Probabilistic Reasoning

Vijay Kumar Meena
Assistant Professor
KIIT University

Acknowledgement

- These slides are taken (as it is or with some modifications) from various resources including -
  - Artificial Intelligence: A Modern Approach (AIMA) by Russell and Norvig.
  - Slides of Prof. Mausam (IITD)
  - Slides of Prof. Rohan Paul (IITD)
  - Slides of Prof. Brian Yu (Harvard University)

Uncertainty in AI

- Till now, we looked at planning and search algorithms.
  - In Search: Effects of actions were deterministic.
  - In CSPs: Constraints were exact.
  - In Expectimax: notion of uncertainty in the agent's decision making.
- Uncertainty:
  - **Observed variables (evidence)**: Agent knows certain things about the state of the world (e.g., sensor measurements or symptoms)
  - **Unobserved variables**: Agent needs to reason about other aspects (e.g. what disease is present, is the car operational, location of the burglar)
  - **Model**: Agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge.

I hear an unusual sound and a burning smell in my car, what fault is there in my engine?

I have fever, loss of smell, loss of taste, do I have Covid?

I hear some footsteps in my house, where is the burglar?

Probability

- Like logical assertions, probabilistic assertions are about possible worlds. Whereas logical assertions say which possible worlds are strictly ruled out (all those in which the assertion is false), probabilistic assertions talk about how probable the various worlds are.
- In probability theory, the set of all possible worlds is called the **sample space**. **The possible worlds are mutually exclusive and exhaustive**—two possible worlds cannot both be the case, and one possible world must be the case.
- For example, if we are about to roll two (distinguishable) dice, there are 36 possible worlds to consider: (1,1), (1,2), ..., (6,6).
- **The Greek letter Ω (uppercase omega) is used to refer to the sample space, and ω (lowercase omega) refers to elements of the space, that is, particular possible worlds.**

Probability

- A fully specified probability model associates a numerical probability P(ω) with each possible world.
- The basic axioms of probability theory say that every possible world has a probability between 0 and 1 and that the total probability of the set of possible worlds is 1.

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1 .$$

For example, when rolling fair dice, we have P(Total = 11) = P((5, 6)) + P((6, 5)) = 1/36 + 1/36 = 1/18.

Conditional Probability

- Probabilities such as P(Total = 11) and P(doubles) are called unconditional or prior probabilities (and sometimes just “priors” for short); they refer to degrees of belief in propositions in the absence of any other information.
- Most of the time, however, we have some information, usually called evidence, that has already been revealed.
- For example, the first die may already be showing a 5 and we are waiting with bated breath for the other one to stop spinning. In that case, we are interested not in the unconditional probability of rolling doubles, but the conditional or posterior probability (or just “posterior” for short) of rolling doubles given that the first die is a 5. This probability is written **P(doubles | Die1 = 5), where the “ | ” is pronounced “given.”**

Conditional Probability

- Mathematically speaking, conditional probabilities are defined in terms of unconditional probabilities as follows: for any propositions a and b, we have

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)},$$

which holds whenever $P(b) > 0$. For example,

$$P(doubles \mid Die_1 = 5) = \frac{P(doubles \wedge Die_1 = 5)}{P(Die_1 = 5)}.$$

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Do I have Covid?
  - T = Engine is faulty or working?
  - D = How long will it take to drive to IIT?
  - L = Where is the person?

- Domains
  - R in {true, false} (often write as {+r, -r})
  - T in {faulty, working}
  - D in $[0, \infty)$
  - L in possible locations in a grid {(0,0), (0,1), …}

I hear an unusual sound and a burning smell in my car, what fault is there in my engine?

I have fever, loss of smell, loss of taste, do I have Covid?

I hear some footsteps in my house, where is the burglar?

Probability Distributions

- Probability distributions associate a probability with each value of the random variable

Temperature:

$$P(T)$$

T	P
hot	0.5
cold	0.5

Weather:

$$P(W)$$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Properties:

$$\forall x \;\; P(X = x) \geq 0$$

$$\sum_x P(X = x) = 1$$

Joint Distributions

- A **joint distribution** over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey:

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Note: Joint distribution can answer all probabilistic queries.
Problem: Table size is d^n.

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: likelihood of assignments (outcomes)
  - *Normalized:* sum to 1.0
- Constraint satisfaction problems:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Note: these constraints are strict.

Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Constraint over T,W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T

Events

- An event is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny? .4
  - Probability that it's hot? .4 + .1
  - Probability that it's hot OR sunny? .4 + .1 + .2

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Marginalization

- From a joint distribution (>1 variable) reduce it to a distribution over a smaller set of variables
- Called marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding likelihoods

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t,s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t,s)$$

$P(W)$

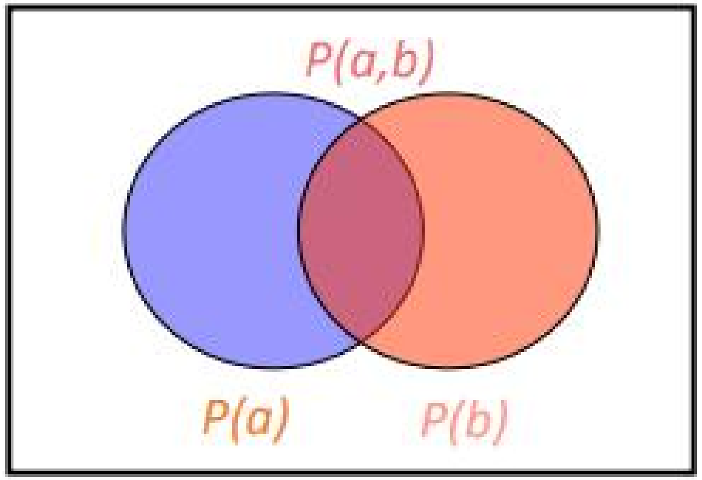
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- Conditional distributions are probability distributions over some variables given fixed values of others

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



Joint Distribution

$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Conditional Distributions

$$P(W|T)$$

$$P(W|T = cold)$$

W	P
sun	0.4
rain	0.6

$$P(W|T = hot)$$

W	P
sun	0.8
rain	0.2

Conditioning and Normalization

$$
\begin{aligned}
P(W=s|T=c) &= \frac{P(W=s,T=c)}{P(T=c)} \\
&= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\
&= \frac{0.2}{0.2+0.3} = 0.4
\end{aligned}
$$

$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the evidence →

$P(c,W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one) →

$P(W|T=c)$

W	P
sun	0.4
rain	0.6

$$
\begin{aligned}
P(W=r|T=c) &= \frac{P(W=r,T=c)}{P(T=c)} \\
&= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\
&= \frac{0.3}{0.2+0.3} = 0.6
\end{aligned}
$$

Probabilistic Inference

- Probabilistic inference
  - Compute a desired (unknown) probability from other known probabilities
  - E.g. conditional distribution from the joint distribution

- We generally compute conditional probabilities
  - P(on time | no reported accidents) = 0.90
  - These represent the agent's *beliefs* given the evidence

- Probabilities change with new evidence:
  - P(on time | no accidents, 5 a.m.) = 0.95
  - P(on time | no accidents, 5 a.m., raining) = 0.80
  - Observing new evidence causes *beliefs to be updated*

Inference by Enumeration

- P(W)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W)?

P(sun)=.3+.1+.1+.15=.65

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W)?

P(sun)=.3+.1+.1+.15=.65
P(rain)=1-.65=.35

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter, hot)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter, hot)?

P(sun|winter,hot)~.1
P(rain|winter,hot)~.05

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter, hot)?

P(sun|winter,hot)~.1
P(rain|winter,hot)~.05
P(sun|winter,hot)=2/3
P(rain|winter,hot)=1/3

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter)?

P(sun|winter)~.1+.15=.25

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter)?

P(rain|winter)~.05+.2=.25

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Inference by Enumeration

- P(W | winter)?

P(sun|winter)~.25
P(rain|winter)~.25
P(sun|winter)=.5
P(rain|winter)=.5

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Computing Conditional Probabilities

- Typically, we are interested in the *posterior joint distribution* of some *query variables* Y given specific values e for some *evidence variables* E
- Let the *hidden variables* be $Z = X - Y - E$
- If we have a joint probability distribution, we can compute the answer by "summing out" the hidden variables:

$$P(Y|e) \propto P(Y, e) = \sum_z P(Y, e, z) \quad \leftarrow \text{summing out}$$

Problem: Space and time complexity of inference by enumeration is $O(d^n)$.

Example: Inference by Enumeration

- Consider medical diagnosis, where there are 100 different symptoms and test results that the doctor could consider.
- A patient comes in complaining of fever, cough and chest pains.
- The doctor wants to compute the probability of pneumonia.
  - The probability table has $>= 2^{100}$ entries!
  - For computing the probability of a disease, we have to sum out over 97 hidden variables!

Product Rule

- Marginal and a conditional provides the joint distribution.

$$P(y)P(x|y) = P(x,y) \quad \Longleftrightarrow \quad P(x|y) = \frac{P(x,y)}{P(y)}$$

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Longleftrightarrow$

$P(D,W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

Chain Rule

Chain rule is derived by successive application of product rule:

$$
\begin{aligned}
& P(X_1, \ldots, X_n) = \\
= \quad & P(X_1, \ldots, X_{n-1}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & P(X_1, \ldots, X_{n-2}) P(X_{n-1} | X_1, \ldots, X_{n-2}) P(X_n | X_1, \ldots, X_{n-1}) \\
= \quad & \ldots \\
= \quad & \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})
\end{aligned}
$$

Constructing a larger distribution by simpler distribution.

Bayes Rule

- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Usefulness
  - Lets us build one conditional from its reverse.
  - Often one conditional is difficult to obtain but the other one is simple.

Example

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{Example givens}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

Independence

- Two variables are *independent* if:

$$\forall x, y : P(x, y) = P(x)P(y)$$

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form: $\forall x, y : P(x|y) = P(x)$
  - We write: $X \perp\!\!\!\perp Y$
  - Example
    - N-independent flips of a fair coin.

n smaller distributions

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

$\vdots$

$P(X_n)$

H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$

2^n

Conditional Independence

- Two variables X and Y are *conditionally independent* given Z if:

$$P(x|y,z) = P(x|z), \forall x, y, z$$

- This means that knowing the value of Y does not change the prediction about X *if the value of Z is known*.

Example

- Consider a patient with three random variables: B (patient has bronchitis), F (patient has fever), C (patient has a cough)
- The full joint distribution has $2^3 - 1 = 7$ independent entries
- If someone has bronchitis, we can assume that, the probability of a cough does *not* depend on whether they have a fever:

$$P(C|B,F) = P(C|B) \tag{1}$$

  I.e., C is *conditionally independent* of F given B
- The same independence holds if the patient does not have bronchitis:

$$P(C|\neg B,F) = P(C|\neg B) \tag{2}$$

  therefore C and F are conditionally independent given B

Example (contd.)

- The full joint distribution can now be written as:

$$
\begin{aligned}
& P(C, F, B) = \\
= \quad & P(C, F|B)P(B) \\
= \quad & P(C|B)P(F|B)P(B)
\end{aligned}
$$

Naïve Bayesian Model

- A common assumption in early diagnosis is that the symptoms are independent of each other given the disease
- Let $s_1, \ldots s_n$ be the symptoms exhibited by a patient (e.g. fever, headache etc)
- Let D be the patient's disease
- Using the naive Bayes assumption:

$$P(D, s_1, \ldots s_n) = P(D)P(s_1|D) \cdots P(s_n|D)$$

Recursive Bayesian Updating

The naive Bayes assumption allows incremental updating of beliefs as more evidence is gathered.

- Suppose that after knowing symptoms $s_1, \ldots s_n$ the probability of D is: $P(D|s_1 \ldots s_n) = P(D)\prod_{i=1}^{n} P(s_i|D)$
- What happens if a new symptom s_{n+1} appears?

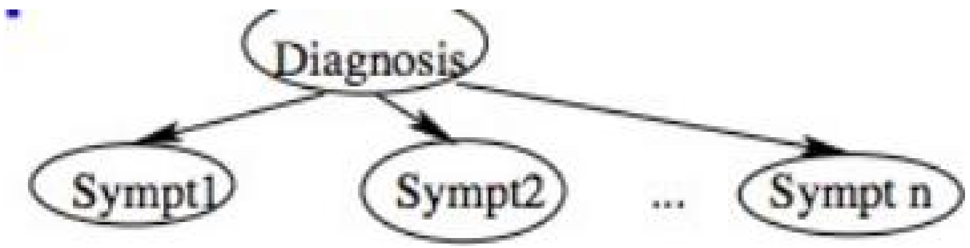
$$
\begin{aligned}
P(D|s_1 \ldots s_n, s_{n+1}) &= P(D)\prod_{i=1}^{n+1} P(s_i|D) \\
&= P(D|s_1 \ldots s_n)P(s_{n+1}|D)
\end{aligned}
$$

- An even nicer formula can be obtained by taking logs:

$$\log P(D|s_1 \ldots s_n, s_{n+1}) = \log P(D|s_1 \ldots s_n) + \log P(s_{n+1}|D)$$

Naïve Bayes: Graphical Representation

- The nodes represent random variables.
- The arcs represent "influences".
- The lack of arcs represents conditional independence.
- Linear number of parameters.
- **Can this be done on general graphs?**



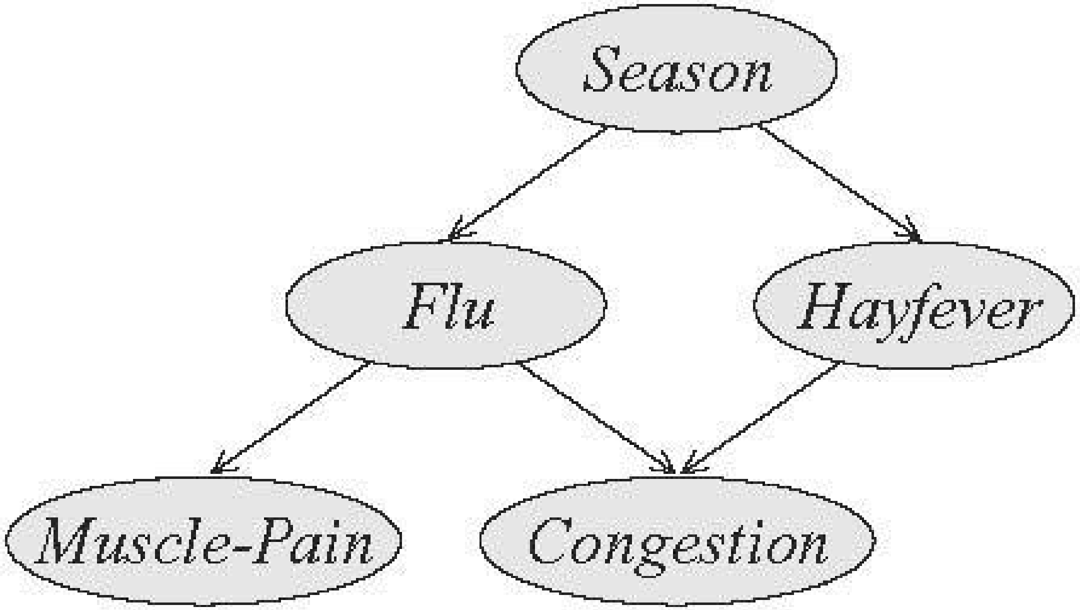
$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\Pi_i \mathbf{P}(Effect_i|Cause)$$

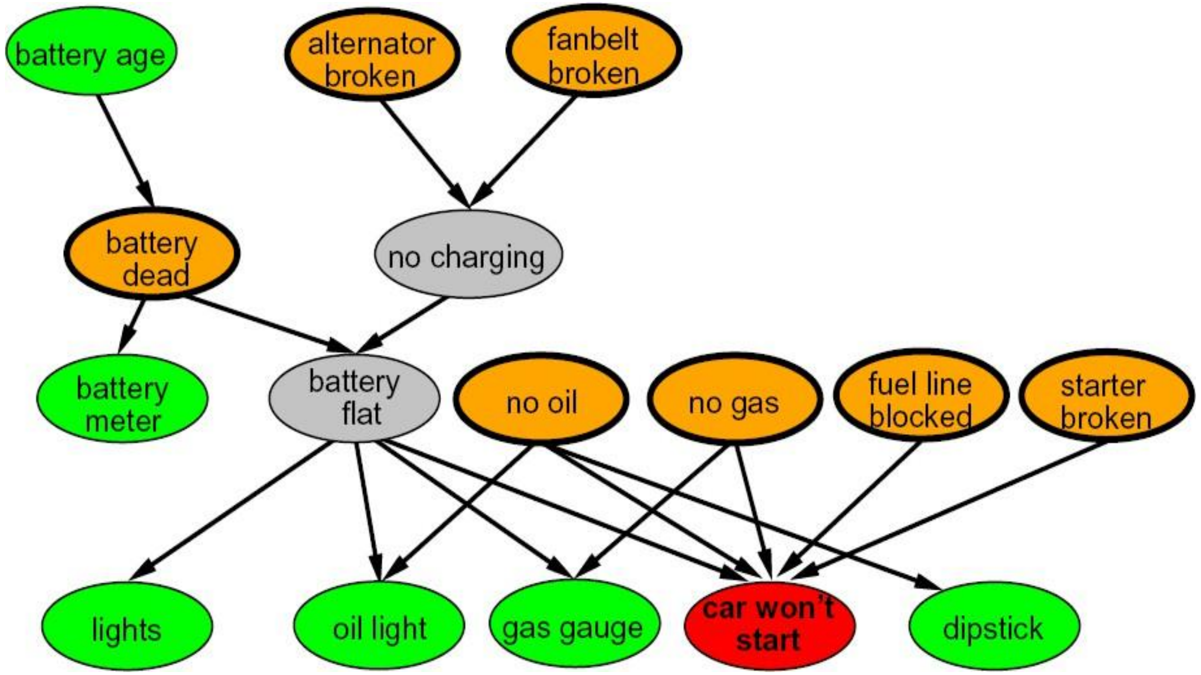
Bayesian Networks

- Problem with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is hard to represent explicitly.

- Bayesian Networks:
  - A technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
    - Also known as probabilistic graphical models
  - Encode how variables locally influence each other. Local interactions chain together to give global, indirect interactions

Examples


Season
Flu
Hayfever
Muscle-Pain
Congestion


battery age
alternator broken
fanbelt broken
battery dead
no charging
battery meter
battery flat
no oil
no gas
fuel line blocked
starter broken
lights
oil light
gas gauge
car won't start
dipstick

Bayesian Networks: Semantics

- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

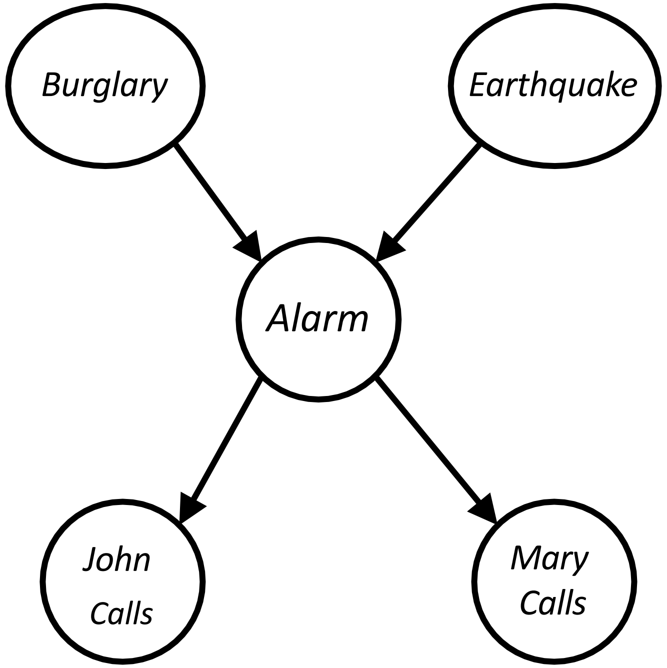
$$P(X|a_1 \ldots a_n)$$

- Bayesian Networks implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals:

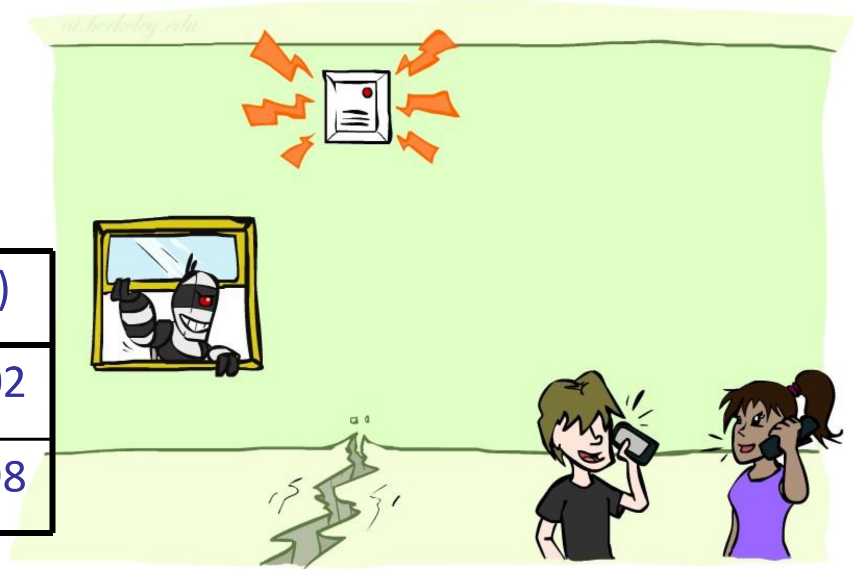
$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i))$$

Example: The Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

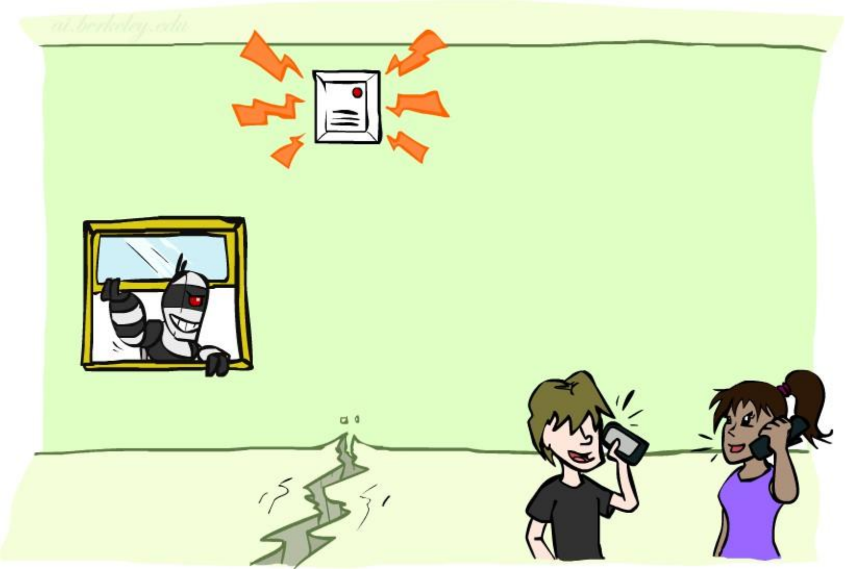


A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

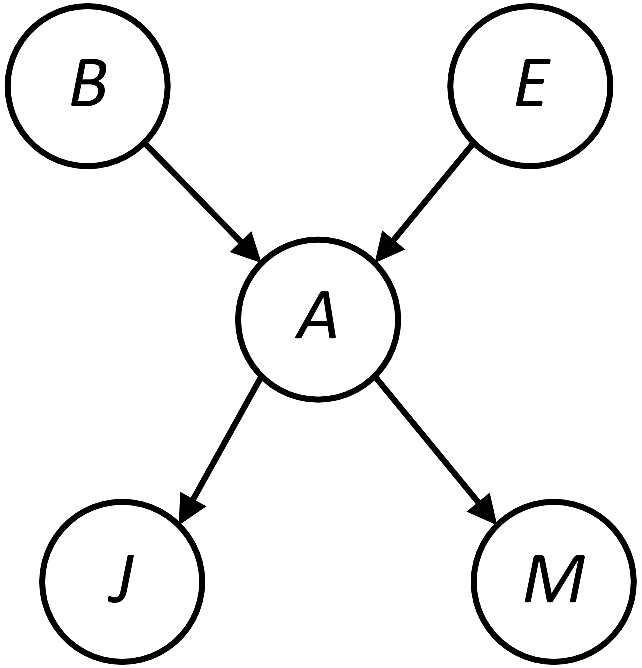
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

Example: The Alarm Network



B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

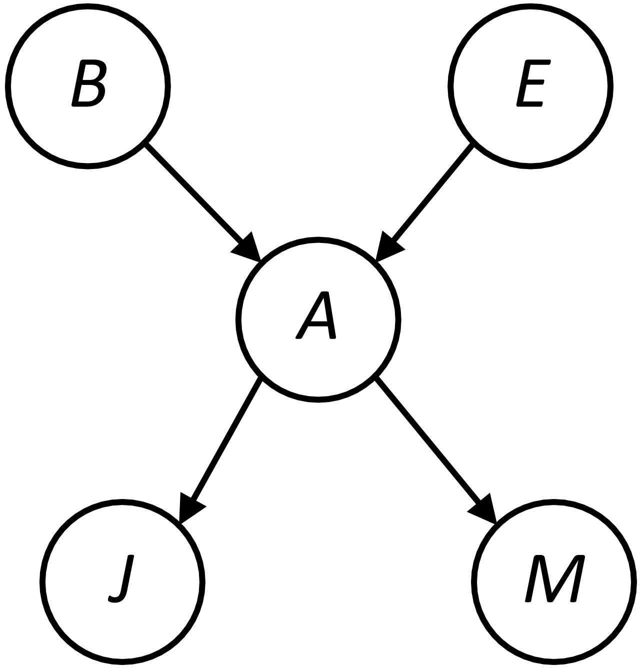
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$

Example: The Alarm Network

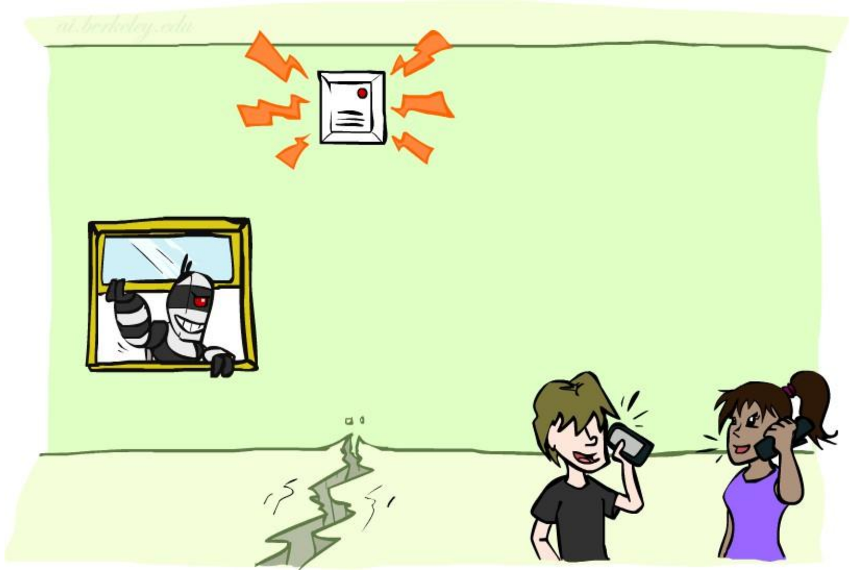
B	P(B)
+b	**0.001**
-b	0.999



E	P(E)
+e	0.002
-e	**0.998**

A	J	P(J|A)
+a	+j	0.9
+a	**-j**	**0.1**
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	**+m**	**0.7**
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99



B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	**-e**	**+a**	**0.94**
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b,-e,+a,-j,+m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Bayesian Network

- A Bayesian network is a directed graph in which each node is annotated with quantitative probability information. The full specification is as follows:
  - Each node corresponds to a random variable, which may be discrete or continuous.
  - A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y , X is said to be a parent of Y. The graph has no directed cycles and hence is a directed acyclic graph, or DAG.
  - Each node X_i has a conditional probability distribution $P(X_i \mid P\ arents(X_i))$that quantifies the effect of the parents on the node.
- The intuitive meaning of an arrow is typically that X has a direct influence on Y, which suggests that causes should be parents of effects.

Example

- You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.
- You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm.
- John nearly always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
- Mary, on the other hand, likes rather loud music and often misses the alarm altogether.
- Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Example

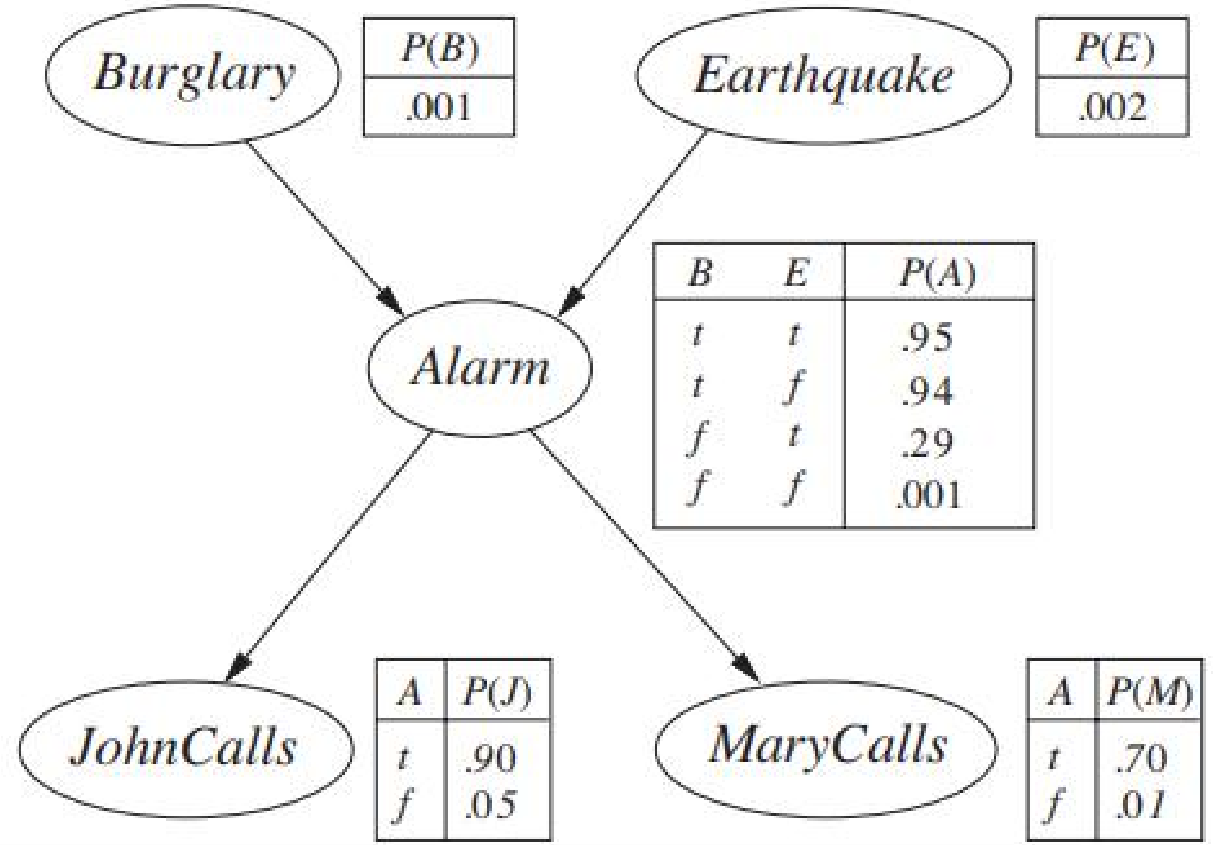


Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

Example

- Calculate the probability that the alarm has sounded, but neither a burglary nor an earthquake has occured, and both John and Mary Call.
- $P(j, m, a, \neg b, \neg e)$
- $P(j, m, a, \neg b, \neg e) = P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b \wedge \neg e)\, P(\neg b)\, P(\neg e)$
- $P(j, m, a, \neg b, \neg e) = 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.000628$