Uncertainly and Probabilistic Reasoning

Overview

- Uncertainty
- Decision theory example
- Probability basics
- Conditional probability
- Axioms of probability
- Joint probability distribution
- Bayes rule
- Bayes rule: Example

Uncertainty

- Problem with first-order logic: agents almost never have full access to the whole truth about their environment.
- Therefore, the agent must act under **uncertainty**.
- Uncertainty can also arise because of incompleteness and incorrectness in the agent's understanding of the properties in the environment.
- Incomplete, because there are too many conditions to explicitly enumerate.

There are trade-offs (playing safe can result in other annoyances), thus the right thing to do depends on both the relative importance of various goals and the likelihood (and degree to which) they will be achieved.

Example: Trying to Catch a Flight

A_t: plan to leave home t minutes before the flight departure time.

- The traveler needs to make a decision in an uncertain environment: car can break down, traffic can be extremely congested, natural disaster, etc.
- Such worst-case scenarios are hard to explicitly enumerate: the list goes on – ran out of gas, spouse/children in an emergency, flight crews goes on a strike, etc. etc.
- Thus the traveler only has an incomplete understanding of the situation.
- The traveler can play safe by going with plan A_{1440}, but this can cause the traveler to wait for a long time at the airport before departure.

Difficulties in Applying F-O-L in Uncertain Domains

For example, application of first-order logic in medical diagnosis domain can fail because of these reasons:

- Laziness: cannot list the complete set of antecedents and consequents needed to ensure an exceptionless rule, and too hard to use the enormous rules that result.
- Theoretical ignorance: medical science has no complete theory.
- Practical ignorance: even though we have all the rules, it is practically impossible to run all the tests.

Similar situation arises in law, business, dating, etc. The agent's knowledge can at best provide only a **degree of belief**. **Probability theory** is well suited for such a domain.

Acquisition of New Information and Probability

- The degree of belief changes as an agent perceives or acquires new information from the world: we call this the **evidence**.
- This is analogous to saying whether or not a given logical sentence is entailed by (i.e. is a logical consequence of) the knowledge base, because the truth value can change when new facts are added to the KB.
- Before the evidence is received, we talk about **prior** or **unconditional** probability.
- After the evidence is obtained, we talk about **posterior** or **conditional** probability.

Example

When playing black jack,

- as new cards are drawn and shown, your degree of belief in the fact that you need more cards can change.

What about poker? or slot machine?

Rational Decisions Under Uncertainty: Decision Theory

- There are trade-offs, and an agent must first have **preferences** between different results when a certain plan was executed.
- **Utility theory** deals with such preferences: how useful is such and such result to the agent?
- **Decision theory** is a general theory of rational decision under uncertainty, combining **probability theory** and **utility theory**.

Decision Theory

- An agent is rational iff it chooses the action that yields the highest expected utility, averaged over all possible outcomes of the action: **Principle of Maximum Expected Utility**
- Example: backgammon (discussed earlier) – min-max trees with probabilistic levels.

Decision Theoretic Agent

function DT-Agent (*percept*) **returns** *action*
static: a set probabilistic *belief* about the state of the world

calculate updated probabilities for current state based on *percept* and past actions

calculate outcome probabilities for actions, given action descriptions and prob of current states.

select *action* with highest expected utility given prob of outcomes and utility information.
return *action*

Decision Theory: Example

Decision theory = Probability theory + Utility theory

	Utility of Resulting State	Probability
Action 1	10	0.2
Action 2	10000	0.001
Action 3	5	0.799

Which action would an optimal Decision Theoretic Agent take?

Decision Theory: Example

Decision theory = Probability theory + Utility theory

	Utility of Resulting State $\times$ Probability	Expected Utility
Action 1	10×0.2	2
Action 2	1000×0.001	1
Action 3	5×0.799	$3.995 \longleftarrow$

Action 3 has the maximum expected utility, thus action 3 will be carried out.

Probability: Notations[a]

- **Random variable**: variable that can take on different values
  - $A, B, ...$: boolean values ($\mathbf{T}$ or $\mathbf{F}$).
  - $X, Y, ...$: numerical values or other multi-valued enumerations (1, 2, 0.5, Cloudy, Rainy, Sunny, ...)
- $P(X = v)$: **probability** of the variable X having value v.
  - This can be viewed as an *event*.
  - For boolean variables, $P(A)$ means $P(A = \mathbf{T})$, and $P(\neg A)$ means $P(A = \mathbf{F})$.
- $\mathbf{P}(X)$: **probability distribution**, a full list of probabilities for all possible values that X can take (note that $\mathbf{P}$ is in **bold**.

[a] All conventions follow Russel & Norvig

Examples

- Boolean:
  $P(Infected) = 0.01, P(\neg Infected) = 0.99.$
- Multi valued:
  $P(Dice = 1) = \frac{1}{6}, P(Dice = 2) = \frac{1}{6}, ...$
- Multi valued:
  $P(Weather = Sunny) = 0.7,$
  $P(Weather = Rainy) = 0.2, ...$

Logical Connectives and Conditional Probability

- Logical connectives can be used:

  $$P(A \vee B), P(A \wedge \neg B), P(Cavity \wedge \neg Insured), \text{ etc.}$$

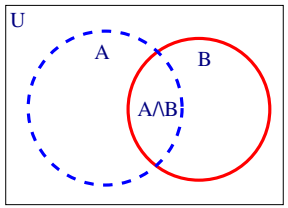
- Conditional Probability $P(A|B)$ (read *probability of* A *given* B):

  $$P(Cavity|Toothache) = 0.8$$

- As new evidence comes in, the conditional probability gets updated:

  $$P(Cavity|\underbrace{Toothache \wedge BadBreath})$$

Conditional Probability



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

- Think about the **area** occupied by each event.
- The bounding rectangle U has an area of 1, thus

  $$P(A) = \frac{\text{Area of } A}{\text{Area of } U} = \frac{\text{Area of } A}{1} = \text{Area of } A$$

- $P(A|B)$ means B now takes on the role of U. Within this limited event space, what is the probability of A.

The Axioms of Probability

All axioms

1. All probabilities are between 0 and 1
$$0 \leq P(A) \leq 1$$
2. For a valid proposition A ($\mathbf{T}$ under all interpretations): $P(A) = 1$, and for a inconsistent proposition A ($\mathbf{F}$ under all interpretations): $P(A) = 0$.
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

Other properties follow from these three axioms.

Other Properties

- From the axioms,
$$\begin{aligned} P(A \vee \neg A) &= P(A) + P(\neg A) - P(A \wedge \neg A) \\ P(\mathbf{T}) &= P(A) + P(\neg A) - P(\mathbf{F}) \\ 1 &= P(A) + P(\neg A) \\ P(\neg A) &= 1 - P(A) \end{aligned}$$

- More generally, the **sum** of probabilities $P(X = v)$ is 1, for all values v the random variable X can take:
$$\left[\sum_{v \in V} P(X = v)\right] = 1,$$
where V is the set of all possible values X can take.

Joint Probability Distribution

For random variables $X_1, X_2, ..., X_n$,

- An **atomic event** is an assignment of particular values to each random variable.
- The **joint probability distribution** $\mathbf{P}(X_1, X_2, ..., X_n)$ completely specifies the probabilities of all **atomic events**.
- Thus,
$$\left[\sum_{(v_1, v_2, ..., v_n) \in \mathbf{V}} P(X_1 = v_1, X_2 = v_2, ..., X_n = v_n)\right] = 1,$$
where $\mathbf{V}$ is a set of all possible $n-$vectors that the vector $(X_1, X_2, .., X_n)$ can assume..

Joint Probability Distribution: Example

	Toothache	¬Toothache	Sum
Cavity	0.04	0.06	$P(C) = 0.1$
¬Cavity	0.01	0.89	$P(\neg C) = 0.9$
Sum	$P(T) = 0.05$	$P(\neg T) = 0.95$	$\sum = 1.0$

Abbreviations: $C = Cavity, T = Toothache$

- $P(C \vee T) = P(C) + P(T) - P(C \wedge T)$
$= 0.1 + 0.05 - 0.04 = 0.11$
- $P(C|T) = \frac{P(C \wedge T)}{P(T)} = \frac{0.04}{0.05} = 0.8$
- $P(T|C) = \frac{P(C \wedge T)}{P(C)} = \frac{0.04}{0.1} = 0.5$

In practice, writing a full joint probability table like this is impossible (or too much effort): for n boolean random variables, you need 2^n entries.

Bayes' Rule

- From $P(A|B) = \frac{P(A \wedge B)}{P(B)}$ and $P(B|A) = \frac{P(A \wedge B)}{P(A)}$, we get

$$P(A|B)P(B) = P(B|A)P(A)$$

and in turn from which we get the **Bayes' Rule**:

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

Extended Bayes' Rule

$$\mathbf{P}(Y|X,E) = \frac{\mathbf{P}(X|Y,E)\mathbf{P}(Y|E)}{\mathbf{P}(X|E)}$$

This rule follows from

$\mathbf{P}(A,B|E) = \mathbf{P}(A|B,E)\mathbf{P}(B|E)$:

$$\begin{aligned}\mathbf{P}(A,B|E) &= \frac{\mathbf{P}(A,B,E)}{\mathbf{P}(E)} \\ &= \frac{\mathbf{P}(A|B,E)\mathbf{P}(B,E)}{\mathbf{P}(E)} \\ &= \mathbf{P}(A|B,E)\mathbf{P}(B|E)\end{aligned}$$

Note: $\mathbf{P}(Y|X,E) = \mathbf{P}(Y|\underbrace{(X,E)}).$

Exercise: text book, exercise 14.5b and 14.6 (p. 434).

Example: Application of Bayes Rule

Exercise 14.3 (p. 433): After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given you don't have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

T: tested positive, $\neg T$: tested negative, D : have disease, $\neg D$: clean.

Solution: Good News and Bad News

These are given:

$$\begin{aligned}P(T|D) &= 0.99 \\ P(\neg T|\neg D) &= 0.99 \\ P(D) &= \frac{1}{10,000} = 0.0001\end{aligned}$$

We want to calculate the probability that you have the disease given a positive test result:

$$P(D|T)$$

We can use Bayes' rule to derive this probability.

Solution: Good News and Bad News (cont'd)

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)}$$

- $P(T|D) = 0.99$, $P(\neg T|\neg D)$, and $P(D) = 0.0001$ are given.
- From these, we can get $P(\neg T|D) = 0.01$, $P(T|\neg D) = 0.01$, and $P(\neg D) = 0.9999$.

Since $P(T|D)$ and $P(D)$ are give, we only need to calculate $P(T)$.

Solution: Good News and Bad News (cont'd)

Observation [a]: $P(T) = P(T|D)P(D) + P(T|\neg D)P(\neg D)$.

Thus, $P(T) = 0.99 \times 0.0001 + 0.01 \times 0.9999 = 0.010098$, and with this,

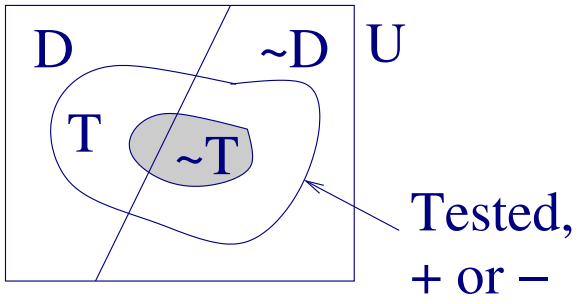
$$P(D|T) = \frac{0.99 \times 0.0001}{0.010098} = 0.0098,$$

which is slightly less than 1%.

Exercise: how accurate should the test be so that $P(D|T)$ is greater than 0.95 (i.e. 95%)?

[a] $P(T) = P(T \wedge D) + P(T \wedge \neg D)$.

Calculating $P(T)$ given $P(T|D)$ and $P(D)$



$$\begin{aligned} P(T) &= P(T \wedge D) + P(T \wedge \neg D) \\ &= P(T|D)P(D) + P(T|\neg D)P(\neg D) \end{aligned}$$

- $\{D\} \cup \{\neg D\}$ completely account for the whole population, but $\{T\} \cup \{\neg T\}$ does not cover the whole population (because you **did not test everyone!**).

Calculating $P(T)$ given $P(T|D)$ and $P(D)$

Another way of deriving $P(T) = P(T|D)P(D) + P(T|\neg D)P(\neg D)$:

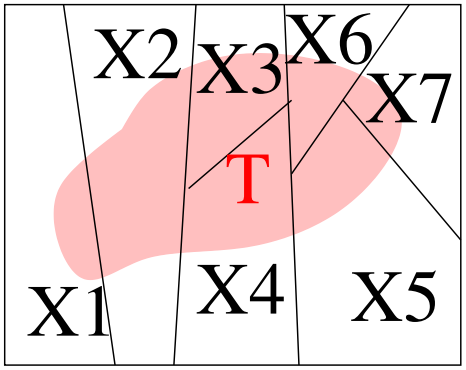
From

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\ P(\neg D|T) &= \frac{P(T|\neg D)P(\neg D)}{P(T)} \end{aligned}$$

and from $P(D|T) + P(\neg D|T) = 1$,

$$\begin{aligned} 1 &= \frac{P(T|D)P(D)}{P(T)} + \frac{P(T|\neg D)P(\neg D)}{P(T)}, \text{ thus} \\ P(T) &= P(T|D)P(D) + P(T|\neg D)P(\neg D) \end{aligned}$$

Calculating $P(T)$: General Case



More Generally, if $\left[\sum_{x\in\{x_1,x_2,\ldots,x_n\}} P(X=x)\right] = 1$ and events $X = x_m$ and $X = x_n$ are disjoint for all $m \neq n$,

$$P(T) = \sum_{x\in\{x_1,x_2,\ldots,x_n\}} P(T|\underbrace{X=x}) P(X=x)$$

What's The Big Deal?

- $P(T|D)$ may be easier to obtain: you can run the test on a small pool of known patients (say 100) at a hospital.
- $P(D|T)$ is much harder to obtain directly. Since the test makes 1 mistake out of 100 tests, if you run the test on 10,000 people, you'll get 100 false-positives, and one genuine patient who tests positive (consider that $P(T) = 0.010098$). So, just to get about 100 people testing positive, you have to run the tests on 10,000 people.
- $P(D)$ serves as a **prior** in this case. In many cases, the prior represents subjective **belief** of the person calculating the probability in case $P(D)$ is not directly measurable.

Key Points

- Application of theorem proving: question answering
- Uncertainty
- Decision theory example: how prob theory and decision theory are combined
- Probability basics: terminology, notations.
- Joint probability distribution: concept
- Conditional probability: definition, various ways of representing conditional prob.
- Axioms of probability: basic axioms, and using them to prove simple equalities.
- Bayes rule: definition and application.

Overview

- Diagnostic vs. causal knowledge
- Calculating $P(T)$ given $P(T|D)$ and $P(D)$
- Ratios of conditional probabilities and causes of an phenomenon
- Example: object recognition
- Bayesian updating

Diagnostic vs. Causal Knowledge

Consider these probabilities:

- $P(Symptom|Disease)$: **causal knowledge**
  - relatively fixed.
- $P(Disease)$: somewhat variable.
- $P(Disease|Symptom)$: **diagnostic knowledge**
  - fluctuates as $P(Disease)$ change.

$P(Disease|Symptom)$ directly measured can be no longer accurate when $P(Disease)$ changes (e.g. an epidemic outburst), however the calculation based on Bayes' rule can be much more robust.

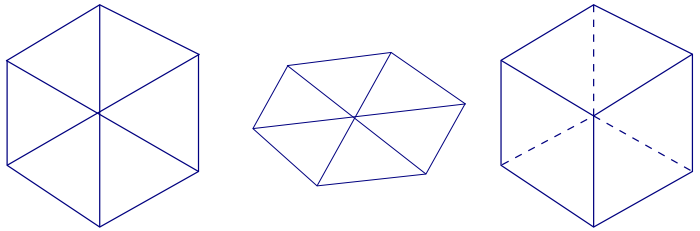
Comparison of Conditional Probabilities

When C_1 or C_2 can cause phenomenon (or *effect*) E, to find out the which is the more probable cause of phenomenon E, we do not need to explicitly calculate $P(E)$:

- $P(C_1|E) = \frac{P(E|C_1)P(C_1)}{P(E)}$
- $P(C_2|E) = \frac{P(E|C_2)P(C_2)}{P(E)}$
- From the above, we get:

$$\frac{P(C_1|E)}{P(C_2|E)} = \frac{P(E|C_1)P(C_1)}{P(E|C_2)P(C_2)} = \frac{a}{b}$$

Example: The Problem of Object Recognition

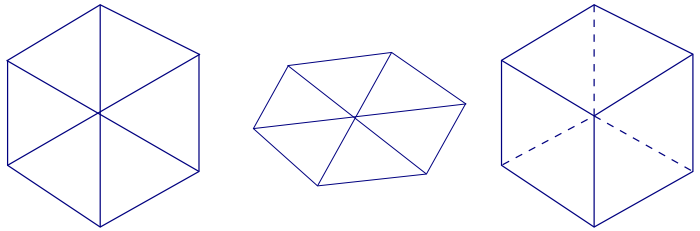


Given an image projected on the retina, what is the more likely cause? the 2D hexagon? or a transparent 3D cube? This is basically a computer vision problem.

$$\frac{P(Hexagon|Image)}{P(Cube|Image)} = \frac{P(Image|Hexagon)P(Hexagon)}{P(Image|Cube)P(Cube)} = \frac{a}{b}$$

A probabilistic vision agent can make a decision based on such a ratio.

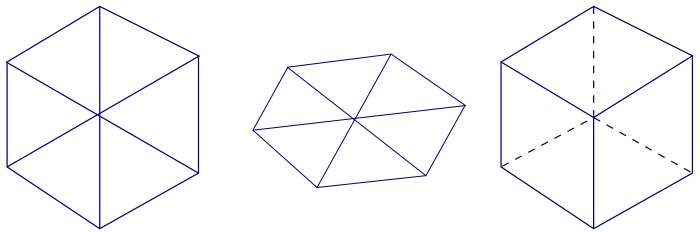
Example: Object Recognition (cont'd)



$$\frac{P(Hexagon|Image)}{P(Cube|Image)} = \frac{P(Image|Hexagon)P(Hexagon)}{P(Image|Cube)P(Cube)} = \frac{a}{b}$$

- Decision: if $a/b > 1$, it is most likely that a hexagon generated the image. If $a/b < 1$, it is most likely that a cube generated the image.

Example: Object Recognition (cont'd)



$$\frac{P(Hexagon|Image)}{P(Cube|Image)} = \frac{P(Image|Hexagon)P(Hexagon)}{P(Image|Cube)P(Cube)} = \frac{a}{b}$$

- Why is $P(Image|Hexagon)$ easier to calculate than $P(Hexagon|Image)$?
- What about $P(Hexagon)$ and $P(Cube)$?

Combining Multiple Evidences

Suppose we have these conditional probabilities $P(Cavity|Toothache)$ and $P(Cavity|Catch)$. What if we want to know $P(Cavity|\underbrace{Toothache \wedge Catch})$? These are the alternatives:

- Look up the joint probability table: not practical or even impossible in most cases
- We can calculate

  $$P(Cav|Ache \wedge Catch) = \frac{P(Ache \wedge Catch|Cav)P(Cav)}{P(Ache \wedge Catch)}$$

  but, calculating the new conditional prob and the normalization factor is a pain.

Bayesian Updating

An Alternative: gradually work in the multiple evidences – **Bayesian Updating**

- Reformulate the Bayes' rule so that conditional probability of events given combined evidences (such as $P(A|B \wedge C)$) are not necessary.
- Use domain knowledge to replace the more complex conditional probabilities with known, simpler ones (utilize **conditional independence**).

Bayesian updating makes combining evidences efficient (more detail next time).

Bayesian Updating: Example

We want to calculate $P(Cavity|Ache \wedge Catch)$:

$$P(Cav|Ache \wedge Catch) = P(Cav|Ache)\frac{P(Catch|Ache \wedge Cav)}{P(Catch|Ache)}$$
$$= \underbrace{P(Cav)\frac{P(Ache|Cav)}{P(Ache)}}_{old}\underbrace{\frac{P(Catch|Ache \wedge Cav)}{P(Catch|Ache)}}_{new}$$

Problem is that $P(Catch|Ache \wedge Cav)$ may be equally hard to calculate. However, we can make these assumptions (**Conditional Independence** of $Ache$ and $Catch$ given Cav):

$$P(Catch|Cav \wedge Ache) = P(Catch|Cav)$$
$$P(Ache|Cav \wedge Catch) = P(Ache|Cav)$$

Only thing that remains is $P(Ache)P(Catch|Ache) = P(Catch \wedge Ache)$, which can be eliminated by normalization (**Exercise:** try this – see Exercise 14.7).

Bayesian Updating: Example (cont'd)

So, after replacing the factors:

$$P(Cav|Ache \wedge Catch) = \alpha \underbrace{P(Cav)P(Ache|Cav)}_{old} \underbrace{P(Catch|Cav)}_{new},$$

where α is the normalization constant needed to make $P(Cav|Ache \wedge Catch)$ add up to 1.

Bayesian Updating: Summary

X and Y are independent given Z (conditional independence):

$$\mathbf{P}(X|Y,Z) = \mathbf{P}(X|Z)$$

Simplified Bayes' rule for multiple evidence is[a]:

$$\mathbf{P}(Z|X,Y) = \alpha \mathbf{P}(Z)\mathbf{P}(X|Z)\mathbf{P}(Y|Z),$$

where α is the normalization constant.

Thus, Bayesian Updating makes combining multiple evidence easy.

[a] **Note:** Z is the cause, and X and Y are the effects.

Key Points

- How is subjective belief utilized in Bayesian analysis?
- Bayesian updating: why does it make probabilistic inference efficient when multiple evidence comes in?

Overview

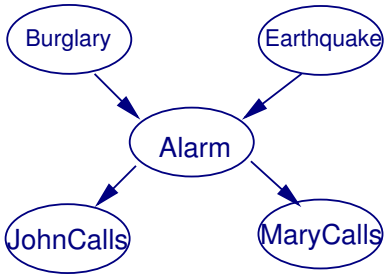
- Belief network

Probabilistic Reasoning

Belief Network represent the dependence between random variables, and give a concise specification of the joint probability distribution. It is represented as a directed acyclic graph (DAG):

1. a set of random variables : nodes of the network
2. a set of directed edges from one node to another
3. each node has a conditional probability table that quantifies the effect the parents have on that node. The parents are the nodes pointing to that node.
4. the graph has no cycles

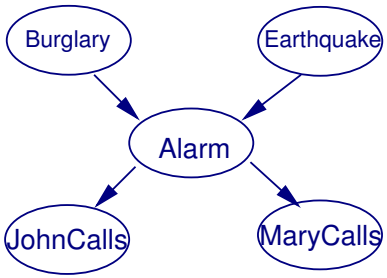
Belief Network: Example



New burglar alarm was installed.

- The alarm can be triggered by either an actual burglary or an earthquake.
- Neighbors John and Mary agreed to call you at work when they hear the alarm.

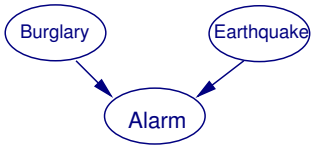
Belief Network: Example (cont'd)



Example question: If you got calls from John and Mary, what is the chance of it being a totally false alarm (not a burglary, nor an earthquake)?

You can ask any conjunctive combination.

Belief Network: Example (cont'd)

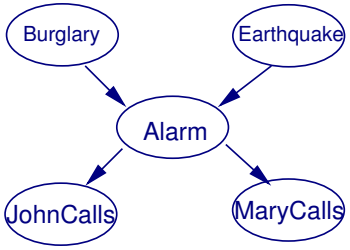


Each node has a conditional probability table fully describing $\mathbf{P}(Current|Parent_1, Parent_2, ..., Parent_n)$:

Burglary	Earthquake	$\mathbf{P}$	$\neg\mathbf{P}$
T	**T**	0.950	0.050
T	**F**	0.950	0.050
F	**T**	0.290	0.710
F	**F**	0.001	0.999

$\mathbf{P}$= $\mathbf{P}(Alarm|Burglary, Earthquake)$

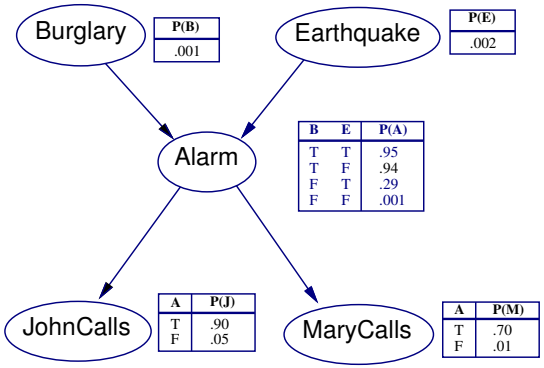
Semantics of Belief Networks



The network can be viewed as

- a representation of the **joint probability distribution** (this view is helpful when **constructing** the network), or
- an encoding of a collection of **conditional independence** statements (this view is helpful when designing effective **inference procedures**).

Belief Network: Representing Joint Prob. Dist.



Each row in the conditional probability tables are:

$$P(X_1 = x_1, ..., X_n = x_n) = P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i | Parent(X_i))$$

Belief Network: Representing Joint Prob. Dist. (cont'd)

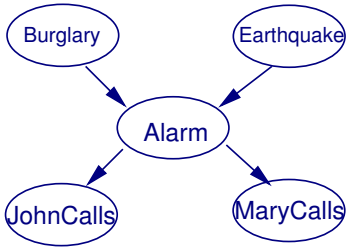
$$P(X_1 = x_1, ..., X_n = x_n) = P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i | Parent(X_i))$$

$Parent(X_j)$ refers to the event when $X_j = x_j$.

Imagine a case where each node in the example has a $\mathbf{T}$ or $\mathbf{F}$ assignment. X_j will then be either $\mathbf{T}$ or $\mathbf{F}$ for all j.

The belief network fully defines a joint probability distribution!

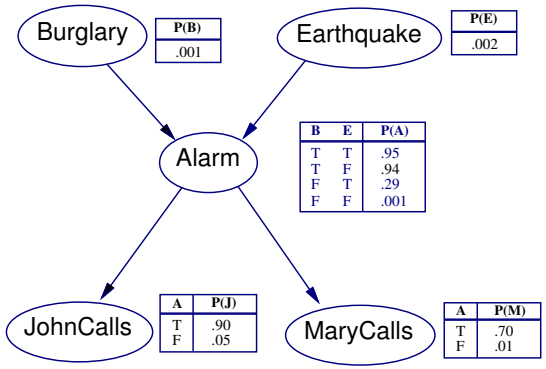
Calculating Probability of a Joint Event



Calculate the probability of the event that the alarm (A) has sounded but neither a burglary ($\neg B$) nor an earthquake ($\neg E$) occurred, and both John (J) and Mary (M) call:

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

Calculating Probability of a Joint Event (cont'd)



$$P(J|A)P(M|A)P(A|\neg B \wedge \neg E)P(\neg B)P(\neg E)$$
$$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 = 0.00062$$

Key Points

- Belief network: definition, semantics, extracting probabilities of certain conjunction of events.

Overview

- Constructing a belief network
- Inference in belief networks
- Knowledge engineering

Joint Probability Distribution Under Conditional Independence

In Belief Networks, the joint probability is given as follows:

$$P(x_1, ..., x_n) = \prod_{i=1}^{N} P(x_i|Parents(X_i)).$$

This is derived from the two following equations:

$$\begin{aligned} P(x_1, ..., x_n) &= P(x_n|x_{n-1}, ..., x_1)P(x_{n-1}, ..., x_1) \\ &= P(x_n|x_{n-1}, ..., x_1)P(x_{n-1}|x_{x-2}, ..., x_1) \\ &\quad \cdots P(x_2|x_1)P(x_1) \\ &= \prod_{i=1}^{N} P(x_i|x_{i-1}, ..., x_1) \quad (1) \end{aligned}$$

$$\mathbf{P}(X_i|X_{i-1}, ..., X_1) = \mathbf{P}(X_i|Parents(X_i)) \quad (2)$$

Belief Network Construction

Given the nodes, we have to find which nodes are **directly** influenced by certain nodes, and from this, find out **conditional independence**.

- For example:
  $\mathbf{P}(MaryCalls|JohnCalls, Alarm, Earthquake, Burglary)$
  $= \mathbf{P}(MaryCalls|Alarm)$
- So, even if $Earthquake$ and $Burglary$ can be found up-stream (e.g. $Earthquake$ causing $Alarm$ to go off, in turn causing $MaryCalls$), those events are conditionally independent from $MaryCalls$.

To construct a Belief Network, we need to find such dependency structure.

Belief Network Construction (cont'd)

The general procedure for Belief Network construction is as follows:

1. Choose the set of relevant variables X_i that describe the domain.
2. Choose an ordering of the variables.
3. While there are variables left:
   (a) Pick a variable X_i and add a node to the network for that variable.
   (b) Set $Parents(X_i)$ to some minimal set of nodes already in the net such that the conditional independence property is satisfied.
   (c) Define the conditional probability table for node X_i.

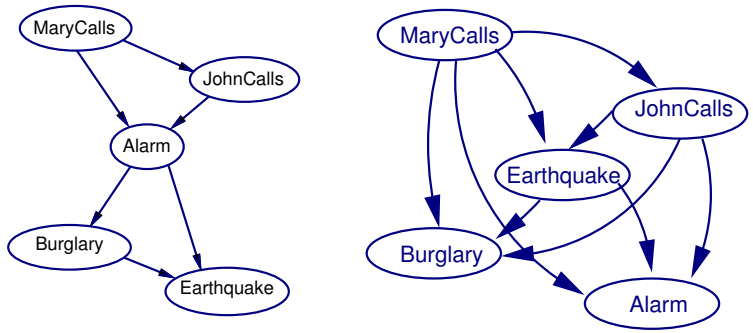
Evaluation of the Construction Algorithm

- Because newly added nodes cannot point to existing nodes, the resulting graph is always **acyclic**.
- Violation of axioms of probability is avoided.
- Compact, compared to the full joint probability table (**locally structured, or sparse**).
  - Belief network with n nodes (binary variables) and k parents per node has $n2^k$ entries in the conditional probability tables:

    n nodes $\times$ 2^k per each node.

  - Full joint probability table: 2^n

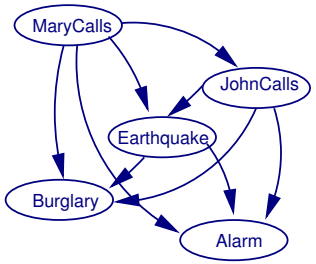
Importance of Node Ordering in Belief Network Construction



The resulting Belief Network can be vastly different when the order of insertion of nodes into the network is different.

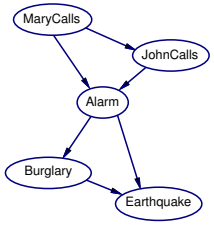
- $MaryCalls, JohnCalls, Alarm, Burglary, Earthquake$
- $MaryCalls, JohnCalls, Earthquake, Burglary, Alarm$

Node Ordering and Joint Probability Tables



- Even with different graphs resulting from different node ordering, you can represent the same joint probability distribution.
- However, some represent the conditional independence relation much better than others.
- For example, the graph above requires the same number of entries as a full joint probability table.

Improper Node Ordering Can Cause Problems



- When adding nodes in the order of $MaryCalls$, $JohnCalls$, ..., conditional independence does not hold:

$$P(JohnCalls|MaryCalls) \neq P(JohnCalls)$$

- Thus, $MaryCalls$ has to become a parent of $JohnCalls$.
- This is because if Mary calls, it probably means that the alarm has gone off, so it makes it more likely that John calls.

Strategy for Better Node Ordering

- For a node A to be a parent of another node B, A must be added to the network before B is added.
- Thus, a node that has a direct influence on other nodes should be added to the network first, i.e. add the **root causes** first.

Strategy for Better Node Ordering (cont'd)

When building the network, stick to a **causal model**, rather than the other way around (e.g. inferring cause given the effect).

Causal model is beneficial because of these reasons:

- conditional probability tables can be made smaller
- the conditional probabilities can be easier to come up with
- easier to reason about the domain using the network

Probabilistic Inference

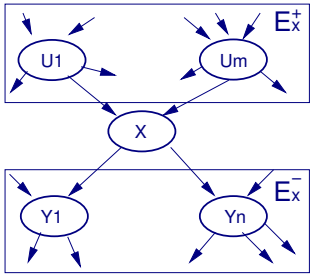
- **Diagnostic inferences**: $P(Cause|Effect)$
- **Causal inferences**: $P(Effect|Cause)$
- **Intercausal inferences**: causes of a common effect (**explaining away**: cause has already been found)

$$P(Cause|Effect) >> P(Cause|Effect \wedge OtherCause)$$

- **Mixed inferences**: combining two or more of the above

$$P(A|CauseOfA \wedge EffectOfA)$$

Answering Queries: A Brief Outline



Given a set of evidence E, find the conditional probability $\mathbf{P}(X|E)$ where X is the query.

- Recursively deterimine the **causal support** E_X^+ for X.
- Recursively deterimine the **evidential support** E_X^- for X.

Note: This is only when the graph is **singly connected**.

Using Belief Networks and Probabilistic Inference

- Making decisions based on the derived probabilities and an agent's utility function.
- Deciding which additional variables to include in the model.
- Performing **sensitivity analysis** to find out which node is most important and thus should be more accurate.
- Explaining the results of reasoning.

Knowledge Engineering for Uncertain Reasoning

- Decide what to talk about (i.e. what to be included in the model). Gradually add more factors that can influence the current collection of events.
- Determine the variables to use and the range of the values.
- Encode general knowledge about the dependence between variables:
  1. qualitative: which variable depends on some other variable
  2. quantitative: probability value of the dependence (from experience, or from data gathered from a sample space)
- Encode a description of the specific problem instance: assign values to the variables.
- Pose queries to the inference procedure and get answers: what is the probability of X? how sensitive are the values in the conditional probability tables to perturbations?

Key Points

- Constructing a belief network: what is the procedure? why does node ordering matter? how to order the nodes?
- Inference in belief networks: what are the kinds of inference? what is the general method?
- Knowledge engineering: how to formulate the idea and design a system.