Lecture 2:

Causal Networks

Probability and Statistics

Statistics emerged as an important mathematical discipline in the late nineteenth and early twentieth century.

Probability is much older and has been studied as long ago as man took an interest in games of chance.

Our story starts relatively recently with the famous theorem of the Rev. Thomas Bayes, published in 1763.

Independent Events

For independent events S and D:

$$P(D\&S) = P(D) \times P(S)$$

(read “disease” for D and “symptom” for S)

Dependent Events

However in cases where S and D are not independent we must write:

$$P(D\&S) = P(D) \times P(S|D)$$

where $P(S|D)$ is the probability of the symptom given that the disease has occurred.

Bayes' Theorem

Now since conjunction is commutative:

$$P(D\&S) = P(S) \times P(D|S) = P(D) \times P(S|D)$$

and re-arranging we get:

$$P(D|S) = P(D) \times P(S|D)/P(S)$$

(Bayes' Theorem)

Bayes' Theorem as an Inference Equation

$$P(D|S) = P(D) \times P(S|D)/P(S)$$

- $P(D|S)$: The probability of the disease given the symptom is what we wish to infer.
- $P(D)$ is the probability of the disease (within a population) this is a measurable quantity.
- $P(S|D)$ is the probability of the symptom given the disease. We can measure this from the case histories of the disease.
- $P(S)$ can also be measured, but fortunately does not need to be.

Notation

Note that:

$$P(D\&S) = P(D) \times P(S)$$

is a scalar equation with two variables: S and D

For much of this course we will use discrete variables. A discrete variable can only have one of a finite number of values (or states), which we denote by lower case letters: s_1, s_2, s_3 etc.

In the simplest case a variable may take just two values (true or false) which we denote by lower case letters, eg:

$$d_t \text{ and } d_f.$$

Normalisation

Suppose that D can take two values (or states): d_t and d_f, and S can take more states: s_1, s_2 etc. Then for any state of S, say s_i we can write:

$$P(d_t|s_i) + P(d_f|s_i) = 1$$

and by applying Bayes' Theorem we can find an expression for $P(s_i)$

$$P(s_i|d_t)P(d_t)/P(s_i) + P(s_i|d_f)P(d_f)/P(s_i) = 1$$
$$P(s_i) = P(s_i|d_t)P(d_t) + P(s_i|d_f)P(d_f)$$

Given values for $P(S|D)$ and $P(D)$ we can calculate $P(S)$ for any state of S. This can be done regardless of the number of states that D and S can take.

Prior and Likelihood Information

We can write $1/P(S)$ as α to remind us it is just a normalising constant:

$$P(D|S) = \alpha \times P(D) \times P(S|D)$$

- $P(D)$ is prior information, since we knew it before we made any measurements.
- $P(S|D)$ is likelihood information, since we find its value from measurement of symptoms.

Bayesian Inference (in its most general form)

- Convert the prior and likelihood information to probabilities;
- Multiply them together;
- Normalise the result to get the posterior probability of each hypothesis given the evidence;
- Select the most probable hypothesis.

More Variables

- When we derived Bayes theorem we had just one hypothesis and one piece of evidence.
- Suppose now that we have evidence from more than one source. Bayes' Theorem is now written:

$$P(D|S_1\&S_2\&S_3\cdots\&S_n) = \frac{P(D)P(S_1\&S_2\cdots S_n|D)}{P(S_1\&S_2\&S_3\cdots S_n)}$$

Conditional Independence

- The term: $P(S_1 \& S_2 \cdots S_n|D)$ is of little use for inference since for large n we are unlikely to be able to estimate it.

- To get round the problem we normally make the assumption that the different S_i are independent given a value for D. This enables us to write:

$$P(S_1 \& S_2 .. S_n|D) = P(S_1|D)P(S_2|D) \cdots P(S_n|D)$$

- However this assumption does not necessarily hold in practice.

Bayesian Inference Equation

As before we can use normalisation to eliminate:

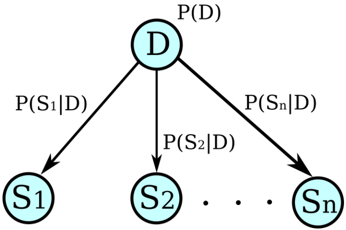
$$P(S_1 \& S_2 \& \cdots \& S_n)$$

and so Bayes theorem becomes:

$$P(D|S_1 \& S_2 .. \& S_n) = \alpha P(D) P(S_1|D) P(S_2|D) \cdots P(S_n|D)$$

Graphical Notation

- We can represent this equation as a graphical model called a Bayesian network.



- Variables (measures or hypothesised) are represented by circles. Nodes are joined to their parents by conditional probabilities. The arrow directions represent causality, in this case the disease is the cause of the symptoms.

Discrete vs Continuous Variables

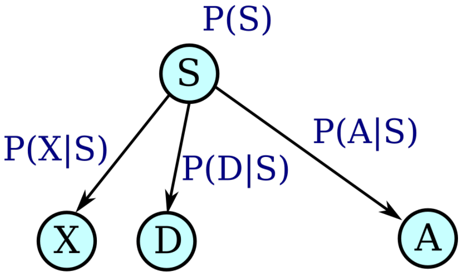
Our variables (hypothesis or evidence) fall into one of two categories:

- Discrete variables take one of a finite number of fixed values or states.
- Continuous variables can take any real value within some range.

Bayesian nets can have a mixture of discrete and continuous variables, but for simplicity we will consider only discrete.

A simple example

A medical expert told me that smoking causes symptoms such as damage to the lungs, dypsnea and aneurysm.



Node	Interpretation	Type	Value
S	Smoking	Discrete (3 states)	Non, Moderate, Heavy
X	Chest XRay	Discrete (4 state)	Clear, Slight, Moderate, Severe
D	Dypsnea	Discrete (12 states)	Baseline Dyspnea Index
A	Aneurysm	Discrete (12 states)	0cm 0.2cm $\cdots$ 2.2cm

Link Matrices

Each node in the network has an associated link matrix (or conditional probability table) which connects it to its immediate parents (or causes). For the link from D to C we have:

$$\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{S}) = \begin{bmatrix} P(x_1|s_1) & P(x_1|s_2) & P(x_1|s_3) \\ P(x_2|s_1) & P(x_2|s_2) & P(x_2|s_3) \\ P(x_3|s_1) & P(x_3|s_2) & P(x_3|s_3) \\ P(x_4|s_1) & P(x_4|s_2) & P(x_4|s_3) \end{bmatrix}$$

Note that the link matrices are written in bold face to distinguish them from the scalar probabilities written, for example, in Bayes' theorem.

Prior Probability of the Roots

- The root nodes of a network do not have any parents. Instead of a link matrix they have a vector giving the prior probabilities of the states, eg:

$$\boldsymbol{P}(\boldsymbol{S}) = [P(s_1), P(s_2), p(s_3)]$$

- This can be thought of a link matrix to empty parents.

Finding the link matrices from data

- We can find the values of the conditional probabilities in the link matrices by experiment.
- To do this we need a large number of cases in which we know the values of all the variables
- For example, the medical records in a hospital might have many patient records of S, X and A.

Suppose that there are $N(x_2 \& d_4)$ points in our data set where variable D is in state d_4 and variable X is in state x_2. Then we can calculate $P(d_4|x_2)$ as follows:

$$P(d_4|x_2) = N(x_2 \& d_4)/N(x_2)$$

Problems in finding the link matrices

- If a network is to represent the variables in an inference problem accurately it may be necessary to have a large number of states for each variable.
- The size of whole network state space is the product of the number of states of each variable. It grows exponentially.
- As the number of conditional probabilities we need to estimate grows, so does the size of the data set that we need to estimate them objectively

Naive Bayesian Network

- Networks of the sort we have considered so far are referred to by a number of names:
  - Naive Bayesian Network
  - Bayesian Classifier
  - Simple Bayesian Network
- They are in many ways the most useful and should be used wherever possible
- They give us a simple way of combining different variables that relate to the same cause.

Using a naive Bayesian network

- Setting a node to a measured value is called instantiation.
- Once a node has been instantiated, we can look up the values for the conditional probabilities in the link matrices.
- Calculating the probabilities of the states of the hypothesis is done by multiplying together the conditional probabilities of the instantiated nodes and the prior probability of the hypothesis node and normalising.

Decision Trees

- The next level of complexity in Bayesian networks is to introduce intermediate nodes between the root node and the leaf nodes.

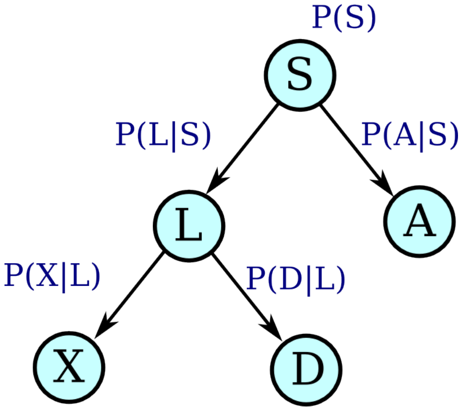
- Suppose our friendly medical expert has just told us:

  "dypsnoea and lung damage frequently go together, their cause being lung cancer"

- We can refine our network into a more complex structure where we explicitly model the relation between them.

The Smoker’s Decision Tree

The nodes X and D are related by a common cause - lung cancer.



Adding the L Node

- In adding a new node we have to decide how many states it has.

- It could be simply binary (true or false), but for better generality we could have three states (or more):

  - l_1 interpreted as probably not cancer
  - l_2 interpreted as could be cancer
  - l_3 interpreted as probably cancer

- To estimate the link matrices we need to consult the hospital data base of patient records.

- Using the network is a little more complex.

Calculating the probability of L

- For the S node and its two children we have:

$$P(S|L\&A) = \alpha P(S)P(L|S)P(A|S)$$

- For the L node we have:

$$P(L|X\&D) = \alpha P(L)P(X|L)P(D|L)$$

- But now we have a problem since we don't want to use a fixed prior probability for L.

- Instead we need to use the evidence that comes from L's parent S.

The Likelihood of L

- From Bayes theorem (last slide) we have

$$P(L|X\&D) = \alpha P(L)P(X|L)P(D|L)$$

- We have likelihood information about L from X and D:

$$\lambda(L|X\&D) = P(X|L)P(D|L)$$

- The likelihood information is evidence, but it does not form a probability distribution. We will call it λ evidence.

The evidence for L from S

Suppose that we are reasoning about a particular patient and we know that he is a heavy smoker. We can now write the probability distribution over the states of S.

$$P'(S) = (0, 0, 1)$$

and this provides specific information that we can send to node L which takes the place of a prior probability. It is:

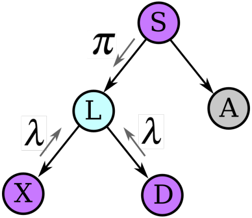
$$\pi(S) = (P(l_1|s_3), P(l_2|s_3), P(l_3|s_3))$$

The total evidence for L

Let us now suppose that we have a value x_2 for X and a value d_3 for D.

$$\lambda(L) = (P(x_2|l_1) \times P(d_3|l_1), P(x_2|l_2) \times P(d_3|l_2), P(x_2|l_3) \times P(d_3|l_3))$$

$$\pi(L) = (P(l_1|s_3), P(l_2|s_3), P(l_3|s_3))$$



and putting it all together:

$$P'(L) = \alpha \times \lambda(L) \times \pi(L)$$

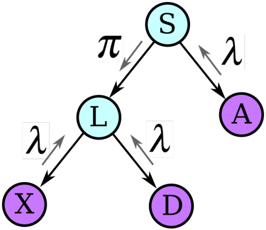
The notation $P'(L)$ is called a posterior probability and just means the probability given all the evidence we have.

Conditioning

Conditioning means calculating probabilities for a given set of conditions.

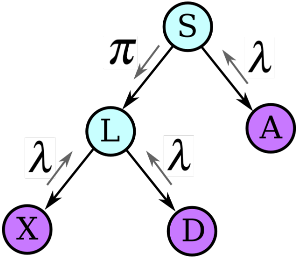
Let's now consider another example in which $X = x_2$, $D = d_3$ and $A = a_1$, and both L and S are unknown.

Evidence for variable L now comes all the way from A.



Conditioning

We do not know whether the patient is a smoker or not, but we do have evidence from node A and from the prior probability of S:



We write the total evidence for S excluding any evidence from L as:

$$\pi_L(S) = P(S)\lambda_A(S)$$

We call this the π message from S to L.

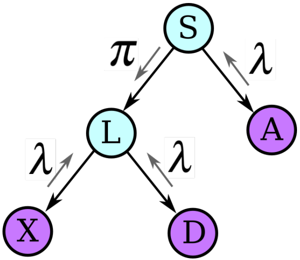
Conditioning Example

$$\text{Given } \boldsymbol{P}(\mathbf{S}) = (0.6, 0.2, 0.2)$$
$$\text{and } \lambda_{\boldsymbol{A}}(\mathbf{S}) = (0.8, 0.2, 0.05)$$

$$\pi_{\boldsymbol{L}}(\mathbf{S}) = (0.48, 0.04, 0.01)$$

The π message from S to L is sent (turned into evidence for L) by multiplying by the conditional probability matrix:

$$\pi(\boldsymbol{L}) = \mathbf{P}(\mathbf{L}|\mathbf{S})\pi_{\boldsymbol{L}}(\mathbf{S})$$



Conditioning Example

Writing out the equation in full we get:

$$\boldsymbol{\pi(L)} = \begin{bmatrix} P(l_1|s_1) & P(l_1|s_2) & P(l_1|s_3) \\ P(l_2|s_1) & P(l_2|s_2) & P(l_2|s_3) \\ P(l_3|s_1) & P(l_3|s_2) & P(l_3|s_2) \end{bmatrix} \begin{bmatrix} 0.48 \\ 0.04 \\ 0.01 \end{bmatrix}$$

so:

$$\pi(l_1) = 0.48 \times P(l_1|s_1) + 0.04 \times P(l_1|s_2) + 0.01 \times P(l_1|s_3)$$

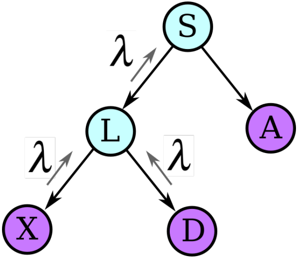
$$\pi(l_2) = 0.48 \times P(l_2|s_1) + 0.04 \times P(l_2|s_2) + 0.01 \times P(l_2|s_3)$$

$$\pi(l_3) = 0.48 \times P(l_3|s_1) + 0.04 \times P(l_3|s_2) + 0.01 \times P(l_3|s_3)$$

The π evidence is a weighted sum of the entries in the link matrix.

What about the evidence from X and D?

We have already seen how the lambda evidence is collected by L from S and D.



This is information about node L, and clearly will have some effect on our belief in its parent S.

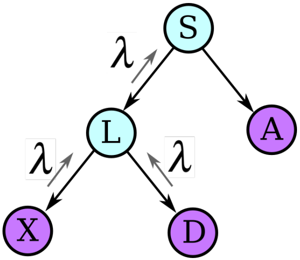
However, we don't have a definite value for L, only a probability distribution over its states.

Conditioning again

Let's suppose that:

$$\boldsymbol{\lambda}(\boldsymbol{L}) = (0.5, 0.2, 0.1)$$

This is all the evidence we need to send to S - it already has the evidence from $\boldsymbol{\pi}(\boldsymbol{L})$. We send it by pre-multiplying the conditional probability matrix by $\lambda(L)$



$$\boldsymbol{\lambda_L}(\mathbf{S}) = \boldsymbol{\lambda}(\boldsymbol{L})\mathbf{P}(\mathbf{L}|\mathbf{S})$$

Conditioning Example

Writing out the equation in full we get:

$$\boldsymbol{\lambda_L(S)} = [0.5, 0.2, 0.1] \begin{bmatrix} P(l_1|s_1) & P(l_1|s_2) & P(l_1|s_3) \\ P(l_2|s_1) & P(l_2|s_2) & P(l_2|s_3) \\ P(l_3|s_1) & P(l_3|s_2) & P(l_3|s_2) \end{bmatrix}$$

so:

$$\lambda_L(s_1) = 0.5 \times P(l_1|s_1) + 0.02 \times P(l_2|s_1) + 0.01 \times P(l_3|s_1)$$

$$\lambda_L(s_2) = 0.5 \times P(l_1|s_2) + 0.02 \times P(l_2|s_2) + 0.01 \times P(l_3|s_2)$$

$$\lambda_L(s_3) = 0.5 \times P(l_1|s_3) + 0.02 \times P(l_2|s_3) + 0.01 \times P(l_3|s_3)$$

The λ evidence is a weighted sum of the entries in the link matrix.

Final Probability Distributions

Having propagated the evidence from all three instantiated nodes X, D and A we can now calculate the probability distributions over the unknown nodes:

$$\boldsymbol{P'(L) = \alpha \times \pi(L) \times \lambda_X(L) \times \lambda_D(L)}$$

$$\boldsymbol{P'(S) = \alpha \times P(S) \times \lambda_L(S) \times \lambda_A(S)}$$

And we can now use these distributions to make inferences.

Evidence

- Every node in the graph will have a vector of λ evidence and a vector of π evidence.
- The evidence is unnormalised probability. If normalised the λ evidence would be the probability of the node given all the information comming from its descendents.
- If normalised the π evidence would become the probability of the node given all the information comming from its parents and their ancestors and descendents.

Instantiation and Evidence

- If a node is instantiated (measured) it is known to be in a particular state and its λ evidence and π evidence will contain 1 for the instantiated state and zero for all others.

$$\lambda(X) = (0, 0, 1, 0)$$

- A node can have virtual evidence. If there is uncertainty about a particular measurement we can express that by specifying a liklihood distribution over the states:

$$\lambda(X) = (0.1, 0.2, 0.5, 0.2)$$

- If a node has no evidence its states are equiprobable:

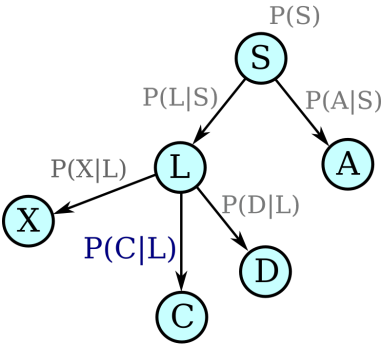
$$\lambda(X) = (1, 1, 1, 1)$$

Incorporating more Nodes

- One of the best features of Bayesian Networks is that we can incorporate new nodes as the data becomes available.
- Suppose we discover that coughing (C) is a symptom of lung cancer.
- This could simply be treated as another node.

Incorporating more Nodes

If we add this new node we only need to find one new conditional probability matrix. All the others remain unchanged.



Summary

Given a set of n discrete random variables $\mathbf{V} = (V_1, V_2 \cdots V_n)$ and a large data sets we can:

- build an affinity matrix from the data set and deduce a dependency graph.
- with the help of an expert we can add causal directions to the graph.
- use our data set to find conditional probability matrices linking nodes to their parents.
- for any measured subset of the variables we can find probability distributions over the states of all the others.

An impressive achievement you will agree - **but !!**

Summary

Just as I was on my way to the pub to celebrate, I met my friendly medical expert again. She said to me:

> *"Haven't you heard? Those brilliant microbiologists have just discovered that cancer is more likely to be caused by a genetic defect than by smoking"*

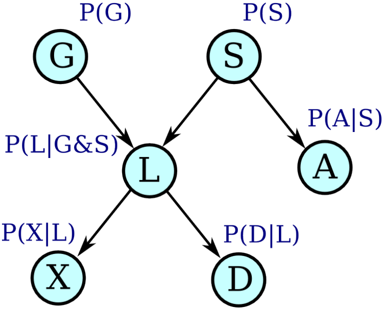
So now it's back to the drawing board . . . again!!!

Multiple Parents

- Up until now our networks have been trees, but in general they need not be. In particular, we need to cope with the possibility of multiple parents.
- Multiple parents can be thought of as representing different possible causes of an outcome.
- In our example we want to express the idea that lung cancer can be caused either by a genetic defect or by smoking.

Multiple Parents

We can incorporate a new two state node G in the graph which indicates a genetic defect.



Notice that we need to replace the link matrix $P(L|S)$ by a new matrix $P(L|G\&S)$.

Multiple Parents

Conditional probabilities, with multiple parents must include all the joint states of the parents. Thus for the $P(L|G\&S)$ node we will have:

$$\begin{bmatrix} P(l_1|g_1\&s_1) & P(l_1|g_1\&s_2) & P(l_1|g_1\&s_3) & P(l_1|g_2\&s_1) & P(l_1|g_2\&s_2) & P(l_1|g_2\&s_3) \\ P(l_2|g_1\&s_1) & P(l_2|g_1\&s_2) & P(l_2|g_1\&s_3) & P(l_2|g_2\&s_1) & P(l_2|g_2\&s_2) & P(l_2|g_2\&s_3) \\ P(l_3|g_1\&s_1) & P(l_3|g_1\&s_2) & P(l_3|g_1\&s_3) & P(l_3|g_2\&s_1) & P(l_3|g_2\&s_2) & P(l_3|g_2\&s_3) \end{bmatrix}$$

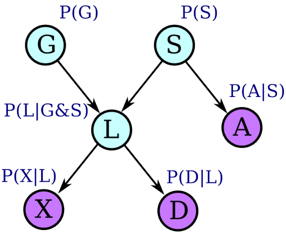
As before each row is for one state of the child node. This time we have one column for each joint state of the parents.

π messages with multiple parents

As before we can calculate a π message from each parent to node L

$$\pi_L(G) = P(G)$$

$$\pi_L(S) = P(S) \times \lambda_A(S)$$



But now we need to pass the messages using the joint conditional probability matrix.

Finding a joint distribution over the parents

We calculate the joint distribution of π messages over the states of the parents by assuming that $\pi_L(G)$ and $\pi_L(S)$ are independent:

$$\pi_L(G\&S) = \pi_L(G) \times \pi_L(S)$$

Remember that this is a scalar equation with variables G and S. In vector form the joint evidence is:

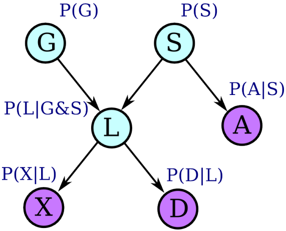
$$\boldsymbol{\pi_L(G\&S)} = \begin{bmatrix} \pi_L(g_1\&s_1) \\ \pi_L(g_1\&s_2) \\ \pi_L(g_1\&s_3) \\ \pi_L(g_2\&s_1) \\ \pi_L(g_2\&s_2) \\ \pi_L(g_2\&s_3) \end{bmatrix} = \begin{bmatrix} \pi_L(g_1) \times \pi_L(s_1) \\ \pi_L(g_1) \times \pi_L(s_2) \\ \pi_L(g_1) \times \pi_L(s_3) \\ \pi_L(g_2) \times \pi_L(s_1) \\ \pi_L(g_2) \times \pi_L(s_2) \\ \pi_L(g_2) \times \pi_L(s_3) \end{bmatrix}$$

The independence of $\pi_L(G)$ and $\pi_L(S)$

- In this simple example G and S have no other path linking them, and hence if we do not consider the evidence from L then $\pi_L(G)$ and $\pi_L(S)$ must be independent.
- If they had, for example, a common parent, then our assumption about the independence of $\pi_L(G)$ and $\pi_L(S)$ would no longer hold.
- We have therefore made an implicit assumption that there are no loops in our network.

π messages with multiple parents

We can now complete sending the π message to L with the same vector equation we used before:



$$\pi(\boldsymbol{L}) = \mathbf{P}(\mathbf{L}|\mathbf{G}\&\mathbf{S})\pi_{\boldsymbol{L}}(\boldsymbol{G}\&\boldsymbol{S})$$

and as before:

$$\boldsymbol{P}'(\boldsymbol{L}) = \alpha \times \pi(\boldsymbol{L}) \times \lambda_{\boldsymbol{X}}(\boldsymbol{L}) \times \lambda_{\boldsymbol{D}}(\boldsymbol{L})$$

λ messages with multiple parents

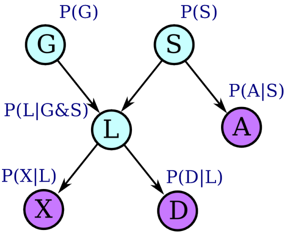
A child with multiple parents sends a message to the joint states of its parents.

For a single parent:

$$\boldsymbol{\lambda_A(S) = \lambda(A)}\mathbf{P(A|S)}$$

For multiple parents:

$$\boldsymbol{\lambda_L(G\&S) = \lambda(L)}\mathbf{P(L|G\&S)}$$



λ messages with multiple parents

The joint message $\lambda_L(G\&S)$ that we have calculated is:

$$[\lambda(g_1\&s_1), \lambda(g_1\&s_2), \lambda(g_1\&s_3), \lambda(g_2\&s_1), \lambda(g_2\&s_2), \lambda(g_2\&s_3)]$$

The evidence we have for G and S excluding any evidence from L are the π messages that we calculated. We use these to condition the joint λ message.

$$\lambda_L(g_1) = \pi_L(s_1)\lambda_L(g_1\&s_1) + \pi_L(s_2)\lambda_L(g_1\&s_2) + \pi_L(s_3)\lambda_L(g_1\&s_3)$$
$$\lambda_L(g_2) = \pi_L(s_1)\lambda_L(g_2\&s_1) + \pi_L(s_2)\lambda_L(g_2\&s_2) + \pi_L(s_3)\lambda_L(g_2\&s_3)$$

λ messages with multiple parents

In more general terms we can write the individual λ messages to the two parents as:

$$\lambda_L(g_i) = \sum_j (\pi_L(s_j) \times \lambda_L(g_i \& s_j))$$

$$\lambda_L(s_k) = \sum_h (\pi_L(g_h) \times \lambda_L(g_h \& s_k))$$

Notice that if the joint λ message contains no evidence ie $\lambda_L(g_i \& s_j) = 1$ for all i, j, then these individual λ messages contain no evidence.

$$\lambda_L(g_i) = \sum_j \pi_L(s_j)$$

$$\lambda_L(s_k) = \sum_h \pi_L(g_h)$$

Summary

Well at last it looks like we have a completely general purpose belief propagation system for causal networks based on just five equations:

- The π message:

$$\pi_C(a_i) = \begin{cases} 1 & \text{if } A \text{ is instantiated for } a_i \\ 0 & \text{if } A \text{ is instantiated but not for } a_i \\ P'(a_i)/\lambda_C(a_i) & \text{if } A \text{ is not instantiated} \end{cases}$$

- The π evidence:

  For a single parent: $\boldsymbol{\pi}(\boldsymbol{L}) = \mathbf{P}(\mathbf{L}|\mathbf{S})\boldsymbol{\pi}_{\boldsymbol{L}}(\boldsymbol{S})$

  For multiple parents: $\boldsymbol{\pi}(\boldsymbol{L}) = \mathbf{P}(\mathbf{L}|\mathbf{G}\&\mathbf{S})\boldsymbol{\pi}_{\boldsymbol{L}}(\boldsymbol{G}\&\boldsymbol{S})$

  (where $\pi_L(G\&S) = \pi_L(G) \times \pi_L(S)$)

Summary

- The λ evidence:

$$\lambda(l_k) = \begin{cases} 1 & \text{if } L \text{ is not instantiated and has received no } \lambda \text{ messages} \\ 1 & \text{if } L \text{ is instantiated as } l_k \\ 0 & \text{if } L \text{ is instantiated but not as } l_k \\ \prod_i \lambda_{D_i}(c_k) & \text{if } L \text{ is not instantiated} \end{cases}$$

- The λ message:

  For a single parent: $\boldsymbol{\lambda_A(S) = \lambda(A)P(A|S)}$

  For multiple parents: $\boldsymbol{\lambda_L(G\&S) = \lambda(L)P(L|G\&S)}$

  (where $\lambda_L(s_k) = \sum_h (\pi_L(g_h) \times \lambda_L(g_h \& s_k)))$

- The posterior probability: $P'(L) = \alpha \times \lambda(L) \times \pi(L)$

Summary

Well that seems to be a satisfactory conclusion to everything - but?? - isn't that my friendly medical expert sitting in the back of the lecture theatre? What is she asking now?

> "Didn't you know that patients with an aortic anuerism often suffer from dypsnoea?

Oh No !!! - that's really put the cat among the pidgeons!

Find out why next lecture.