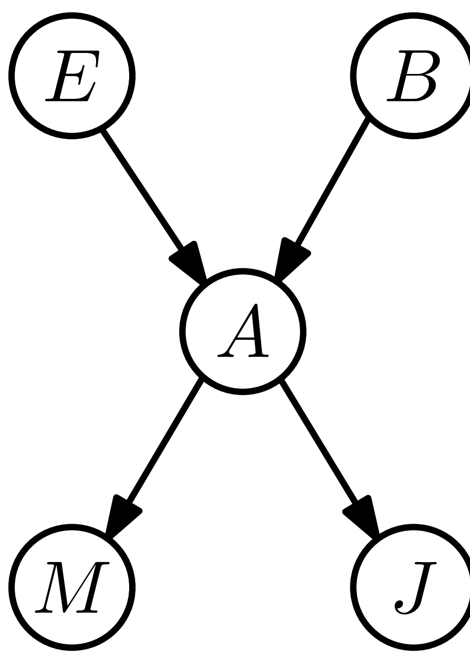
A	$P(J \mid A)$
a	0.7
$\neg a$	0.01

Our First Bayesian Network

$$P(a \wedge b \wedge \neg e \wedge m \wedge \neg j) = P(\neg j|a)P(m|a)P(a|b \wedge \neg e)P(\neg e)P(b)$$

$P(B) = 0.001$

$P(E) = 0.002$



E	B	$P(A|B \wedge E)$
e	b	0.95
$\neg e$	b	0.94
e	$\neg b$	0.29
$\neg e$	$\neg b$	0.001

A	$P(M|A)$
a	0.9
$\neg a$	0.05

A	$P(J|A)$
a	0.7
$\neg a$	0.01

Our First Bayesian Network

$$P(a \wedge b \wedge \neg e \wedge m \wedge \neg j) = P(\neg j|a)P(m|a)P(a|b \wedge \neg e)P(\neg e)P(b)$$

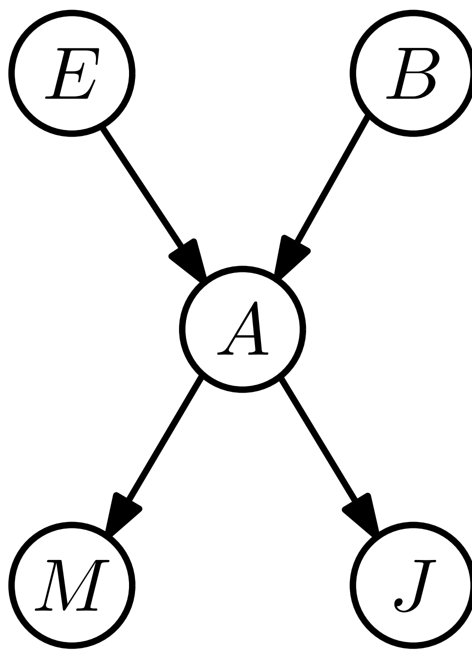
$$= \boxed{0.3} \times \boxed{0.9} \times \boxed{0.94} \times \boxed{0.998} \times \boxed{0.001}$$

$$= 0.0002532924$$

$P(B) = \boxed{0.001}$

E	B	$P(A \mid B \wedge E)$
e	b	0.95
$\neg e$	b	$\boxed{0.94}$
e	$\neg b$	0.29
$\neg e$	$\neg b$	0.001

$P(E) = \boxed{0.002}$



A	$P(M \mid A)$
a	$\boxed{0.9}$
$\neg a$	0.05

A	$P(J \mid A)$
a	$\boxed{0.7}$
$\neg a$	0.01

Why is this useful?/Decision Theory

We want to have agents make best decision given
the information they know

Suppose there are two tests for a disease

Test A works 100% of the time but costs $10 to administer

Test B works 90% of the time but costs only $5 to administer

Assume no false negatives only false positives

If Test B is positive we always need to run A to confirm

Why is this useful?/Decision Theory

We want to have agents make best decision given
the information they know

Suppose there are two tests for a disease

Test A works 100% of the time but costs \$10 to administer

Test B works 90% of the time but costs only \$5 to administer

Assume no false negatives only false positives

If Test B is positive we always need to run A to confirm

If you have the disease with probability p
then cost of running Test B is:
$p \times 15 + (1-p)(5 + .1 \times 10) = 6 + 9p$

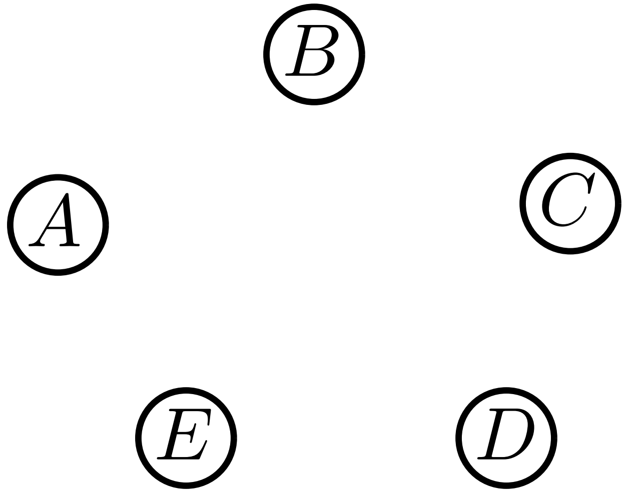
Need to compute $P($ disease $|$ symptoms $)$
to pick which test to run

Bayesian Networks

Alternate point of view:

Pick a factoring order for the variables

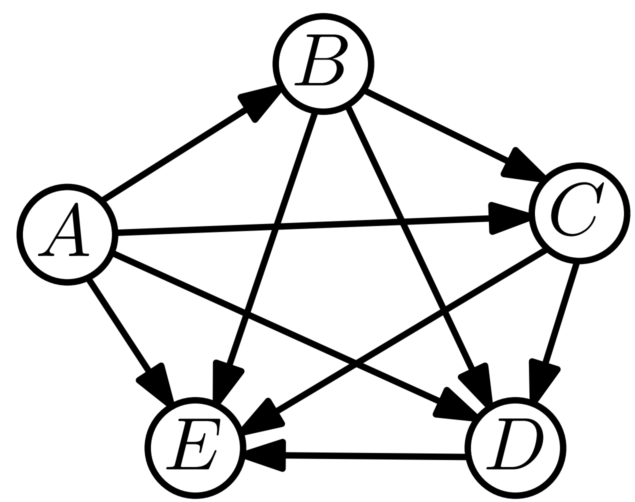
Create a node for each variable



Bayesian Networks

Alternate point of view:

- Pick a factoring order for the variables
- Create a node for each variable
- Add an edge from earlier variables to later variables

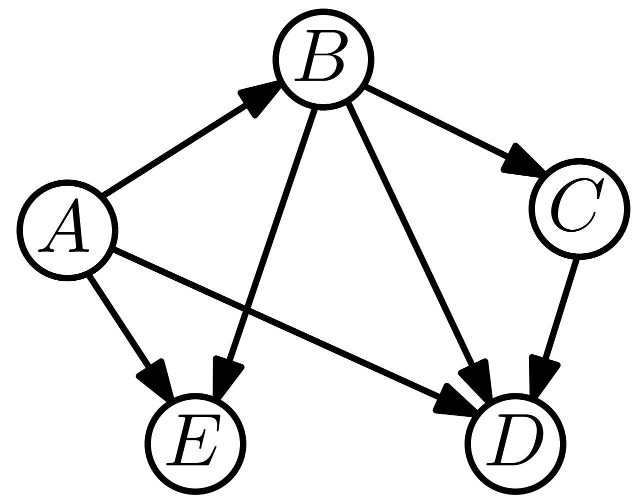


$$P(A \wedge B \wedge C \wedge D \wedge E) = \\ P(E|A \wedge B \wedge C \wedge D)P(D|A \wedge B \wedge C)P(C|B \wedge A)P(B|A)P(A)$$

Bayesian Networks

Alternate point of view:

- Pick a factoring order for the variables
- Create a node for each variable
- Add an edge from earlier variables to later variables
- Delete edges based on conditional independence



$$P(A \wedge B \wedge C \wedge D \wedge E) =$$
$$P(E|A \wedge B)P(D|A \wedge B \wedge C)P(C|B)P(B|A)P(A)$$

Bayesian Networks

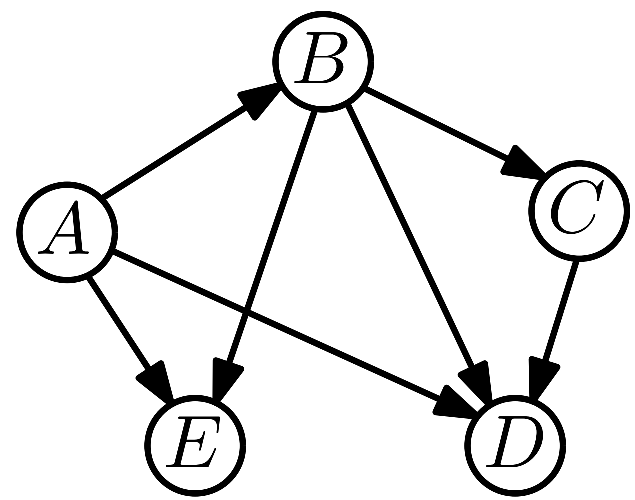
Alternate point of view:

Pick a factoring order for the variables

Create a node for each variable

Add an edge from earlier variables to later variables

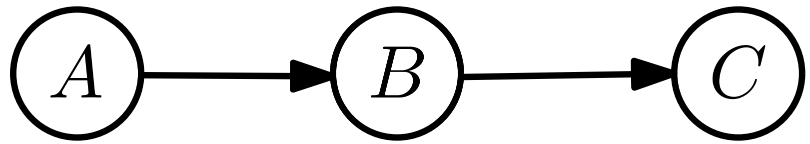
Delete edges based on conditional independence



$$P(A \wedge B \wedge C \wedge D \wedge E) =$$
$$P(E|A \wedge B)P(D|A \wedge B \wedge C)P(C|B)P(B|A)P(A)$$

Note: no cycles at third step so all Bayesian networks are acyclic

Simple Bayesian Networks



$$P(A \wedge B \wedge C) = P(C|B)P(B|A)P(A)$$

Known as Markov dependence

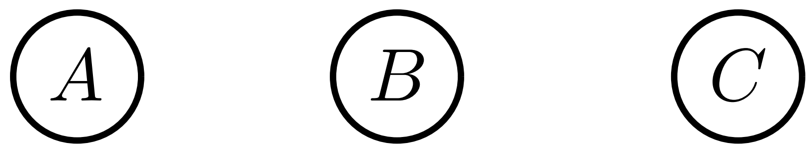
Example

A - did it rain yesterday?

B - is it raining today?

C - will it rain tomorrow?

Simple Bayesian Networks



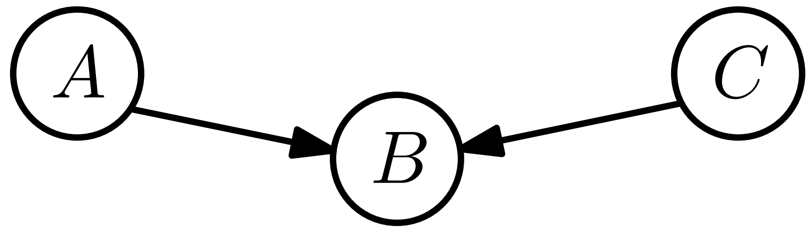
$$P(A \wedge B \wedge C) = P(C)P(B)P(A)$$

Marginal independence

Example

- 3 coin flips

Simple Bayesian Networks

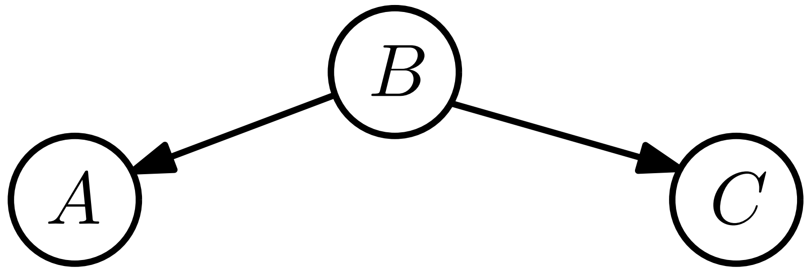


$$P(A \wedge B \wedge C) = P(B|A \wedge C)P(C)P(A)$$

Example

Earthquake, burglary, alarm

Simple Bayesian Networks



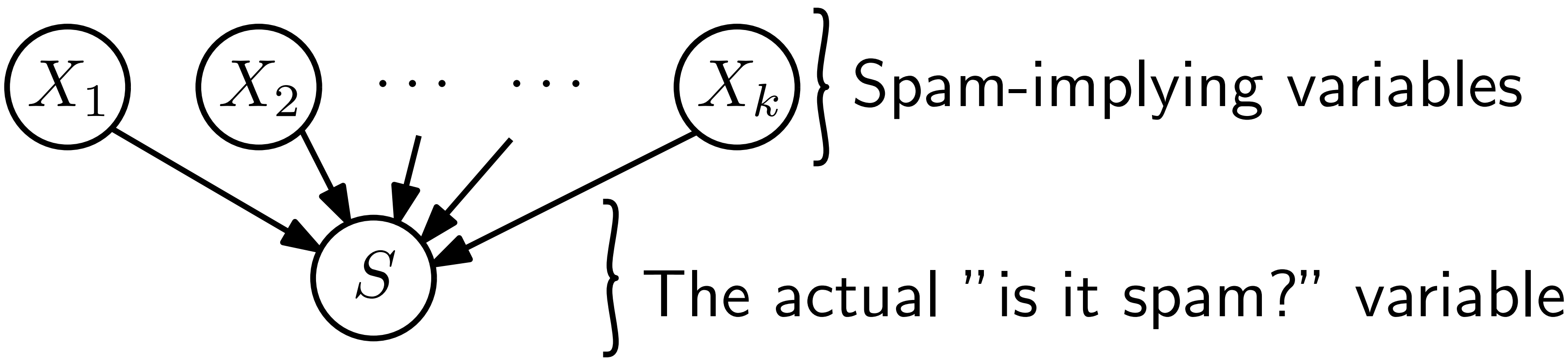
$$P(A \wedge B \wedge C) = P(A|B)P(C|B)P(B)$$

Example

Height, reading ability, age

Applications of Bayesian Networks

Spam filtering



$$P(S \wedge X_1 \wedge \ldots \wedge X_k) = P(X_1|S) \ldots P(X_k|S)P(S)$$

The spam-implying variables are conditionally independent once you know whether or not a message is spam

Conclusions

Logic has troubles with uncertainty

It is useful to represent and quantify uncertainty

In full generalization, probability is intractable

Conditional independence helps simplify the world

Bayesian networks are nice simple representations

- Encodes conditional probabilities in edges of a graph