Uncertainty I: Probability as Degree of Belief

We now examine:

- how **probability theory** might be used to represent and reason with knowledge when we are **uncertain** about the world;
- how **inference** in the presence of uncertainty can in principle be performed using only basic results along with the **full joint probability distribution**;
- how this approach **fails** in practice;
- how the notions of **independence** and **conditional independence** may be used to solve this problem.

Reading: Russell and Norvig, chapter 13.

Uncertainty in AI

The (predominantly logic-based) methods covered so far have assorted shortcomings:

- limited epistemological commitment—true/false/unknown;
- actions are possible when sufficient knowledge is available...
- ...but this is not generally the case;
- in practice there is a need to cope with **uncertainty**.

For example in the Wumpus World:

- we can not make observations further afield than the current locality;
- consequently inferences regarding pit/wumpus location *etc* will not usually be possible.

Uncertainty in AI

A couple of more subtle problems have also presented themselves:

- The **Qualification Problem:** it is not generally possible to guarantee that an action will succeed—only that it will succeed if **many other preconditions** do/don't hold.
- **Rational action** depends on the **likelihood** of achieving different goals, and their **relative desirability**.

Logic (as seen so far) has major shortcomings

An example:

$$\forall x\ \texttt{symptom}(x, \texttt{toothache}) \rightarrow \texttt{problem}(x, \texttt{cavity})$$

This is plainly incorrect. Toothaches can be caused by things other than cavities.

$$\begin{aligned}\forall x\ \texttt{symptom}(x, \texttt{toothache}) \rightarrow &\texttt{problem}(x, \texttt{cavity}) \vee \\ &\texttt{problem}(x, \texttt{abscess}) \vee \\ &\texttt{problem}(x, \texttt{gum-disease}) \vee \\ &\ldots\end{aligned}$$

BUT:

- it is impossible to complete the list;
- there's no clear way to take account of the relative likelihoods of different causes.

Logic (as seen so far) has major shortcomings

If we try to make a **causal rule**

$$\forall x \; \texttt{problem}(x, \texttt{abscess}) \rightarrow \texttt{symptom}(x, \texttt{toothache})$$

it's still wrong—abscesses do not always cause pain.

We need further information in addition to

$$\texttt{problem}(x, \texttt{abscess})$$

and it's still not possible to do this correctly.

Logic (as seen so far) has major shortcomings

FOL can fail for essentially three reasons:

1. **Laziness:** it is not feasible to assemble a set of rules that is sufficiently exhaustive.
   If we could, it would not be feasible to apply them.
2. **Theoretical ignorance:** insufficient knowledge **exists** to allow us to write the rules.
3. **Practical ignorance:** even if the rules have been obtained there may be insufficient information to apply them.

Truth, falsehood, and belief

Instead of thinking in terms of the **truth** or **falsity** of a statement we want to deal with an agent's **degree of belief** in the statement.

- **Probability theory** is the perfect tool for application here.
- **Probability theory** allows us to **summarise** the uncertainty due to laziness and ignorance.

An important distinction

There is a fundamental difference between **probability theory** and **fuzzy logic**:

- when dealing with probability theory, statements remain **in fact** either **true** or **false**;
- a probability denotes an agent's **degree of belief** one way or another;
- fuzzy logic deals with **degree of truth**.

In practice the use of probability theory has proved spectacularly successful.

Belief and evidence

An agent's beliefs will depend on what it has **perceived**: probabilities are based on **evidence** and may be altered by the acquisition of new evidence:

- **Prior (unconditional) probability** denotes a degree of belief in the absence of evidence;
- **Posterior (conditional) probability** denotes a degree of belief after evidence is perceived.

As we shall see **Bayes' theorem** is the fundamental concept that allows us to update one to obtain the other.

Making rational decisions under uncertainty

When using **logic**, we concentrated on finding an action sequence guaranteed to achieve a goal, and then executing it.

When dealing with **uncertainty** we need to define **preferences** among states of the world and take into account the **probability** of reaching those states.

Utility theory is used to assign preferences.

Decision theory combines probability theory and utility theory.

A **rational** agent should act in order to maximise expected utility.

Probability

We want to assign degrees of belief to propositions about the world.

We will need:

- **Random variables** with associated **domains**—typically boolean, discrete, or continuous;
- all the usual concepts—events, atomic events, sets *etc*;
- probability distributions and densities;
- probability axioms (Kolmogorov);
- conditional probability and Bayes' theorem.

So if you've forgotten this stuff now is a good time to re-read it.

Probability

The standard axioms are:

- Range
$$0 \leq \Pr(x) \leq 1$$
- Always true propositions
$$\Pr(\texttt{always true proposition}) = 1$$
- Always false propositions
$$\Pr(\texttt{always false proposition}) = 0$$
- Union
$$\Pr(x \vee y) = \Pr(x) + \Pr(y) - \Pr(x \wedge y)$$

Origins of probabilities I

Historically speaking, probabilities have been regarded in a number of different ways:

- **Frequentist:** probabilities come from measurements;
- **Objectivist:** probabilities are actual "properties of the universe" which frequentist measurements seek to uncover.

  An excellent example: quantum phenomena.

  A bad example: coin flipping—the uncertainty is due to our uncertainty about the initial conditions of the coin.
- **Subjectivist:** probabilities are an agent's degrees of belief.

  This means the agent is allowed to make up the numbers!

Origins of probabilities II

The **reference class problem**: even frequentist probabilities are subjective.

Example: Say a doctor takes a frequentist approach to diagnosis. She examines a large number of people to establish the prior probability of whether or not they have heart disease.

To be accurate she tries to measure "similar people". (She knows for example that gender might be important.)

Taken to an extreme, **all** people are **different** and there is therefore no **reference class**.

Origins of probabilities III

The **principle of indifference** (Laplace).

- Give equal probability to all propositions that are syntactically symmetric with respect to the available evidence.
- Refinements of this idea led to the attempted development by Carnap and others of **inductive logic**.
- The aim was to obtain the correct probability of any proposition from an arbitrary set of observations.

It is currently thought that no unique inductive logic exists.

Any inductive logic depends on prior beliefs and the effect of these beliefs is overcome by evidence.

Prior probability

A **prior probability** denotes the probability (degree of belief) assigned to a proposition **in the absence of any other evidence**.

For example

$$\Pr(\texttt{Cavity} = \texttt{true}) = 0.05$$

denotes the degree of belief that a random person has a cavity **before we make any actual observation of that person**.

Notation

To keep things compact, we will use

$$\Pr(\texttt{Cavity})$$

to denote the entire probability distribution of the random variable `Cavity`.

Instead of

$$\Pr(\texttt{Cavity} = \texttt{true}) = 0.05$$
$$\Pr(\texttt{Cavity} = \texttt{false}) = 0.95$$

write

$$\Pr(\texttt{Cavity}) = (0.05, 0.95)$$

Notation

A similar convention will apply for joint distributions. For example, if `Decay` can take the values `severe`, `moderate` or `low` then

$$\Pr(\texttt{Cavity}, \texttt{Decay})$$

is a 2 by 3 table of numbers.

	`severe`	`moderate`	`low`
`true`	0.26	0.1	0.01
`false`	0.01	0.02	0.6

Similarly

$$\Pr(\texttt{true}, \texttt{Decay})$$

denotes 3 numbers *etc*.

The full joint probability distribution

The **full joint probability distribution** is the joint distribution of **all** random variables that describe the state of the world.

This can be used to answer **any query**.

(But of course life's not really that simple!)

Conditional probability

We use the **conditional probability**

$$\Pr(x|y)$$

to denote the probability that a proposition x holds given that **all the evidence we have so far** is contained in proposition y.

From basic probability theory

$$\Pr(x|y) = \frac{\Pr(x \wedge y)}{\Pr(y)}$$

Conditional probability and implication are distinct

Conditional probability is **not** analogous to **logical implication**.

- $\Pr(x|y) = 0.1$ does **not** mean that if y is true then $\Pr(x) = 0.1$.
- $\Pr(x)$ is a prior probability.
- The notation $\Pr(x|y)$ is for use when y is the entire evidence.
- $\Pr(x|y \wedge z)$ might be very different.

Using the full joint distribution to perform inference

We can regard the full joint distribution as a **knowledge base**.

We want to use it to obtain answers to questions.

	CP		¬ CP	
	HBP	¬ HBP	HBP	¬ HBP
HD	0.09	0.05	0.07	0.01
¬ HD	0.02	0.08	0.03	0.65

We'll use this medical diagnosis problem as a running example.

- $\mathtt{HD} = \mathtt{Heart\ disease}$
- $\mathtt{CP} = \mathtt{Chest\ pain}$
- $\mathtt{HBP} = \mathtt{High\ blood\ pressure}$

Using the full joint distribution to perform inference

The process is nothing more than the application of basic results:

- Sum atomic events:

$$\begin{aligned}\Pr(\mathsf{HD} \vee \mathsf{CP}) = &\Pr(\mathsf{HD} \wedge \mathsf{CP} \wedge \mathsf{HBP}) \\ &+ \Pr(\mathsf{HD} \wedge \mathsf{CP} \wedge \neg\mathsf{HBP}) \\ &+ \Pr(\mathsf{HD} \wedge \neg\mathsf{CP} \wedge \mathsf{HBP}) \\ &+ \Pr(\mathsf{HD} \wedge \neg\mathsf{CP} \wedge \neg\mathsf{HBP}) \\ &+ \Pr(\neg\mathsf{HD} \wedge \mathsf{CP} \wedge \mathsf{HBP}) \\ &+ \Pr(\neg\mathsf{HD} \wedge \mathsf{CP} \wedge \neg\mathsf{HBP}) \\ &= 0.09 + 0.05 + 0.07 + 0.01 + 0.02 + 0.08 \\ &= 0.32\end{aligned}$$

- Marginalisation: if A and B are sets of variables then

$$\Pr(A) = \sum_b \Pr(A \wedge b) = \sum_b \Pr(A|b)\Pr(b)$$

Using the full joint distribution to perform inference

Usually we will want to compute the conditional probability of some variable(s) given some evidence.

For example

$$\Pr(\mathsf{HD}|\mathsf{HBP}) = \frac{\Pr(\mathsf{HD} \wedge \mathsf{HBP})}{\Pr(\mathsf{HBP})} = \frac{0.09 + 0.07}{0.09 + 0.07 + 0.02 + 0.03} = 0.76$$

and

$$\Pr(\neg\mathsf{HD}|\mathsf{HBP}) = \frac{\Pr(\neg\mathsf{HD} \wedge \mathsf{HBP})}{\Pr(\mathsf{HBP})} = \frac{0.02 + 0.03}{0.09 + 0.07 + 0.02 + 0.03} = 0.24$$

Using the full joint distribution to perform inference

The process can be simplified slightly by noting that

$$\alpha = \frac{1}{\Pr(\mathsf{HBP})}$$

is a constant and can be regarded as a normaliser making relevant probabilities sum to 1.

So a short cut is to avoid computing it as above. Instead:

$$\Pr(\mathsf{HD}|\mathsf{HBP}) = \alpha \Pr(\mathsf{HD} \wedge \mathsf{HBP}) = (0.09 + 0.07)\alpha$$

$$\Pr(\neg\mathsf{HD}|\mathsf{HBP}) = \alpha \Pr(\neg\mathsf{HD} \wedge \mathsf{HBP}) = (0.02 + 0.03)\alpha$$

and we need

$$\Pr(\mathsf{HD}|\mathsf{HBP}) + \Pr(\neg\mathsf{HD}|\mathsf{HBP}) = 1$$

so

$$\alpha = \frac{1}{0.09 + 0.07 + 0.02 + 0.03}$$

Using the full joint distribution to perform inference

The general inference procedure is as follows:

$$\Pr(Q|e) = \frac{1}{Z}\Pr(Q \wedge e) = \frac{1}{Z}\sum_u \Pr(Q, e, u)$$

where

- Q is the query variable
- e is the evidence
- u are the unobserved variables
- $1/Z$ normalises the distribution.

Using the full joint distribution to perform inference

Simple eh?

Well, no...

- For n Boolean variables the table has 2^n entries.
- Storage and processing time are both $O(2^n)$.
- You need to establish 2^n numbers to work with.

In reality we might well have $n > 1000$, and of course it's even worse if variables are non-Boolean.

How can we get around this?

Exploiting independence

If I toss a coin and roll a dice, the full joint distribution of outcomes requires $2 \times 6 = 12$ numbers to be specified.

	1	2	3	4	5	6
Head	0.014	0.028	0.042	0.057	0.071	0.086
Tail	0.033	0.067	0.1	0.133	0.167	0.2

Here $\Pr(\texttt{Coin} = \texttt{head}) = 0.3$ and the dice has probability $i/21$ for the ith outcome.

BUT: if we assume the outcomes are independent then

$$\Pr(\texttt{Coin}, \texttt{Dice}) = \Pr(\texttt{Coin})\Pr(\texttt{Dice})$$

Where $\Pr(\texttt{Coin})$ has two numbers and $\Pr(\texttt{Dice})$ has six.

So instead of 12 numbers we only need 8.

Exploiting independence

Similarly, say instead of just considering HD, HBP and CP we also consider the outcome of the Oxford versus Cambridge tiddlywinks competition TC

$$\Pr(\texttt{TC = Oxford}) = 0.2$$
$$\Pr(\texttt{TC = Cambridge}) = 0.7$$
$$\Pr(\texttt{TC = Draw}) = 0.1$$

Now

$$\Pr(\texttt{HD}, \texttt{HBP}, \texttt{CP}, \texttt{TC}) = \Pr(\texttt{TC}|\texttt{HD}, \texttt{HBP}, \texttt{HD}) \Pr(\texttt{HD}, \texttt{HBP}, \texttt{HD})$$

Assuming that the patient is not an extraordinarily keen fan of tiddlywinks, their cardiac health has nothing to do with the outcome, so

$$\Pr(\texttt{TC}|\texttt{HD}, \texttt{HBP}, \texttt{HD}) = \Pr(\texttt{TC})$$

and $2 \times 2 \times 2 \times 3 = 24$ numbers has been reduced to $3 + 8 = 11$.

Exploiting independence

In general you need to identify such independence through knowledge of the problem.

BUT:

- it generally does not work as clearly as this;
- the independent subsets themselves can be big.

Bayes theorem

From first principles

$$\Pr(x, y) = \Pr(x|y) \Pr(y)$$
$$\Pr(x, y) = \Pr(y|x) \Pr(x)$$

so

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

The most important equation in modern AI?

When evidence e is involved this can be written

$$\Pr(Q|R, e) = \frac{\Pr(R|Q, e) \Pr(Q|e)}{\Pr(R|e)}$$

Bayes theorem

Taking another simple medical diagnosis example: does a patient with a fever have malaria? A doctor might know that

$$\Pr(\texttt{fever}|\texttt{malaria}) = 0.99$$
$$\Pr(\texttt{malaria}) = \frac{1}{10000}$$
$$\Pr(\texttt{fever}) = \frac{1}{20}$$

Consequently we can try to obtain $\Pr(\texttt{malaria}|\texttt{fever})$ by direct application of Bayes theorem

$$\Pr(\texttt{malaria}|\texttt{fever}) = \frac{0.99 \times 0.0001}{0.05} = 0.00198$$

or using the alternative technique

$$\Pr(\texttt{malaria}|\texttt{fever}) = \alpha \Pr(\texttt{fever}|\texttt{malaria}) \Pr(\texttt{malaria})$$

if the relevant further quantity $\Pr(\texttt{fever}|\neg\texttt{malaria})$ is known.

Bayes theorem

- Sometimes the first possibility is easier, sometimes not.
- **Causal knowledge** such as

  $$\Pr(\texttt{fever}|\texttt{malaria})$$

  might well be available when **diagnostic knowledge** such as

  $$\Pr(\texttt{malaria}|\texttt{fever})$$

  is not.
- Say the incidence of malaria, modelled by $\Pr(\texttt{Malaria})$, suddenly changes. Bayes theorem tells us what to do.
- The quantity

  $$\Pr(\texttt{fever}|\texttt{malaria})$$

  would not be affected by such a change.

  Causal knowledge can be more robust.

Conditional independence

What happens if we have multiple pieces of evidence?

We have seen that to compute

$$\Pr(\mathsf{HD}|\mathsf{CP},\mathsf{HBP})$$

directly might well run into problems.

we could try using Bayes theorem to obtain

$$\Pr(\mathsf{HD}|\mathsf{CP},\mathsf{HBP}) = \alpha \Pr(\mathsf{CP},\mathsf{HBP}|\mathsf{HD}) \Pr(\mathsf{HD})$$

However while HD is probably manageable, a quantity such as $\Pr(\mathsf{CP},\mathsf{HBP}$ might well still be problematic especially in more realistic cases.

Conditional independence

However although in this case we might not be able to exploit independence directly we **can** say that

$$\Pr(\mathsf{CP},\mathsf{HBP}|\mathsf{HD}) = \Pr(\mathsf{CP}|\mathsf{HD}) \Pr(\mathsf{HBP}|\mathsf{HD})$$

which simplifies matters.

Conditional independence:

- $\Pr(A,B|C) = \Pr(A|C) \Pr(B|C)$
- If we know that C is the case then A and B are independent.

Although CP and HBP are **not** independent, they do not directly influence one another in a patient known to have heart disease.

This is much nicer!

$$\Pr(\mathsf{HD}|\mathsf{CP},\mathsf{HBP}) = \alpha \Pr(\mathsf{CP}|\mathsf{HD}) \Pr(\mathsf{HBP}|\mathsf{HD}) \Pr(\mathsf{HD})$$

Naive Bayes

Conditional independence is often assumed even when it does not hold.

Naive Bayes:

$$\Pr(A,B_1,B_2,\ldots,B_n) = \Pr(A) \prod_{i=1}^{n} \Pr(B_i|A)$$

Also known as **Idiot's Bayes**.

Despite this, it is often surprisingly effective.

Uncertainty II - Bayesian Networks

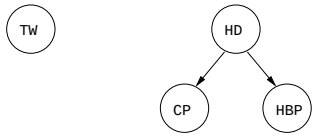
Having seen that in principle, if not in practice, the full joint distribution alone can be used to perform any inference of interest, we now examine a practical technique.

- We introduce the Bayesian network (BN) as a compact representation of the full joint distribution.
- We examine the way in which a BN can be constructed.
- We examine the semantics of BNs.
- We look briefly at how inference can be performed.

Reading: Russell and Norvig, chapter 14.

Bayesian networks

Also called **probabilistic/belief/causal networks** or **knowledge maps**.



- Each node is a random variable (RV).
- Each node N_i has a distribution

$$\Pr(N_i|\texttt{parents}(N_i))$$

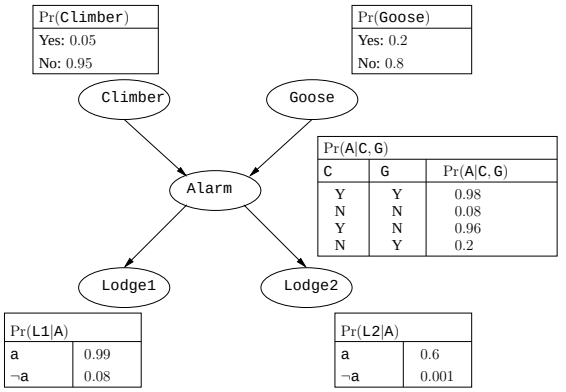
- A Bayesian network is a directed acyclic graph.
- Roughly speaking, an arrow from N to M means N directly affects M.

Bayesian networks

After a regrettable incident involving an inflatable gorilla, a famous College has decided to install an alarm for the detection of roof climbers.

- The alarm is **very** good at detecting climbers.
- Unfortunately, it is also sometimes triggered when one of the extremely fat geese that lives in the College lands on the roof.
- One porter's lodge is near the alarm, and inhabited by a chap with excellent hearing and a **pathological hatred** of roof climbers: he **always** reports an alarm. His hearing is so good that he sometimes thinks he hears an alarm, even when there isn't one.
- Another porter's lodge is a good distance away and inhabited by an old chap with dodgy hearing who likes to watch his collection of **videos** with the sound turned up.

Bayesian networks



Bayesian networks

Note that:

- in the present example all RVs are discrete (in fact Boolean) and so in all cases $\Pr(N_i|\text{parents}(N_i))$ can be represented as a table of numbers;
- `Climber` and `Goose` have only prior probabilities;
- all RVs here are Boolean, so a node with p parents requires 2^p numbers.

Semantics

A BN with n nodes represents the full joint probability distribution for those nodes as

$$\Pr(N_1 = n_1, N_2 = n_2, \ldots, N_n = n_n) = \prod_{i=1}^{n} \Pr(N_i = n_i|\text{parents}(N_i)) \quad (1)$$

For example

$$\begin{aligned}\Pr(\neg\mathtt{C}, \neg\mathtt{G}, \mathtt{A}, \mathtt{L1}, \mathtt{L2}) &= \Pr(\mathtt{L1}|\mathtt{A}) \Pr(\mathtt{L2}|\mathtt{A}) \Pr(\mathtt{A}|\neg\mathtt{C}, \neg\mathtt{G}) \Pr(\neg\mathtt{C}) \Pr(\neg\mathtt{G}) \\ &= 0.99 \times 0.6 \times 0.08 \times 0.95 \times 0.8\end{aligned}$$

Semantics

In general $\Pr(A, B) = \Pr(A|B) \Pr(B)$ so abbreviating $\Pr(N_1 = n_1, N_2 = n_2, \ldots, N_n = n_n)$ to $\Pr(n_1, n_2, \ldots, n_n)$ we have

$$\Pr(n_1, \ldots, n_n) = \Pr(n_n|n_{n-1}, \ldots, n_1) \Pr(n_{n-1}, \ldots, n_1)$$

Repeating this gives

$$\begin{aligned}\Pr(n_1, \ldots, n_n) &= \Pr(n_n|n_{n-1}, \ldots, n_1) \Pr(n_{n-1}|n_{n-2}, \ldots, n_1) \cdots \Pr(n_1) \\ &= \prod_{i=1}^{n} \Pr(n_i|n_{i-1}, \ldots, n_1)\end{aligned} \quad (2)$$

Now compare equations 1 and 2. We see that BNs make the assumption

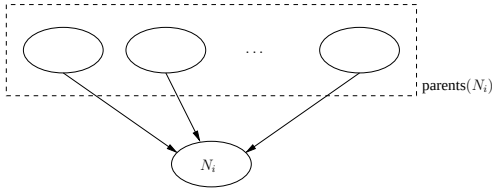
$$\Pr(N_i|N_{i-1}, \ldots, N_1) = \Pr(N_i|\text{parents}(N_i))$$

for each node, assuming that $\text{parents}(N_i) \subseteq \{N_{i-1}, \ldots, N_1\}$.

Each N_i is conditionally independent of its predecessors given its parents

Semantics

- When constructing a BN we want to make sure the preceding property holds.
- This means we need to take care over ordering.
- In general causes should directly precede effects.



Here, $\text{parents}(N_i)$ contains all preceding nodes having a direct influence on N_i.

Semantics

Deviation from this rule can have major effects on the complexity of the network.

That's bad! We want to keep the network simple:

- if each node has at most p parents and there are n Boolean nodes, we need to specify at most $n2^p$ numbers...
- ...whereas the full joint distribution requires us to specify 2^n numbers.

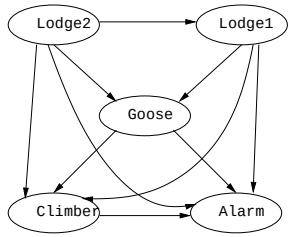
So: there is a trade-off attached to the inclusion of tenuous although strictly-speaking correct edges.

Semantics

As a rule, we should include the most basic causes first, then the things they influence directly *etc*.

What happens if you get this wrong?

Example: add nodes in the order `L2`, `L1`, `G`, `C`, A.



Semantics

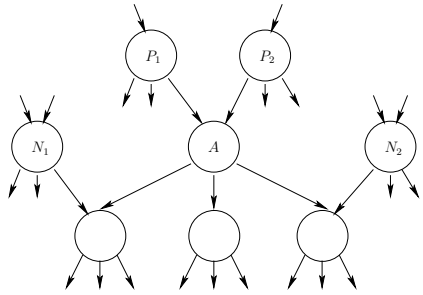
In this example:

- increased connectivity;
- many of the probabilities here will be quite unnatural and hard to specify.

Once again: causal knowledge is preferred to diagnostic knowledge.

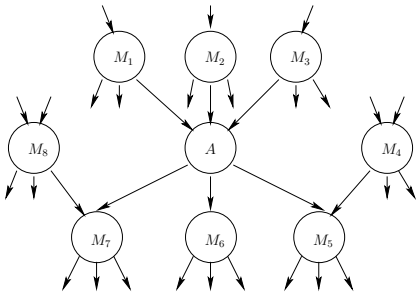
Semantics

As an alternative we can say directly what conditional independence assumptions a graph should be interpreted as expressing. There are two common ways of doing this.



Any node A is conditionally independent of the N_i—its non-descendants—given the P_i—its parents.

Semantics



Any node A is conditionally independent of all other nodes given the Markov blanket M_i—that is, its parents, its children, and its children's parents.

More complex nodes

How do we represent

$$\Pr(N_i|\text{parents}(N_i))$$

when nodes can denote general discrete and/or continuous RVs?

- BNs containing both kinds of RV are called hybrid BNs.
- Naive discretisation of continuous RVs tends to result in both a reduction in accuracy and large tables.
- $O(2^p)$ might still be large enough to be unwieldy.
- We can instead attempt to use standard and well-understood distributions, such as the Gaussian.
- This will typically require only a small number of parameters to be specified.

More complex nodes

Example: functional relationships are easy to deal with.

$$N_i = f(\text{parents}(N_i))$$

$$\Pr(N_i = n_i|\text{parents}(N_i)) = \begin{cases} 1 & \text{if } n_i = f(\text{parents}(N_i)) \\ 0 & \text{otherwise} \end{cases}$$

More complex nodes

Example: a continuous RV with one continuous and one discrete parent.

$$\Pr(\texttt{Speed of car}|\texttt{Throttle position}, \texttt{Tuned engine})$$

where SC and TP are continuous and TE is Boolean.

- For a specific setting of $\texttt{ET} = \texttt{true}$ it might be the case that SC increases with TP, but that some uncertainty is involved

$$\Pr(\text{SC}|\text{TP}, \texttt{et}) = N(g_{\texttt{et}}\text{TP} + c_{\texttt{et}}, \sigma^2_{\texttt{et}})$$

- For an un-tuned engine we might have a similar relationship with a different behaviour

$$\Pr(\text{SC}|\text{TP}, \neg\texttt{et}) = N(g_{\neg\texttt{et}}\text{TP} + c_{\neg\texttt{et}}, \sigma^2_{\neg\texttt{et}})$$

There is a set of parameters $\{g, c, \sigma\}$ for each possible value of the discrete RV.

More complex nodes

Example: a discrete RV with a continuous parent

$$\Pr(\texttt{Go roofclimbing}|\texttt{Size of fine})$$

We could for example use the probit distribution

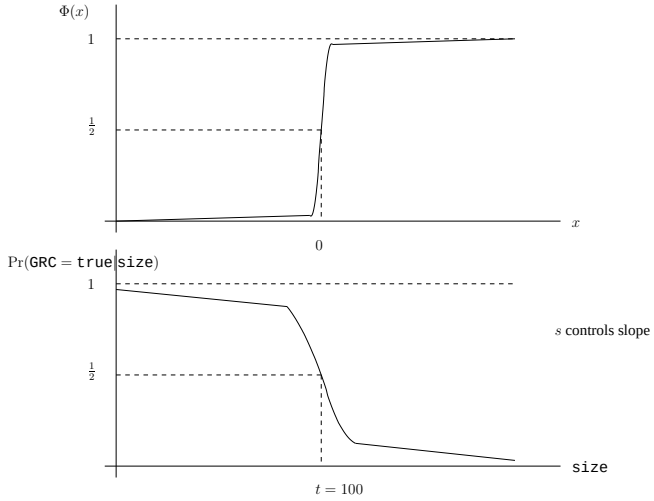
$$\Pr(\texttt{Go roofclimbing} = \texttt{true}|\texttt{size}) = \Phi\left(\frac{t - \texttt{size}}{s}\right)$$

where

$$\Phi(x) = \int_{-\infty}^{x} N(y)dy$$

and $N(x)$ is the Gaussian distribution with zero mean and variance 1.

More complex nodes



More complex nodes

Alternatively, for this example we could use the logit distribution

$$\Pr(\texttt{Go roofclimbing} = \texttt{true}|\texttt{size}) = \frac{1}{1 + e^{(-2(t-\texttt{size})/s)}}$$

which has a similar shape.

- Tails are longer for the logit distribution.
- The logit distribution tends to be easier to use...
- ...but the probit distribution is often more accurate.

Basic inference

We saw earlier that the full joint distribution can be used to perform all inference tasks:

$$\Pr(Q|e) = \frac{1}{Z}\Pr(Q \wedge e) = \frac{1}{Z}\sum_{u}\Pr(Q, e, u)$$

where

- Q is the query variable
- e is the evidence
- u are the unobserved variables
- $1/Z$ normalises the distribution.

Basic inference

As the BN fully describes the full joint distribution

$$\Pr(Q, u, e) = \prod_{i=1}^{n} \Pr(N_i | \text{parents}(N_i))$$

It can be used to perform inference in the obvious way.

$$\Pr(Q|e) = \frac{1}{Z} \sum_{u} \prod_{i=1}^{n} \Pr(N_i | \text{parents}(N_i))$$

- More sophisticated algorithms aim to achieve this more efficiently.
- For complex BNs we resort to approximation techniques.

Other approaches to uncertainty: Default reasoning

One criticism made of probability is that it is numerical whereas human argument seems fundamentally different in nature:

- on the one hand this seems quite defensible. I certainly am not aware of doing logical thought through direct manipulation of probabilities, but;
- on the other hand, neither am I aware of solving differential equations in order to walk!

Default reasoning:

- does not maintain degrees of belief;
- allows something to be believed until a reason is found not to.

Other approaches to uncertainty: rule-based systems

Rule-based systems have some desirable properties:

- Locality: if we establish the evidence X and we have a rule $X \rightarrow Y$ then Y can be concluded regardless of any other rules.
- Detachment: once any Y has been established it can then be assumed. (It's justification is irrelevant.)
- Truth-functionality: truth of a complex formula is a function of the truth of its components.

These are not in general shared by probabilistic systems. What happens if:

- we try to attach measures of belief to rules and propositions;
- we try to make a truth-functional system by, for example, making belief in $X \wedge Y$ a function of beliefs in X and Y?

Other approaches to uncertainty: rule-based systems

Problems that can arise:

1. Say I have the causal rule

   $$\text{Heart disease} \xrightarrow{0.95} \text{Chest pain}$$

   and the diagnostic rule

   $$\text{Chest pain} \xrightarrow{0.7} \text{Heart disease}$$

   Without taking very great care to keep track of the reasoning process, these can form a loop.
2. If in addition I have

   $$\text{Chest pain} \xrightarrow{0.6} \text{Recent physical exertion}$$

   then it is quite possible to form the conclusion that with some degree of certainty heart disease is explained by exertion, which may well be incorrect.

Other approaches to uncertainty: rule-based systems

In addition, we might argue that because heart disease is an explanation for chest pain the belief in physical exertion should decrease.

In general when such systems have been successful it has been through very careful control in setting up the rules.

Other approaches to uncertainty: Dempster-Shafer theory

Dempster-Shafer theory attempts to distinguish between uncertainty and ignorance.

Whereas the probabilistic approach looks at the probability of X, we instead look at the probability that the available evidence supports X.

This is denoted by the belief function Bel(X).

Example: given a coin but no information as to whether it is fair I have no reason to think one outcome should be preferred to another

$$\text{Bel}(\text{outcome} = \text{head}) = \text{Bel}(\text{outcome} = \text{tail}) = 0$$

Other approaches to uncertainty: Dempster-Shafer theory

These beliefs can be updated when new evidence is available. If an expert tells us there is n percent certainty that it's a fair coin then

$$\text{Bel}(\text{outcome} = \text{head}) = \text{Bel}(\text{outcome} = \text{tail}) = \frac{n}{100} \times \frac{1}{2}.$$

We may still have a "gap" in that

$$\text{Bel}(\text{outcome} = \text{head}) + \text{Bel}(\text{outcome} = \text{tail}) \neq 1.$$

Dempster-Shafer theory provides a coherent system for dealing with belief functions.

Other approaches to uncertainty: Dempster-Shafer theory

Problems:

- the Bayesian approach deals more effectively with the quantification of how belief changes when new evidence is available;
- the Bayesian approach has a better connection to the concept of utility, whereas the latter is not well-understood for use in conjunction with Dempster-Shafer theory.

Uncertainty III: exact inference in Bayesian networks

We now examine:

- the basic equation for inference in Bayesian networks, the latter being hard to achieve if approached in the obvious way;
- the way in which matters can be improved a little by a small modification to the way in which the calculation is done;
- the way in which much better improvements might be possible using a still more informed approach, although not in all cases.

Reading: Russell and Norvig, chapter 14, section 14.4.

Performing exact inference

We know that in principle any query Q can be answered by the calculation

$$\Pr(Q|e) = \frac{1}{Z}\sum_{u}\Pr(Q, e, u)$$

where Q denotes the query, e denotes the evidence, u denotes unobserved variables and $1/Z$ normalises the distribution.

The naive implementation of this approach yields the Enumerate-Joint-Ask algorithm, which unfortunately requires $O(2^n)$ time and space for n Boolean random variables (RVs).

Performing exact inference

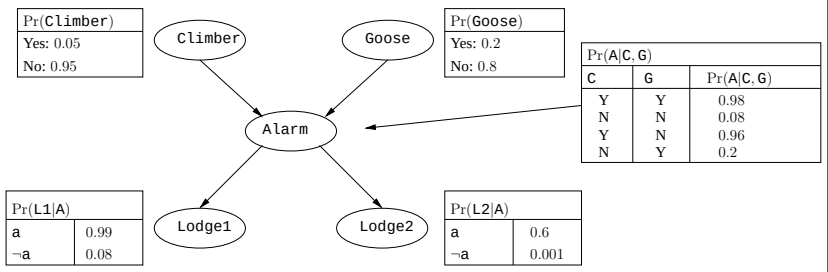
In what follows we will make use of some abbreviations.

- C denotes `Climber`
- G denotes `Goose`
- A denotes `Alarm`
- $L1$ denotes `Lodge1`
- $L2$ denotes `Lodge2`

Instead of writing out $\Pr(C = \top)$, $\Pr(C = \bot)$ *etc* we will write $\Pr(c)$, $\Pr(\neg c)$ and so on.

Performing exact inference

Also, for a Bayesian network, $\Pr(Q, e, u)$ has a particular form expressing conditional independences in the problem. For our earlier example:



$$\Pr(C, G, A, L1, L2) = \Pr(C)\Pr(G)\Pr(A|C, G)\Pr(L1|A)\Pr(L2|A)$$

Performing exact inference

Consider the computation of the query $\Pr(C|l1, l2)$

We have

$$\Pr(C|l1, l2) = \frac{1}{Z} \sum_A \sum_G \Pr(C)\Pr(G)\Pr(A|C, G)\Pr(l1|A)\Pr(l2|A)$$

Here there are 5 multiplications for each set of values that appears for summation, and there are 4 such values.

In general this gives time complexity $O(n2^n)$ for n Boolean RVs.

Performing exact inference

Looking more closely we see that

$$\begin{aligned}\Pr(C|l1, l2) &= \frac{1}{Z} \sum_A \sum_G \Pr(C)\Pr(G)\Pr(A|C, G)\Pr(l1|A)\Pr(l2|A) \\ &= \frac{1}{Z}\Pr(C) \sum_A \Pr(l1|A)\Pr(l2|A) \sum_G \Pr(G)\Pr(A|C, G) \\ &= \frac{1}{Z}\Pr(C) \sum_G \Pr(G) \sum_A \Pr(A|C, G)\Pr(l1|A)\Pr(l2|A)\end{aligned} \quad (3)$$

So for example...

Performing exact inference

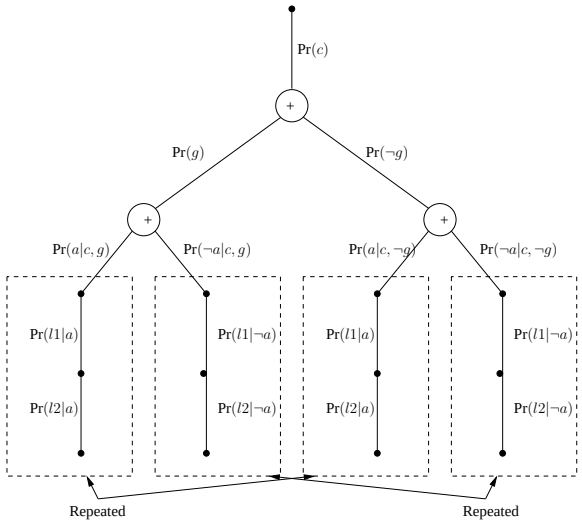
So for example...

$$\Pr(c|l1, l2) = \frac{1}{Z}\Pr(c) \left(\Pr(g) \left\{ \begin{array}{c} \Pr(a|c, g)\Pr(l1|a)\Pr(l2|a) \\ +\Pr(\neg a|c, g)\Pr(l1|\neg a)\Pr(l2|\neg a) \end{array} \right\} + \Pr(\neg g) \left\{ \begin{array}{c} \Pr(a|c, \neg g)\Pr(l1|a)\Pr(l2|a) \\ +\Pr(\neg a|c, \neg g)\Pr(l1|\neg a)\Pr(l2|\neg a) \end{array} \right\} \right)$$

with a similar calculation for $\Pr(\neg c|l1, l2)$.

Basically straightforward, BUT optimisations can be made.

Performing exact inference



Optimisation 1: Enumeration-Ask

The enumeration-ask algorithm improves matters to $O(2^n)$ time and $O(n)$ space by performing the computation depth-first.

However matters can be improved further by avoiding the duplication of computations that clearly appears in the example tree.

Optimisation 2: variable elimination

Looking again at the fundamental equation (3)

$$\frac{1}{Z}\underbrace{\mathsf{Pr}(C)}_{C}\sum_{G}\underbrace{\mathsf{Pr}(G)}_{G}\sum_{A}\underbrace{\mathsf{Pr}(A|C,G)}_{A}\underbrace{\mathsf{Pr}(l1|A)}_{L1}\underbrace{\mathsf{Pr}(l2|A)}_{L2}$$

where $C, G, A, L1, L2$ denote the relevant factors.

The basic idea is to evaluate (3) from right to left (or in terms of the tree, bottom up) storing results as we progress and re-using them when necessary.

$\mathsf{Pr}(l1|A)$ depends on the value of A. We store it as a table $\mathbf{F}_{L1}(A)$. Similarly for $\mathsf{Pr}(l2|A)$.

$$\mathbf{F}_{L1}(A) = \begin{pmatrix} 0.99 \\ 0.08 \end{pmatrix} \quad \mathbf{F}_{L2}(A) = \begin{pmatrix} 0.6 \\ 0.001 \end{pmatrix}$$

as $\mathsf{Pr}(l1|a) = 0.99$, $\mathsf{Pr}(l1|\neg a) = 0.08$ and so on.

Optimisation 2: variable elimination

Similarly for $\mathsf{Pr}(A|C,G)$, which is dependent on A, C and G

$\mathbf{F}_A(A,C,G) =$

A	C	G	$\mathbf{F}_A(A,C,G)$
$\top$	$\top$	$\top$	0.98
$\top$	$\top$	$\bot$	0.96
$\top$	$\bot$	$\top$	0.2
$\top$	$\bot$	$\bot$	0.08
$\bot$	$\top$	$\top$	0.02
$\bot$	$\top$	$\bot$	0.04
$\bot$	$\bot$	$\top$	0.8
$\bot$	$\bot$	$\bot$	0.92

Can we write

$$\mathsf{Pr}(A|C,G)\mathsf{Pr}(l1|A)\mathsf{Pr}(l2|A) \tag{4}$$

as

$$\mathbf{F}_A(A,C,G)\mathbf{F}_{L1}(A)\mathbf{F}_{L2}(A) \tag{5}$$

in a reasonable way?

Optimisation 2: variable elimination

The answer is "yes" provided multiplication of factors is defined correctly. Looking at (3)

$$\frac{1}{Z}\mathsf{Pr}(C)\sum_{G}\mathsf{Pr}(G)\sum_{A}\mathsf{Pr}(A|C,G)\mathsf{Pr}(l1|A)\mathsf{Pr}(l2|A)$$

note that the values of the product (4) in the summation depend on the values of C and G external to it, and the values of A themselves. So (5) should be a table collecting values for (4) where correspondences between RVs are maintained.

This leads to a definition for multiplication of factors best given by example.

Optimisation 2: variable elimination

$$\mathbf{F}(A,B)\mathbf{F}(B,C) = \mathbf{F}(A,B,C)$$

where

A	B	$\mathbf{F}(A,B)$
$\top$	$\top$	0.3
$\top$	$\bot$	0.9
$\bot$	$\top$	0.4
$\bot$	$\bot$	0.1

B	C	$\mathbf{F}(B,C)$
$\top$	$\top$	0.1
$\top$	$\bot$	0.8
$\bot$	$\top$	0.8
$\bot$	$\bot$	0.3

A	B	C	$\mathbf{F}(A,B,C)$
$\top$	$\top$	$\top$	0.3×0.1
$\top$	$\top$	$\bot$	0.3×0.8
$\top$	$\bot$	$\top$	0.9×0.8
$\top$	$\bot$	$\bot$	0.9×0.3
$\bot$	$\top$	$\top$	0.4×0.1
$\bot$	$\top$	$\bot$	0.4×0.8
$\bot$	$\bot$	$\top$	0.1×0.8
$\bot$	$\bot$	$\bot$	0.1×0.3

Optimisation 2: variable elimination

This process gives us

$$\mathbf{F}_A(A,C,G)\mathbf{F}_{L1}(A)\mathbf{F}_{L2}(A) =$$

A	C	G	
$\top$	$\top$	$\top$	$0.98 \times 0.99 \times 0.6$
$\top$	$\top$	$\bot$	$0.96 \times 0.99 \times 0.6$
$\top$	$\bot$	$\top$	$0.2 \times 0.99 \times 0.6$
$\top$	$\bot$	$\bot$	$0.08 \times 0.99 \times 0.6$
$\bot$	$\top$	$\top$	$0.02 \times 0.08 \times 0.001$
$\bot$	$\top$	$\bot$	$0.04 \times 0.08 \times 0.001$
$\bot$	$\bot$	$\top$	$0.8 \times 0.08 \times 0.001$
$\bot$	$\bot$	$\bot$	$0.92 \times 0.08 \times 0.001$

Optimisation 2: variable elimination

How about

$$\mathbf{F}_{\overline{A},L1,L2}(C,G) = \sum_A \mathbf{F}_A(A,C,G)\mathbf{F}_{L1}(A)\mathbf{F}_{L2}(A)$$

To denote the fact that A has been summed out we place a bar over it in the notation.

$$\sum_A \mathbf{F}_A(A,C,G)\mathbf{F}_{L1}(A)\mathbf{F}_{L2}(A) = \mathbf{F}_A(a,C,G)\mathbf{F}_{L1}(a)\mathbf{F}_{L2}(a) + \mathbf{F}_A(\neg a,C,G)\mathbf{F}_{L1}(\neg a)\mathbf{F}_{L2}(\neg a)$$

where

$$\mathbf{F}_A(a,C,G) =$$

C	G	
$\top$	$\top$	0.98
$\top$	$\bot$	0.96
$\bot$	$\top$	0.2
$\bot$	$\bot$	0.08

$\mathbf{F}_{L1}(a) = 0.99$ $\mathbf{F}_{L2}(a) = 0.6$

and similarly for $\mathbf{F}_A(\neg a,C,G)$, $\mathbf{F}_{L1}(\neg a)$ and $\mathbf{F}_{L2}(\neg a)$.

Optimisation 2: variable elimination

$$\mathbf{F}_A(a,C,G)\mathbf{F}_{L1}(a)\mathbf{F}_{L2}(a) =$$

C	G	
$\top$	$\top$	$0.98 \times 0.99 \times 0.6$
$\top$	$\bot$	$0.96 \times 0.99 \times 0.6$
$\bot$	$\top$	$0.2 \times 0.99 \times 0.6$
$\bot$	$\bot$	$0.08 \times 0.99 \times 0.6$

$$\mathbf{F}_A(\neg a,C,G)\mathbf{F}_{L1}(\neg a)\mathbf{F}_{L2}(\neg a) =$$

C	G	
$\top$	$\top$	$0.02 \times 0.08 \times 0.001$
$\top$	$\bot$	$0.04 \times 0.08 \times 0.001$
$\bot$	$\top$	$0.8 \times 0.08 \times 0.001$
$\bot$	$\bot$	$0.92 \times 0.08 \times 0.001$

$$\mathbf{F}_{\overline{A},L1,L2}(C,G) =$$

C	G	
$\top$	$\top$	$(0.98 \times 0.99 \times 0.6) + (0.02 \times 0.08 \times 0.001)$
$\top$	$\bot$	$(0.96 \times 0.99 \times 0.6) + (0.04 \times 0.08 \times 0.001)$
$\bot$	$\top$	$(0.2 \times 0.99 \times 0.6) + (0.8 \times 0.08 \times 0.001)$
$\bot$	$\bot$	$(0.08 \times 0.99 \times 0.6) + (0.92 \times 0.08 \times 0.001)$

Optimisation 2: variable elimination

Now, say for example we have $\neg c$, g. Then doing the calculation explicitly would give

$$\sum_A \mathsf{Pr}(A|\neg c, g)\mathsf{Pr}(l1|A))\mathsf{Pr}(l2|A)$$
$$= \mathsf{Pr}(a|\neg c, g)\mathsf{Pr}(l1|a)\mathsf{Pr}(l2|a) + \mathsf{Pr}(\neg a|\neg c, g)\mathsf{Pr}(l1|\neg a)\mathsf{Pr}(l2|\neg a)$$
$$= (0.2 \times 0.99 \times 0.6) + (0.8 \times 0.08 \times 0.001)$$

which matches!

Continuing in this manner form

$$\mathbf{F}_{G,\overline{A},L1,L2}(C, G) = \mathbf{F}_G(G)\mathbf{F}_{\overline{A},L1,L2}(C, G)$$

sum out G to obtain $\mathbf{F}_{\overline{G},\overline{A},L1,L2}(C) = \sum_G \mathbf{F}_G(G)\mathbf{F}_{\overline{A},L1,L2}(C, G)$, form

$$\mathbf{F}_{C,\overline{G},\overline{A},L1,L2} = \mathbf{F}_C(C)\mathbf{F}_{\overline{G},\overline{A},L1,L2}(C)$$

and normalise.

Optimisation 2: variable elimination

What's the computational complexity now?

- for Bayesian networks with suitable structure we can perform inference in linear time and space;
- however in the worst case it is $\#P$-hard, which is worse than NP-hard.

Consequently, we may need to resort to approximate inference.

Exercises

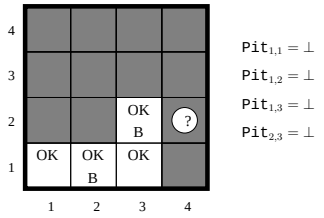
Exercise 1: This question revisits the Wumpus World, but now our hero (let's call him Wizzo the Incompetent), having learned some probability by attending **Artificial Intelligence II**, will use probabilistic reasoning instead of situation calculus.

Wizzo, through careful consideration of the available knowledge on Wumpus caves, has established that each square contains a pit with prior probability 0.3, and pits are independent of one-another. Let $\texttt{Pit}_{i,j}$ be a Boolean random variable (RV) denoting the presence of a pit at row i, column j. So for all i, j

$$\mathsf{Pr}(\texttt{Pit}_{i,j} = \top) = 0.3 \quad (6)$$
$$\mathsf{Pr}(\texttt{Pit}_{i,j} = \bot) = 0.7 \quad (7)$$

In addition, after some careful exploration of the current cave, our hero has discovered the following.

Exercises



B denotes squares where a breeze is perceived. Let $\texttt{Breeze}_{i,j}$ be a Boolean RV denoting the presence of a breeze at i, j

$$\texttt{Breeze}_{1,2} = \texttt{Breeze}_{2,3} = \top \quad (8)$$
$$\texttt{Breeze}_{1,1} = \texttt{Breeze}_{1,3} = \bot \quad (9)$$

Wizzo is considering whether to explore the square at $2, 4$. He will do so if the probability that it contains a pit is less than 0.4. Should he?

Exercises

Hint: The RVs involved are $\texttt{Breeze}_{1,2}, \texttt{Breeze}_{2,3}, \texttt{Breeze}_{1,1}, \texttt{Breeze}_{1,3}$ and $\texttt{Pit}_{i,j}$ for all the i, j. You need to calculate

$$\Pr(\texttt{Pit}_{2,4} | \text{all the evidence you have so far})$$

Exercises

Exercise 2: Continuing with the running example of the roof-climber alarm...

The porter in lodge 1 has left and been replaced by a somewhat more relaxed sort of chap, who doesn't really care about roof-climbers and therefore acts according to the probabilities

$$\Pr(l1|a) = 0.3 \qquad \Pr(\neg l1|a) = 0.7$$
$$\Pr(l1|\neg a) = 0.001 \quad \Pr(\neg l1|\neg a) = 0.999$$

Your intrepid roof-climbing buddy is on the roof. What is the probability that lodge 1 will report him? Use the variable elimination algorithm to obtain the relevant probability. Do you learn anything interesting about the variable $L2$ in the process?

Exercises

Exercise 3: Exam question, 2006, paper 8, question 9.