Lecture 2

Simple Bayesian Networks

Limitations of Simple Bayesian Inference

- Simple Bayesian inference is inadequate to deal with more complex models of prior knowledge.
- Once several factors affect a decision they need to be combined somehow

The Catness Measure again

$$Catness = |(R_l - R_r)/R_r| + |(S_i - 2 \times (R_l + R_r))/R_r|$$

- We are currently weighting the two error terms equally, but perhaps this is not a good idea.
- Moreover we may want new terms, for example fur colour around the putative eyes.

A more complex Catness measure???

$$Catness = \alpha|(R_l - R_r)/R_r| + \beta|(S_i - 2\ (R_l + R_r))/R_r| + \gamma(ColourMatch) + \&c.$$

- α, β and γ are constants to be determined.
- The whole process becomes very heuristic and we need to look for better methods for representing our prior models.
- One approach is to go to a Bayesian Network.

More Evidence

- When we used Bayes theorem in lecture 1 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:

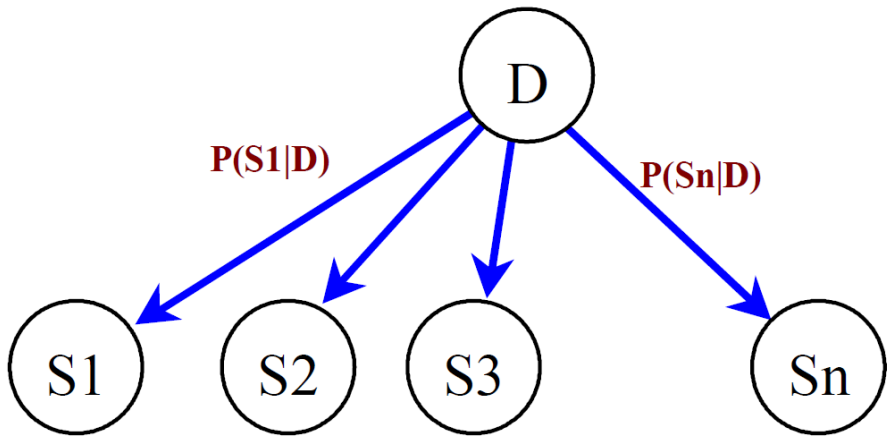
$$P(S_1 \& S_2 \& \cdots \& S_n)$$

and so Bayes theorem becomes:

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

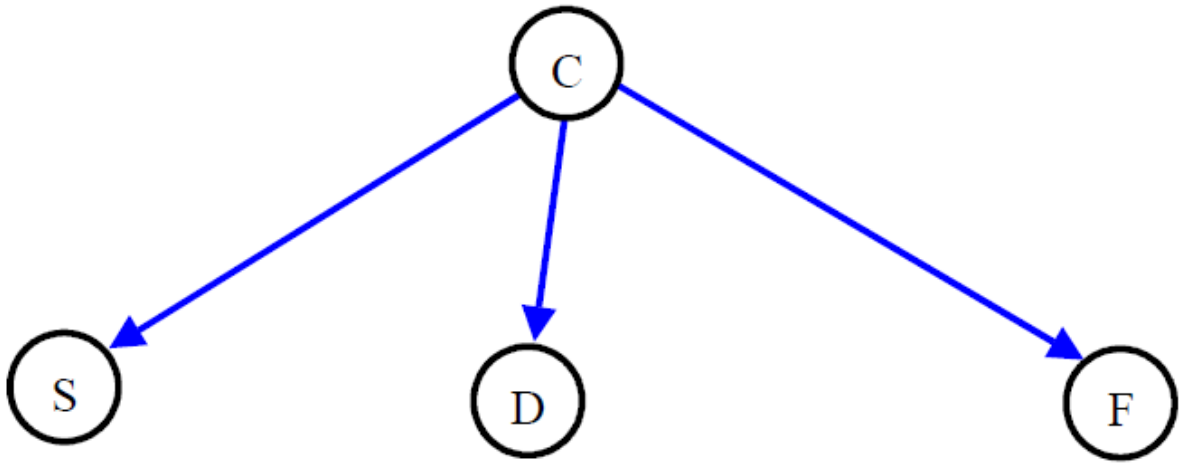
Graphical Notation

- 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.

Let's clarify the notation with the cat example



Variable	Interpretation	Type	Value
C	Cat	Discrete (2 states)	True or False
S	Eye separation	Continuous	$S = (S_i - 2 * (R_l + R_r))/R_r$
D	Eye difference	Continuous	$\lvert(R_l - R_r)/R_r\rvert$
F	Fur colour	Discrete (20 states)	Histogram of pixel hues

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

We will come back to continuous variables later, but for the first part of the course we will consider mainly discrete variables.

Quantisation

- To use continuous variables in a discrete framework we must quantise them.
- For example in practice the *Eye separation* variable might vary between -1.5 (eyes very close) to 1.5 (eyes very far apart). We could therefore quantise it to have a reasonable resolution in the range of interest:

$$[below - 1.5][-1.5 \cdots -0.75][-0.75 \cdots -0.25]$$
$$[-0.25 \cdots 0.25][0.25 \cdots 0.75][0.75 \cdots 1.5][above 1.5]$$

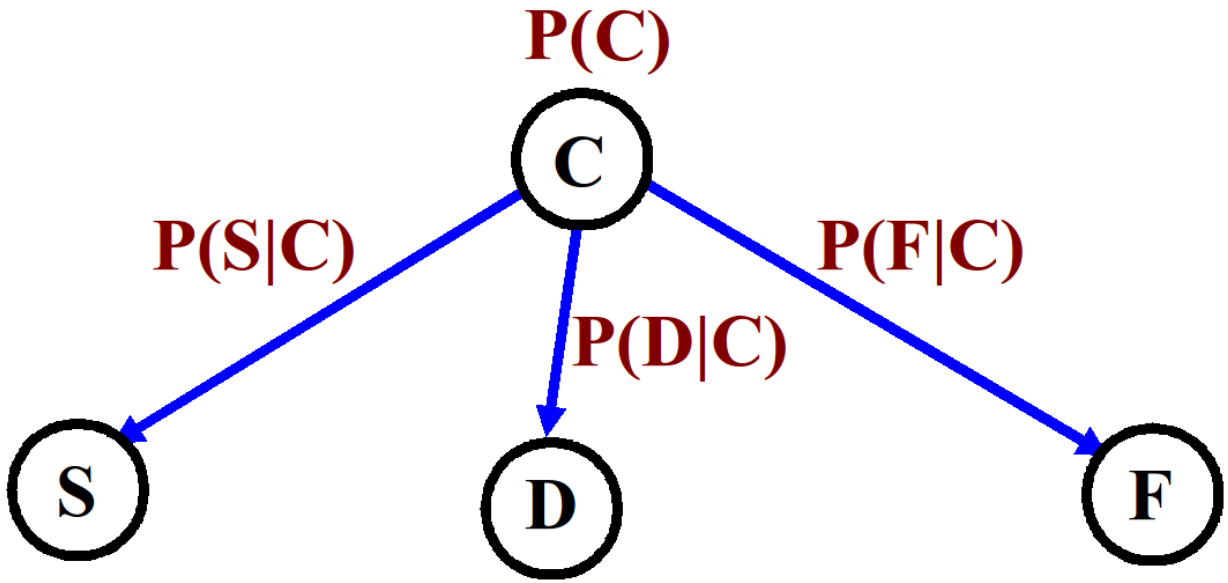
Some Observations about Quantisastion

- We can quantise variables in a large number of ways, and indeed this forms an important area of research.

- An important class of methods uses data and information theory to try to obtain the best representation of the variable in as few states as possible

- For our purposes we will simply use heuristic methods.

This is what our simple network now looks like



Variable	Interpretation	Type	Value
C	Cat	Discrete (2 states)	c_1, c_2
S	Eye separation	Discrete (7 states)	$s_1, s_2, s_3, s_4, s_5, s_6, s_7$
D	Eye difference	Discrete (4 states)	d_1, d_2, d_3, d_4
F	Fur colour	Discrete (20 states)	$f_1, f_2 \cdots f_{20}$

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:

$$\boldsymbol{P}(\boldsymbol{D}|\boldsymbol{C}) = \begin{bmatrix} P(d_1|c_1) & P(d_1|c_2) \\ P(d_2|c_1) & P(d_2|c_2) \\ P(d_3|c_1) & P(d_3|c_2) \\ P(d_4|c_1) & P(d_4|c_2) \end{bmatrix}$$

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

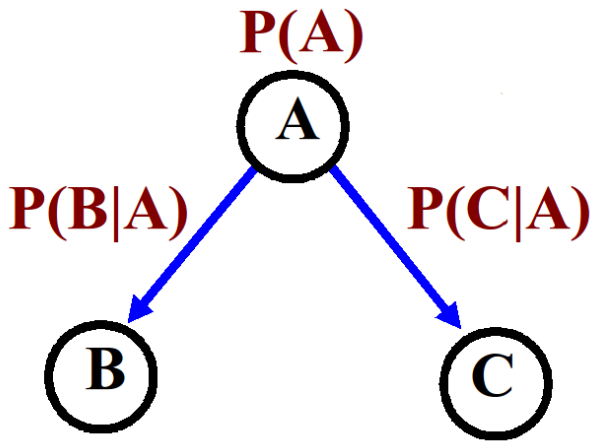
- The root nodes of a network do not have any parents. Instead of a link matrix they have a vector giving the prior probabilities of the states, eg:

$$\boldsymbol{P}(\boldsymbol{C}) = [P(c_1), P(c_2)]$$

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

Problem Break

Given the simple network:



If each node has two states: $[a_1, a_2], [b_1, b_2], [c_1, c_2]$ find an expression for $P(a_1|b_2\&c_1)$ in terms of the probabilities found in the link matrices.

Hint: Use Bayes' Theorem and the fact that:

$$P(a_1|b_2\&c_1) + P(a_2|b_2\&c_1) = 1$$

Solution

$$P(a_1|b_2\&c_1) + P(a_2|b_2\&c_1) = 1$$

Apply Bayes' theorem (with the independence assumption):

$$\frac{P(a_1)P(b_2|a_1)P(c_1|a_1)}{P(b_2\&c_1)} + \frac{P(a_2)P(b_2|a_2)P(c_1|a_2)}{P(b_2\&c_1)} = 1$$

$$P(b_2\&c_1) = P(a_1)P(b_2|a_1)P(c_1|a_1) + P(a_2)P(b_2|a_2)P(c_1|a_2)$$

thus:

$$P(a_1|b_2\&c_1) = \frac{P(a_1)P(b_2|a_1)P(c_1|a_1)}{P(a_1)P(b_2|a_1)P(c_1|a_1) + P(a_2)P(b_2|a_2)P(c_1|a_2)}$$

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, in our problem we might process many real pictures for the leaf nodes, S, D and F, and get expert advice on the state of C.

Suppose that there are $N(c_2 \& d_4)$ points in our data set where variable D is in state d_4 and variable C is in state $c2$. Then we can calculate $P(d_4|c_2)$ as folows:

$$P(d_4|c_2) = N(c_2 \& d_4)/N(c_2)$$

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.
- Already the state space of our problem is quite big, and it grows exponentially.
- As the number of conditional probabilities 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.
- Clearly they give us a much more accurate way of expressing how each term in the catness measure relates to the presence of a cat.

Using a naive Bayesian network

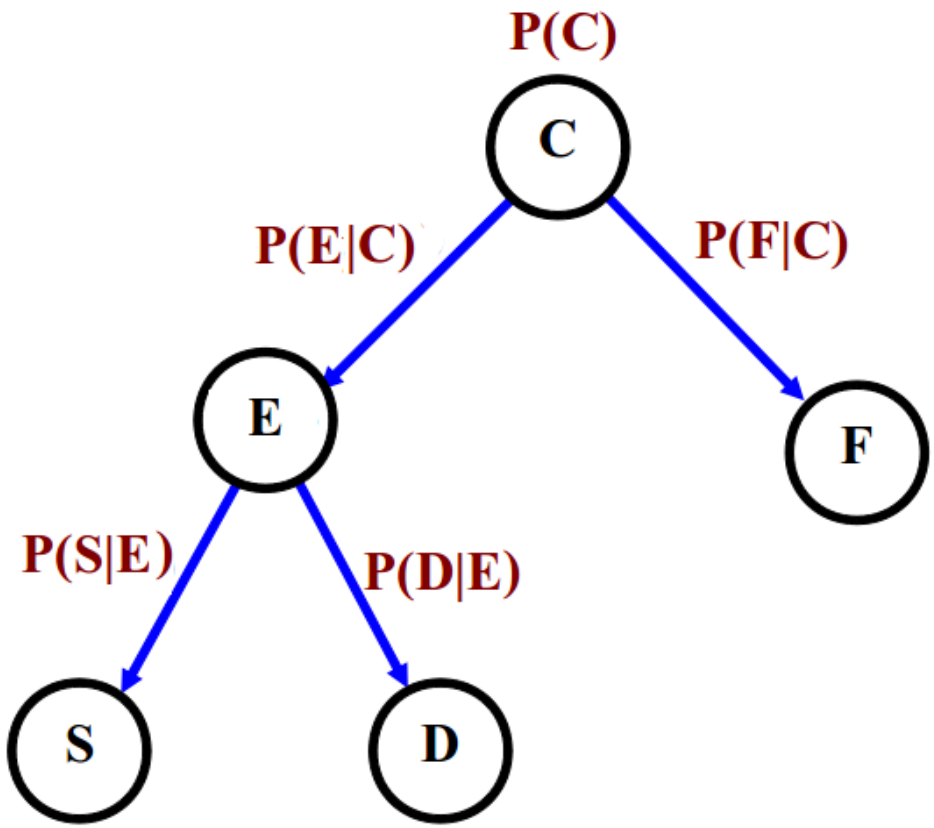
- Setting a node to a measured value is called instantiation.
- Once a node has been instantiated, we can look up the values for the conditional probabilities in the link matrices.
- Calculating the probabilities of the states of the hypothesis is done by multiplying together the conditional probabilities 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 a simple decision tree.
- Reasoning about our variables we could argue that, given there was a cat in the picture, the separation and the difference variables might not be completely conditionally independent.
- Thus we might refine our network into a more complex structure where we explicitly model the relation between them.

The Cat Decision Tree

The nodes *S* and *D* are related by a common cause - the presence of eyes - which we now include in the model. Eyes might be present but not caused by a cat.



Adding the Eyes 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:
  - e_1 interpreted as probably not eyes
  - e_2 interpreted as could be eyes
  - e_3 interpreted as probably eyes
- To estimate the link matrices we need expert advice on both the non terminal node E and the hypothesis node C.
- Using the network is a little more complex.

Calculating the probability of a cat

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

$$P(C|E\&F) = \alpha P(C)P(E|C)P(F|C)$$

- For the Eyes node we have:

$$P(E|\mathrm{S}\&D) = \alpha P(E)P(\mathrm{S}|E)P(D|E)$$

- But now we have a problem since we don't have a prior probability for E.

- We will see shortly that $P(E)$ is evidence sent from E's parent C.

The Likelihood of the Eyes

- From Bayes theorem (last slide) we have

$$P(E|S\&D) = \alpha P(E)P(S|E)P(D|E)$$

- Although we don't have any direct prior information $P(E)$ but we do have likelihood information about E from S and D:

$$L(E|S\&D) = P(S|E)P(D|E)$$

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

Calculating the Probability of a Cat

$$P(C|E\&F) = \alpha P(C)P(E|C)P(F|C)$$

- Suppose we measure F, say $F = f_5$, we can look up $P(f_5|c_1)$ and $P(f_5|c_2)$ directly from the link matrix.
- However, we don't have an instantiated value for E, but we have likelihood evidence over the states of E, and we can use this to calculate a weighted average of the conditional probabilities:

$$P(e|c_1) = P(e_1|c_1)L(e_1) + P(e_2|c_1)L(e_2) + P(e_3|c_1)L(e_3)$$
$$P(e|c_2) = P(e_1|c_2)L(e_1) + P(e_2|c_2)L(e_2) + P(e_3|c_2)L(e_3)$$

Calculating the Probability of a Cat

$$P(C|E\&F) = \alpha P(C)P(E|C)P(F|C)$$

Substituting in the instantiated value for F and the likelihood evidence for E calculated from S and D we get:

$$\begin{aligned} P(c_1|e\&f_5) = \quad & \alpha \times P(c_1) \times \\ & \{P(e_1|c_1)L(e_1) + P(e_2|c_1)L(e_2) + P(e_3|c_1)L(e_3)\} \\ & \times P(f_5|c_1) \end{aligned}$$

$$\begin{aligned} P(c_2|e\&f_5) = \quad & \alpha \times P(c_2) \times \\ & \{P(e_1|c_2)L(e_1) + P(e_2|c_2)L(e_2) + P(e_3|c_2)L(e_3)\} \\ & \times P(f_5|c_2) \end{aligned}$$

Posterior Probabilities

- The term posterior probability, indicated P', is used to indicate the probability calculated after some instantiation.
- For our previous example we write $P'(C)$ instead of $P(C|E\&F)$
- It is called *posterior* because it is the probability *after* some evidence has been collected.

Calculating the Probability of Eyes

- We already have the likelihood evidence from S and D for the states of the Eyes node.
- Although we don't have a Prior probability for it we have evidence that comes from its parent C.
- However, this is not from the posterior probability $P'(C)$, but rather from the evidence for C excluding anything that came from E itself. We write (for the time being):

$$P_E(C) = \alpha P(C)P(F|C)$$

Passing Evidence from C to E

- Let's suppose we calculate $P_E(C)$ as $[0.4, 0.6]$. We use the link matrix to calculate the equivalent of a prior probability for E:

$$\begin{bmatrix} P(e_1) \\ P(e_2) \\ P(e_3) \end{bmatrix} = \begin{bmatrix} P(e_1|c_1) & P(e_1|c_2) \\ P(e_2|c_1) & P(e_2|c_2) \\ P(e_3|c_1) & P(e_3|c_2) \end{bmatrix} \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} =$$

$$\begin{bmatrix} 0.4P(e_1|c_1) + 0.6P(e_1|c_2) \\ 0.4P(e_2|c_1) + 0.6P(e_2|c_2) \\ 0.4P(e_3|c_1) + 0.6P(e_3|c_2) \end{bmatrix}$$

- Note that this is a probability distribution since the columns of the link matrix sum to 1

Finally E can be Calculated

- Suppose we have instantiations $S = s_3$ and $D = d_2$, then we calculate $P'(E)$ as

$$P'(e_1) = \alpha P(e_1)P(s_3|e_1)P(d_2|e_1)$$
$$P'(e_2) = \alpha P(e_2)P(s_3|e_2)P(d_2|e_2)$$
$$P'(e_3) = \alpha P(e_3)P(s_3|e_3)P(d_2|e_3)$$

- and as usual α is calculated by normalisation.

$$P'(e_1) + P'(e_2) + P'(e_3) = 1$$

In Conclusion

- The whole process looks very complex, and this is just a tiny network.
- However, don't despair - every node in a big network is calculated in the same way.
- Next time we will start to develop a general method for calculating probabilities in any network which uses just 5 equations.