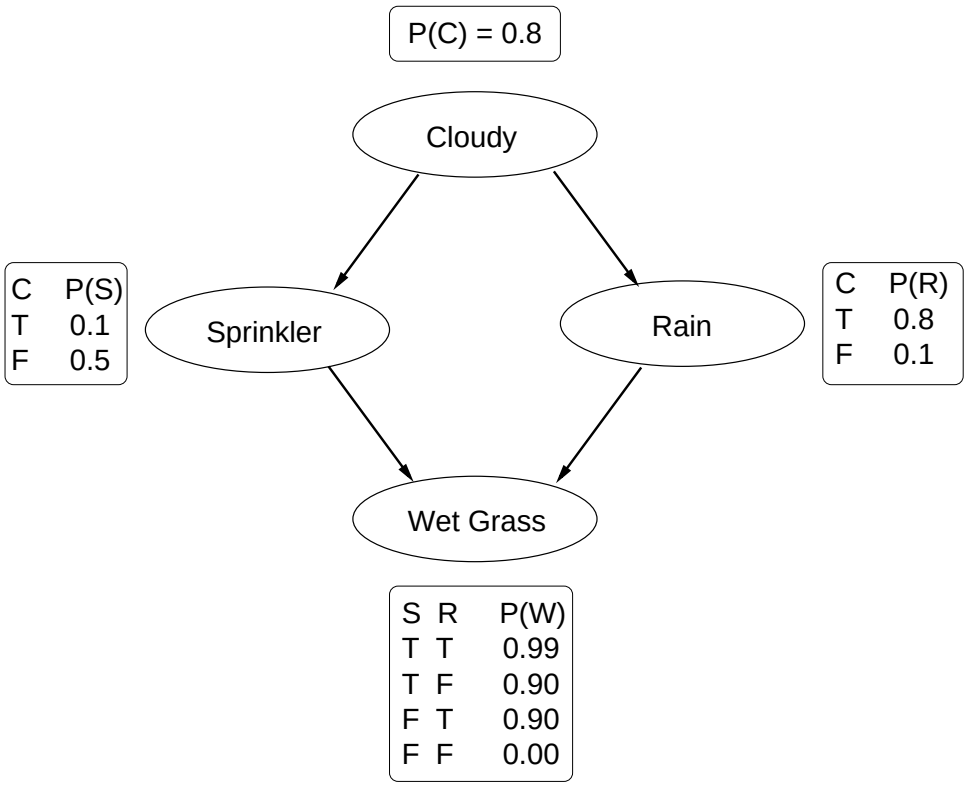
University of Cambridge
Engineering Part IIB & EIST Part II

Paper 4F10: Statistical Pattern Processing

Lectures 10 & 11: Graphical Models and Bayesian Networks



Mark Gales
mjfg@eng.cam.ac.uk
October 2005

Graphical Models

Graphical models have their origin in several areas of research. They are a union of *graph* theory and *probability* theory. They are a useful framework for representing, reasoning with and learning complex problems.

The techniques are useful for multivariate (multiple variable) probabilistic systems and encompass many standard schemes, for example

- mixture models;
- factor analysis;
- hidden Markov models;
- Kalman filters

This, and the next, lecture will look at the basics of graphical models. In particular a specific form of graphical model, *Bayesian networks* will be examined.

Basic Probability Revisited (again)

These lectures will make extensive use of the following standard probability concepts:

- discrete random variables: one of a possible set of events occurs (e.g. rolling a die). Associated *probability mass function* satisfies

$$\sum_A P(A) = 1; \quad P(A) \geq 0$$

- continuous random variables: the RV can take any value within a, possibly, infinite, range. Associated *probability density function* satisfies

$$\int p(A)dA = 1; \quad p(A) \geq 0$$

- joint probability $P(A, B)$
- joint independence $P(A, B) = P(A)P(B)$
- conditional probability $P(A|B)$
- Bayes' rule

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

- marginal probability

$$P(A) = \sum_B P(A, B)$$

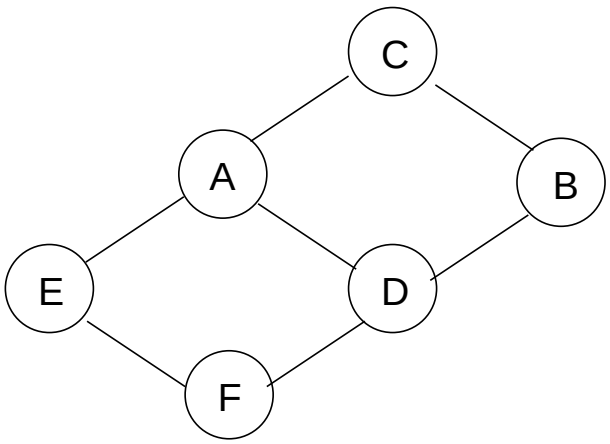
where the summation is over all possible values of B.

Basic Notation

A *graph* consists of a collection of *nodes* and *edges*.

- Nodes, or vertices, are usually associated with the variables (some of which may not be observed, or latent).
- Edges connect nodes to one another.

The *absence* of an edge between two nodes indicates conditional independence. The graphical model can be considered as representing dependencies in the system. This will discussed in more detail later.



Here there are 6 nodes, $\{A, B, C, D, E, F\}$ and 7 edges.

Further Notation

In this work we will consider collections of nodes. So using the graph from the previous slide as an example, we can consider

$$\mathcal{C}_1 = \{A, B, C, D\}; \quad \mathcal{C}_2 = \{A, D, E, F\}$$

Various operations can then be performed. For example

- the *union* of two sets

$$\mathcal{S} = \mathcal{C}_1 \cup \mathcal{C}_2 = \{A, B, C, D, E, F\}$$

- the *intersection* of the two sets

$$\mathcal{S} = \mathcal{C}_1 \cap \mathcal{C}_2 = \{A, D\}$$

- removing elements from a set

$$\mathcal{C}_1 \setminus \mathcal{S} = \{B, C\}$$

Conditional Independence

One of the fundamental aspects of graphical models is the concept of *conditional independence*.
Consider three variables, A, B and C. We can write

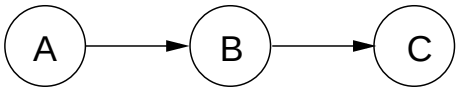
$$P(A, B, C) = P(A)P(B|A)P(C|B, A)$$

If C is conditionally independent of A given B, then we can write

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

The value of A does not affect the distribution of C if B is known.

Graphically this can be described as



Conditional independence is very important when modelling highly complex systems.

Consider the case above where each of A, B and C can take one of 3 values, $\{-1, 0, 1\}$. Modelling the complete joint distribution requires $3^3 - 1 = 26$ parameters. Using the conditional independence above requires $3^2 + 3^2 - 2 = 16$ parameters.

Bayesian Networks

In these lectures only *Bayesian networks* will be considered. They are a special case of graphical models, *directed acyclic graphs* (DAGs):

- directed: all connections have arrows associated with them;
- acyclic: following the arrows around it is not possible to complete a loop

The main problems with BNs are:

- inference (from observations ‘‘it’s cloudy’’ infer the probability of the wet grass).
- training the models;
- determining the structure of the network (i.e. what is connected to what)

The first two issues will be addressed in these lectures.

The final problem of determining the appropriate structure is an area of on-going research.

Observed and Unobserved Variables

In general the variables (nodes) may be split into two groups:

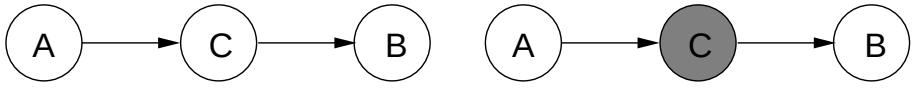
- observed variables are the ones we have knowledge about.
- unobserved variables are ones we don't know about and therefore have to infer the probability.

We need to find efficient algorithms that allow rapid inference to be made. Preferably a very general scheme that will allow inference over any Bayesian network.

First three basic structures are described and the effects of observing one of the variables on them are described.

Standard Structures

- Structure 1



If C is not observed

$$P(A,B) = \sum_C P(A,B,C) = P(A) \sum_C P(C|A)P(B|C)$$

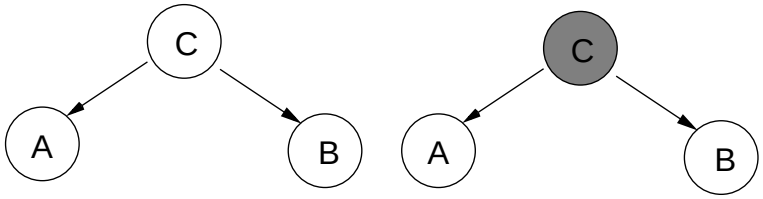
then A and B are dependent on each other.

If C is observed (indicated by the shading) then

$$P(A,B|C) \;=\; P(A)P(B|C)$$

A and B are then independent. The path is sometimes called *blocked*.

- Structure 2



If C is not observed

$$P(A,B) = \sum_C P(A,B,C) = \sum_C P(C)P(A|C)P(B|C)$$

then A and B are dependent on each other.

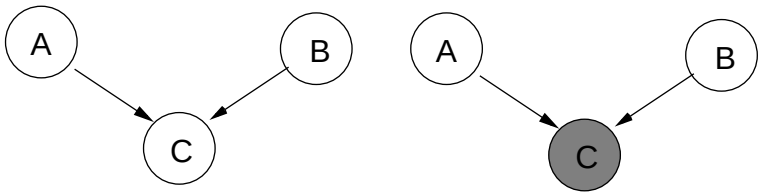
If C is observed then

$$P(A,B|C) \;=\; P(A|C)P(B|C)$$

A and B are then independent.

Standard Structures (cont)

- Structure 3



If C is not observed

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

A and B are independent of each other.

If C is observed

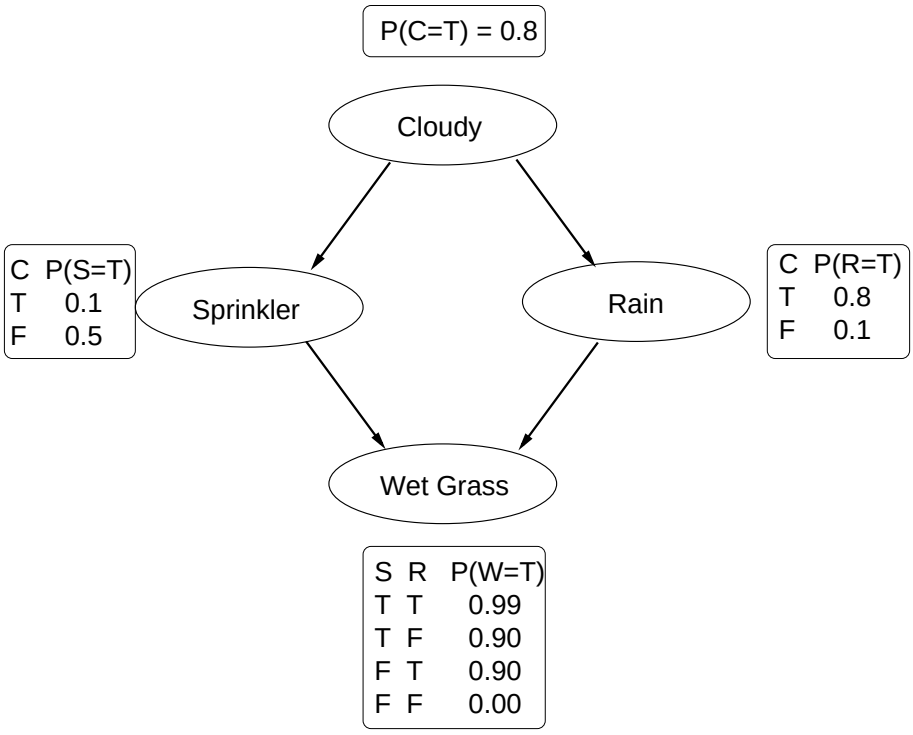
$$\begin{aligned} P(A,B|C) &= \frac{P(A,B,C)}{P(C)} \\ &= \frac{P(C|A,B)P(A)P(B)}{P(C)} \end{aligned}$$

A and B are not independent of each other if C is observed.

This phenomenom that two varibles are dependent if a common child is observed is sometimes called explaining away

Simple Example

Consider the following Bayesian network



Whether the grass is wet, W, depends on whether the sprinkler has been used, S, or whether it has rained, R. Whether the sprinkler is used depends on whether it is cloudy, similarly for whether it has rained.

The probability of the grass being wet is conditionally independent of it being cloudy, given information about the sprinklers and whether it has rained. This joint probability may be expressed as

$$P(C, S, R, W) = P(C)P(S|C)P(R|C)P(W|S, R)$$

Inference Example

The basic task is given some *observation* infer the probability of an event. So a question may be

> `It is cloudy, what's the probability that the grass is wet?`

- so we want to compute $P(W = \mathtt{T}|C = \mathtt{T})$. (Note: for simplicity of notation $P(W_\mathtt{T}|C_\mathtt{T})$ will be used for $P(W = \mathtt{T}|C = \mathtt{T})$.)

Re-expressing this request in terms of the joint probability

$$P(W_\mathtt{T}|C_\mathtt{T}) = \frac{P(W_\mathtt{T}, C_\mathtt{T})}{P(C_\mathtt{T})}$$

The denominator is known (0.8). The numerator may expressed as a marginal distribution

$$\begin{aligned} P(W_\mathtt{T}, C_\mathtt{T}) &= \sum_S \sum_R P(W_\mathtt{T}, S, R, C_\mathtt{T}) \\ &= \sum_S \sum_R P(W_\mathtt{T}|S, R) P(S|C_\mathtt{T}) P(R|C_\mathtt{T}) P(C_\mathtt{T}) \end{aligned}$$

where the summation are over the variable being `T`, or `F`. From the simple example this is (note $P(C_\mathtt{T})$ has simply been cancelled from the numerator and denominator)

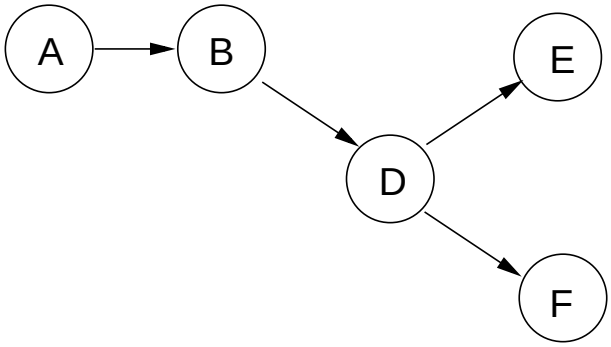
$$\begin{aligned} P(W_\mathtt{T}|C_\mathtt{T}) &= 0.99 \times 0.1 \times 0.8 + 0.90 \times 0.1 \times 0.2 \\ &\quad +0.90 \times 0.9 \times 0.8 + 0.00 \times 0.9 \times 0.2 \\ &= 0.7452 \end{aligned}$$

Message Passing for Trees

Rather than using the standard inference for the `wet grass` example, it is useful to introduce *message passing*. The aim is to convert the standard inference process into:

- local calculation on connected nodes;
- message passing between nodes;

To initially simplify the process only *trees* will be considered. Here each node has only one *parent*.



An example tree is given above.

For this example consider calculating $P(F)$. This can be written as a marginal distribution. First we need to chose the order to do the summations in. Pick (A, B, E, D).

$$\begin{aligned} P(F) &= \sum_D \sum_E \sum_B \sum_A P(A, B, D, E, F) \\ &= \sum_D P(F|D) \sum_E P(E|D) \sum_B P(D|B) \left(\sum_A P(A)P(B|A) \right) \end{aligned}$$

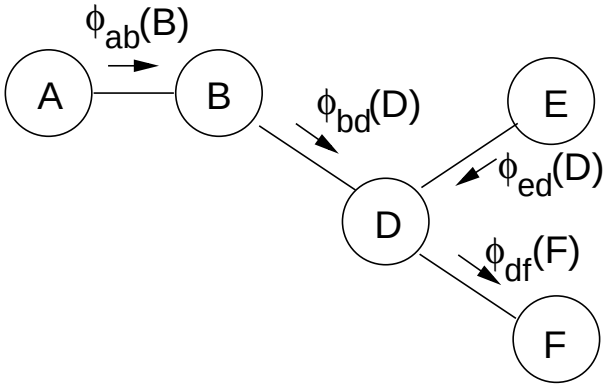
Single Variable Messages

The marginal probability can be calculated in terms of *messages* being passed between nodes, where the message consists of information about the parent.

The message from node I to node J is written as $\phi_{ij}(J)$. The marginal probability can be written as:

$$\begin{aligned} P(F) &= \sum_D P(F|D) \sum_E P(E|D) \left(\sum_B P(D|B)\phi_{ab}(B) \right) \\ &= \sum_D P(F|D) \left(\sum_E P(E|D)\phi_{bd}(D) \right) \\ &= \sum_D P(F|D)\phi_{bd}(D) \left(\sum_E P(E|D) \right) \\ &= \left(\sum_D P(F|D)\phi_{bd}(D)\phi_{ed}(D) \right) \\ &= \phi_{df}(F) \end{aligned}$$

This can be summarised as



Note: $\phi_{ed}(D) = \Sigma_E P(E|D) = 1$

Inference can now be viewed in terms of local computation and routing of messages.

Observed Variable

The message passing has not considered whether observations are observed or not. What happens if a variable is observed?

If the node is a parent to the node of interest, then *blocking* (see structure 1) will occur. If B is observed then the message is changed to

$$\phi_{bd}(D) = P(D|B = \mathtt{T})$$

(assuming that B was observed to be $\mathtt{T}$).

If the node is a child of the variable of interest this will also alter the probability. Let E be observed as $\mathtt{T}$. It is necessary to find

$$P(F|E = \mathtt{T}) = P(F, E = \mathtt{T})/P(E = \mathtt{T})$$

Thus all propagation the same, except

$$\phi_{ed}(D) = P(E = \mathtt{T}|D)$$

and a normalisation term is also required to be computed $P(E = \mathtt{T})$. This can be computed in a similar fashion to finding $P(F)$.

Message Passing

The general expression for a "forward" message is (for trees)

$$\phi_{ij}(J) = \sum_I P(J|I) \prod_{k \in \mathcal{N}_i \setminus j} \phi_{ki}(I)$$

where $\mathcal{N}_i \setminus j$ is the set of nodes that I is connected to excluding node J.

This can be checked by looking at the message from D to F. The *neighbours* of node D are

$$\mathcal{N}_d = \{B, E, F\}$$

So

$$\mathcal{N}_d \setminus F = \{B, E\}$$

The message is therefore

$$\phi_{df}(F) = \sum_D P(F|D) \left(\phi_{bd}(D)\phi_{ed}(D)\right)$$

This form of message passing is quite general, but only allows single marginal probabilities to be computed. A more general form would be useful.

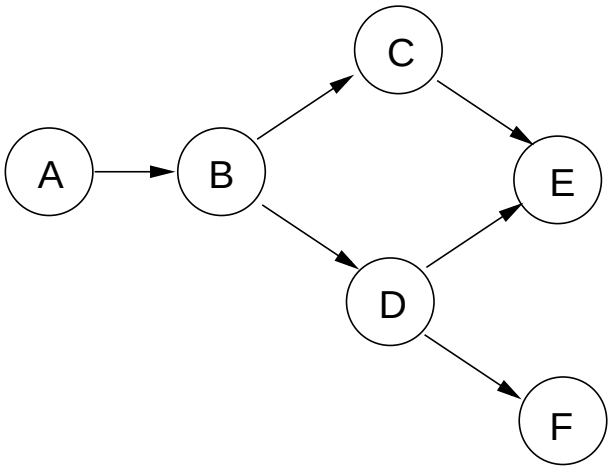
Cliques and Separators

Before defining the algorithm some basic terminology is required.

- Cliques $\mathcal{C}$: fully connected (every node is connected to every other node) subset of all the nodes.
- Separators $\mathcal{S}$: the subset of the nodes of a clique that are connected to nodes outside the clique.
- Neighbours $\mathcal{N}$: the set of neighbours for a particular clique.

Thus given the value of the separators for a clique it is conditionally independent of *all* other variables.

A simple example illustrates this concept.



An additional node C has been added this allows a brief discussion of *moral graphs*.

It is possible to use cliques and separators to generalise message passing to allow efficient inference with marginal distributions of multiple variables.

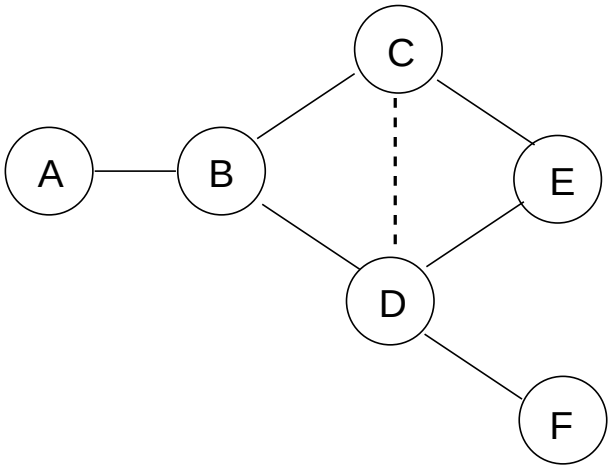
Moral Graphs

The first step in generating the cliques is to convert the DAG into a *moral graph*. This process involves the following steps:

1. connect the parents of each node,
2. remove the directions of the graph.

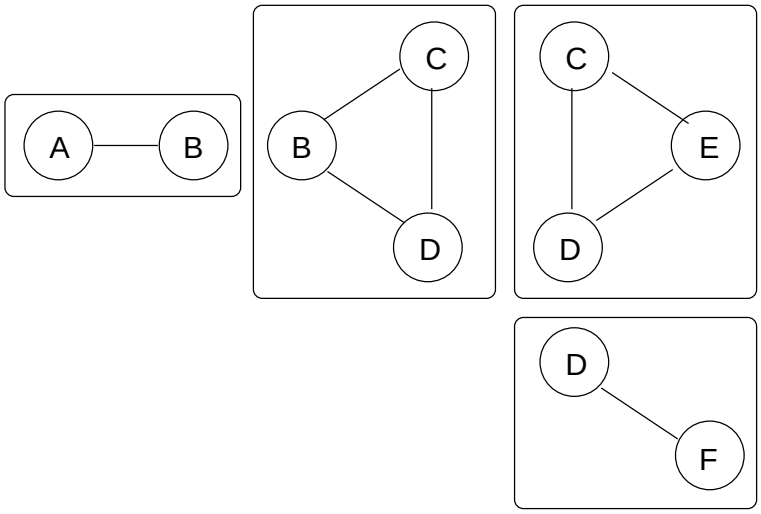
This yields an undirected graph.

An example of this is converting the example into a moral graph.



In step (1) the parents of `E`, `C` and `D` are now connected with the dotted line. In the second step the directions of the arrow are then removed.

Cliques and Separators (cont)



For this example the set of *cliques* are

$$\begin{aligned}\mathcal{C}_1 &= \{A, B\} \\ \mathcal{C}_2 &= \{B, C, D\} \\ \mathcal{C}_3 &= \{C, D, E\} \\ \mathcal{C}_4 &= \{D, F\}\end{aligned}$$

Each of these are fully connected subsets of all the variables. The separators are the *intersection* of the cliques.

$$\begin{aligned}\mathcal{S}_{12} &= \mathcal{C}_1 \cap \mathcal{C}_2 = \{B\} \\ \mathcal{S}_{23} &= \mathcal{C}_2 \cap \mathcal{C}_3 = \{C, D\} \\ \mathcal{S}_{13} &= \mathcal{C}_1 \cap \mathcal{C}_3 = \{\} \\ \mathcal{S}_{34} &= \mathcal{C}_3 \cap \mathcal{C}_4 = \{D\}\end{aligned}$$

and

$$\begin{aligned}\mathcal{N}_1 &= \{2\} \\ \mathcal{N}_2 &= \{1, 3\}\end{aligned}$$

Message Passing with Cliques

The previous message passing between nodes is now replaced by message passing between cliques.

The probability can now be expressed as

$$\begin{aligned} P(A,B,C,D,E,F) &= P(A,B)P(C,D|B)P(E|C,D)P(F|D) \\ &= P(\mathcal{C}_1)P(\mathcal{C}_2|\mathcal{S}_{12})P(\mathcal{C}_3|\mathcal{S}_{23})P(\mathcal{C}_4|\mathcal{S}_{34}) \end{aligned}$$

This can now be expressed as function of the Cliques and separators separately. For example

$$P(\mathcal{C}_3|\mathcal{S}_{23}) = \frac{P(\mathcal{C}_3,\mathcal{S}_{23})}{P(\mathcal{S}_{23})} = \frac{P(\mathcal{C}_3)}{P(\mathcal{S}_{23})}$$

Hence it is also possible to write

$$P(A,B,C,D,E,F) = \frac{\Pi_{i=1}^{4} P(\mathcal{C}_i)}{\Pi_{i=1}^{4}\Pi_{j=1}^{4} P(\mathcal{S}_{ij})}$$

Note here the probability of the empty separator will be 1.

Using the previous equalities we could for example write

$$P(\mathcal{C}_4) = P(\mathcal{C}_4|\mathcal{S}_{34}) \sum_{\mathcal{C}_3\setminus\mathcal{S}_{34}} P(\mathcal{C}_3|\mathcal{S}_{23}) \sum_{\mathcal{C}_2\setminus\mathcal{S}_{23}} P(\mathcal{C}_2|\mathcal{S}_{12}) \sum_{\mathcal{C}_1\setminus\mathcal{S}_{12}} P(\mathcal{C}_1)$$

Previously a message was from one node to another with information about the node. With cliques we get:

- messages from one clique, $\mathcal{C}_i$, to another, $\mathcal{C}_j$;
- information about the the separator $\mathcal{S}_{ij}$.

General Message Passing

General inference can be performed by using the following process:

1. Add undirected *edges* to all co-parents that are not currently connected (*marrying parents*).
2. Drop all directions in the graph to form the *moral graph*.
3. *Triangulate* the moral graph.
4. Identify the *cliques* in the triangulated graph.
5. Join the cliques together to form a *junction tree*.

Steps (1), (2) and (4) have already been illustrated in the simple example.

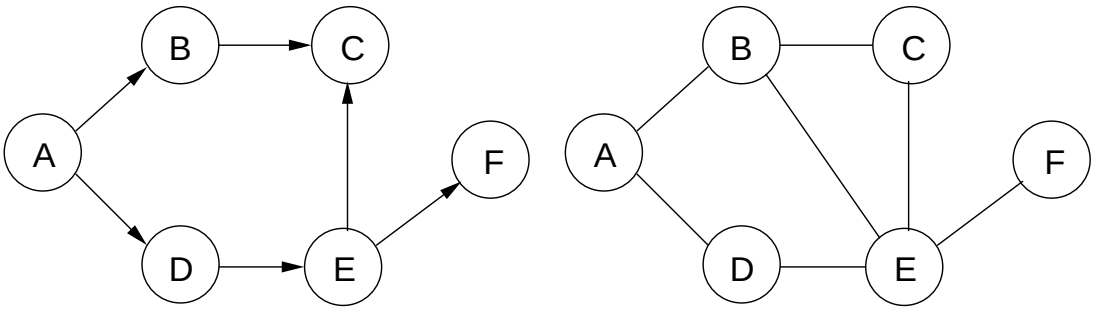
For the general process we need to add steps (3) and (5).

Triangulating a Graph

Triangulating a graph involves the following process

> add sufficient additional undirected links between nodes such that there are no *cycles* (i.e. closed paths) of length 4 or more distinct nodes without a shortcut.

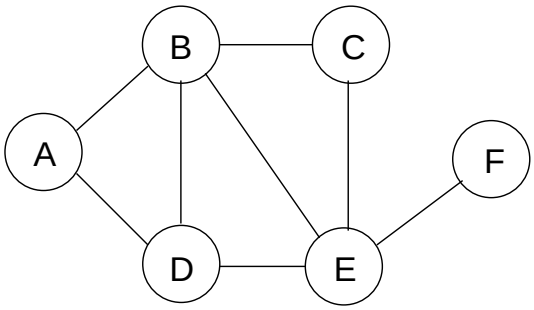
The previous example had no such cycles so there was no need to triangulate the graph. Consider the modified graph:



Bayesian Network Moral Graph

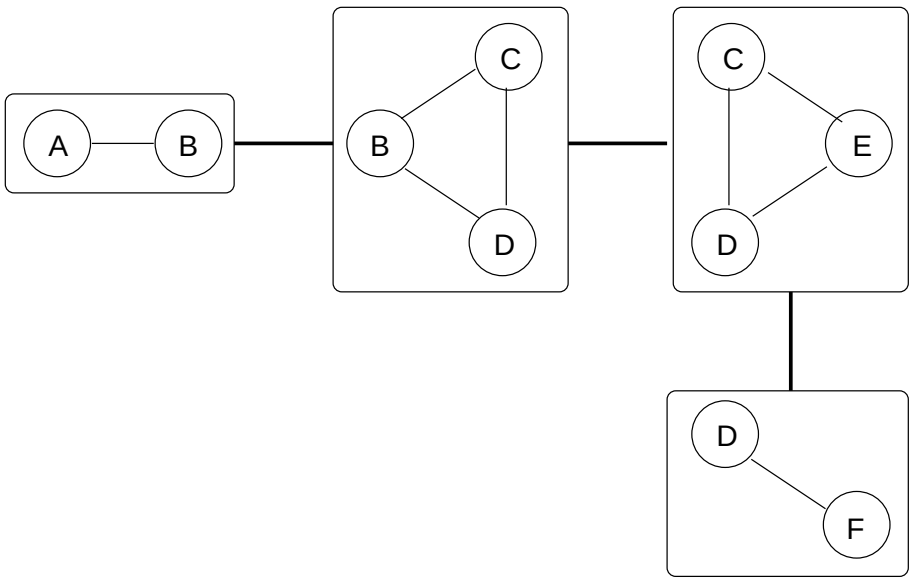
The moral graph has a cycle of 4, $\{A, B, E, D\}$. It is therefore necessary to add an additional edge - here B to D. (Note an edge A to E would also of satisfied the problem).

The associated triangulated graph is



Junction Tree

In the probability calculation an arbitrary ordering of the cliques was used. In general the order can affect the computational cost.



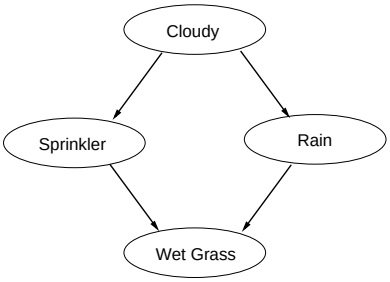
Junction trees have the following property

> Any node that appears in two different cliques must also appear in all the cliques along the path that connects the two cliques.

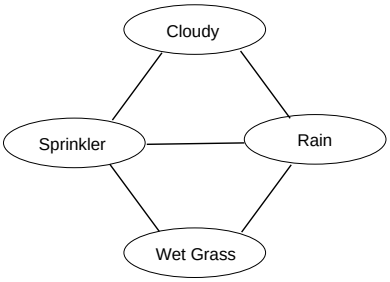
For example D appears in both $\mathcal{C}_2$ and $\mathcal{C}_4$. To be a valid junction tree it must also appear in $\mathcal{C}_3$, which it does.

Message Passing Example

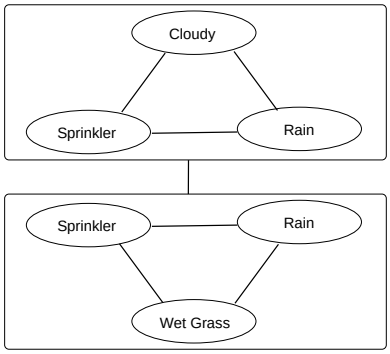
Using the sprinkler rain example to illustrate the process.



Bayesian Network



Moral Graph



Junction Tree

There are two cliques

$$\begin{aligned} \mathcal{C}_1 &= \{C, S, R\} \\ \mathcal{C}_2 &= \{S, R, W\} \\ \mathcal{S}_{12} &= \{S, R\} \end{aligned}$$

We want the message

$$\phi_{12}(\mathcal{S}_{12}) = \sum_C P(\mathcal{C}_1)$$

But it is know to be cloudy, so

$$\phi_{12}(\mathcal{S}_{12}) = P(S, R|C = \texttt{T}) = P(S|C = \texttt{T})P(R|C = \texttt{T})$$

The message propagated is

S	R	$P()$
T	T	0.08
T	F	0.02
F	T	0.72
F	F	0.18

Message Passing Example (cont)

The propagated message contains information about the separator between the two cliques, $\mathcal{S}_{12}$. From the basic layout there is already the information about the $P(\mathcal{C}_2|\mathcal{S}_{12})$. The probability of the clique $\mathcal{C}_2$ is given by

W	S	R	$P()$
T	T	T	0.0792
T	T	F	0.0180
T	F	T	0.6480
T	F	F	0.0000
F	T	T	0.0008
F	T	F	0.0020
F	F	T	0.0720
F	F	F	0.1800

From this complete table of the clique probabilities it is simple to determine the appropriate marginal probabilities.

Training

The estimation of a graphical model depends on the problem defined. They may be partitioned as

Structure	Observability	Method
Known	full	Sample statistics
Known	partial	EM or gradient ascent
Unknown	full	Search through model space
Unknown	partial	Structural EM

In this course only the known structure case will be considered.

For the fully observable case all the values of all the variables are observed at each time instance. The maximum likelihood estimate of the parameters are based on counts. For example

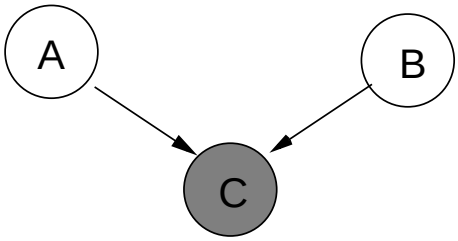
$$\begin{aligned} P(A|B,C) &= \frac{P(A,B,C)}{P(B,C)} \\ &\approx \frac{\text{Count}(A,B,C)}{\text{Count}(B,C)} \end{aligned}$$

The approximate sign is required as there are (normally!) only a finite number of training examples.

Partially Observable

The approach described here is based on EM (see lectures 3 & 4 of the course). However standard gradient descent schemes can also be used.

For the partially observable case the unobserved variables can be treated as latent variables in the EM algorithm. This is best illustrated by a simple example. Consider the following Bayesian network (structure 3)



Only C is observed. The overall probability of the observation may be written as

$$P(C) = \sum_A \sum_B P(A, B, C) = \sum_A \sum_B P(C|A, B)P(A)P(B)$$

There are 2 latent variables, A and B. Using EM the auxiliary function for this may be written as

$$\mathcal{Q}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k+1)}) = \sum_A \sum_B P(A, B|C, \boldsymbol{\theta}^{(k)}) \log \left(P(A, B, C|\boldsymbol{\theta}^{(k+1)}) \right)$$

where $\boldsymbol{\theta}^{(k)}$ are the model parameter estimates at the k^{th} iteration.

Simple Example

Consider the case where C is binary valued, $\{0, 1\}$, and an equal number of zeros and ones are observed. In addition A and B are also binary valued. The initial parameter estimates, $\boldsymbol{\theta}^{(0)}$, for the model are:

$$P(A_1) = 0.4; \quad P(B_1) = 0.4,$$

A	B	$P(C_1)$
0	0	0.1
0	1	0.9
1	0	0.8
1	1	0.1

Given the observations (value of C) and the current model parameters the posteriors of the latent variables are needed. Note these are not independent given C. If $C = 1$ (C_1)

$$P(A_1, B_1|C_1) = \frac{P(C_1|A_1, B_1)P(B_1)P(A_1)}{P(C_1)}$$

From the current model $P(C_1) = 0.46$. So

$$P(A_1, B_1|C_1) = \frac{0.1 \times 0.4 \times 0.4}{0.46} = 0.0348$$

It is possible to generate all the probabilities for $\boldsymbol{\theta}^{(0)}$

C	A	B	$P(A, B\|C)$
0	0	0	0.6000
1	0	0	0.0783
0	0	1	0.0444
1	0	1	0.4696

C	A	B	$P(A, B\|C)$
0	1	0	0.0889
1	1	0	0.4174
0	1	1	0.2667
1	1	1	0.0348

Simple Example (cont)

The training data is

$$\{0, 1, 1, 0, 0, 1, 0, 1\}$$

The estimate for the new model $P(A)$ is found from getting the expected counts

$$\begin{aligned} P(A_1) &= \frac{\text{Exp. Count}(A_1)}{\text{Exp. Count}(A = 1 \text{ or } A = 0)} \\ &= \frac{4 \times (0.4522 + 0.3556)}{8} = 0.4039 \end{aligned}$$

and to get $P(C_1|A_1, B_1)$

$$\begin{aligned} P(C_1|A_1, B_1) &= \frac{\text{Exp. Count}(C_1, A_1, B_1)}{\text{Exp. Count}(C = 1 \text{ or } C = 0, A_1, B_1)} \\ &= \frac{4 \times 0.0348}{4 \times (0.2667 + 0.0348)} = 0.1154 \end{aligned}$$

So the estimates, $\boldsymbol{\theta}^{(1)}$, are

$$P(A_1) = 0.4039, \quad P(B_1) = 0.4078$$

A	B	$P(C_1)$
0	0	0.1154
0	1	0.9136
1	0	0.8244
1	1	0.1154

This now describes the data as

$$P(C_1) = 0.479$$

EM can then be repeated. Any obvious solutions?

Continuous and Discrete RVs

It is possible to combine both continuous and discrete random variables. In this case it is helpful to distinguish between discrete and continuous variables in the network. Also it is helpful to show which variables (if any) are observed (measured) and which are unobserved (must be inferred).

The notation used here is:

- discrete random variables are represented as a square;
- continuous random variables are represented as a circle;
- observed variables will be indicated by shading the associated circle or square.
- unobserved variables will be indicated by not shading the associated circle or square.

In this module you have already come across two forms of Bayesian network, *Gaussian mixture models* and *factor analysis*.

GMMs and FA

Gaussian Mixture Model Factor Analysis

- Gaussian mixture models: these have the form

$$p(\boldsymbol{o}) = \sum_{m=1}^{M} c_m \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}^{(m)}, \boldsymbol{\Sigma}^{(m)})$$

this may be though as selecting a component from the PMF (formed of the component priors). Given the selected component w the observation is generated from the specified Gaussian component.

- Factor analysis: this is best described in terms of a generative model

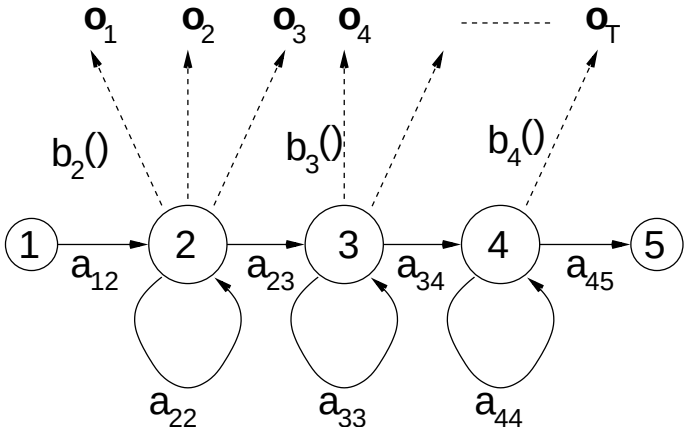
$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(\mathbf{0}; \mathbf{I}) \\ \boldsymbol{o} &= \mathbf{C}\boldsymbol{z} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(w)}) \end{aligned}$$

Here $\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ means distributed according to a multivariate Gaussian distribution of mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The overall covariance matrix is given by

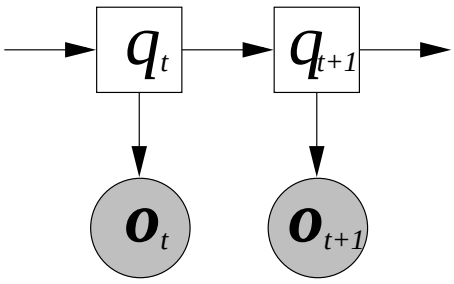
$$\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}' + \boldsymbol{\Sigma}^{(w)}$$

Hidden Markov Models

Some sources of data, such as speech, have a variable amount of data associated with each training (and test) sample. One method to handle this form of data is to use hidden Markov models.



HMM Structure



Dynamic BN

The structure of a HMM is shown above. The basic process may be described as

1. Perform a transition from the current state i to some state j determined by the transition matrix $\mathbf{A}$.
2. On entering a state an observation is generated. The probability of the observation depends only on the current state.

The HMM can be represented as a dynamic Bayesian network (DBN). The probability of the observation only depends on the current state. This indicates a form of conditional independence.

Precision Matrices

Another application of Graphical models is describing precision matrices (inverse covariance matrix modelling).

Consider a simple generative model

$$\begin{aligned} x_1 &= w_1, \qquad & w_1 &\sim \mathcal{N}(0,1) \\ x_2 &= w_2, \qquad & w_2 &\sim \mathcal{N}(x_1,1) \\ x_3 &= w_3, \qquad & w_3 &\sim \mathcal{N}(x_2,1) \end{aligned}$$

Here $\sim \mathcal{N}(0,1)$ means distributed according to a Gaussian distribution of mean 0 and variance 1.

What are the covariance matrix and precision matrix for $\boldsymbol{x}$?

The covariance matrix is simple to obtain using the standard formulae and noting that each of the noise sources are independent of one another

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{bmatrix}$$

Inverting this to get the precision matrix yields

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$$

It is simple to obtain the elements of the covariance matrix, what does the precision matrix tell us?

Some Matrix Equalities

Consider the partitioned covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

The following equalities apply (for reference):

- the covariance matrix of the conditional distribution

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

- the precision matrix of the partitioned matrix may be given by

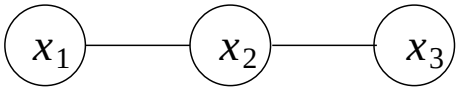
$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \mathbf{E}^{-1} & -\mathbf{E}^{-1}\mathbf{G} \\ -\mathbf{F}\mathbf{E}^{-1} & \mathbf{D}^{-1} + \mathbf{F}\mathbf{E}^{-1}\mathbf{G} \end{bmatrix}$$

  where

$$\begin{aligned} \mathbf{E} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ \mathbf{F} &= \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ \mathbf{G} &= \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \end{aligned}$$

Conditional Independence (again)

From the generative model for this process it is clear that x_1 and x_3 are conditionally independent given x_2. This yield a graphical model of the form



It is interesting to compute the covariance matrix for variables x_1 and x_3 given x_3. Using the previous equality and the covariance matrix (partitioned that matrix 1 has variables 1 and 3, 2 has 2)

$$\begin{aligned}\boldsymbol{\Sigma}_{1|2} &= \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} \frac{1}{2} \begin{bmatrix} 1 & 2 \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}\end{aligned}$$

The precision matrix can be represented as a graphical model. The lack of a connection (a zero in the precision matrix) indicates conditional independence.

Summary

The last two lectures have examined the use of Bayesian Networks. In particular:

- graphical models and conditional independence;
- Bayesian networks;
- inference in trees;
- cliques, separators and neighbours;
- general inference;
- training for fully and partially observed networks;
- examples of standard networks.