Bayesian Methods for Machine Learning

Lecture 3 (part 1) - Bayesian networks

Simon Leglaive

CentraleSupélec, 2020-2021

Bayesian networks

A Probabilistic graphical model (PGM) comprises nodes (also called vertices) connected by links (also known as edges or arcs).

Each node represents a random variable, and the links express probabilistic relationships between these variables.

- In Bayesian networks or directed graphical models (focus of today's lecture), the links of the graphs have a particular directionality indicated by arrows.
- In Markov random fields, or undirected graphical models, the links do not carry arrows and have no directional significance.

Example in medical diagnosis

In 1998 the LDS Hospital in Salt Lake City, Utah developed a Bayesian network to distinguish patients with pneumonia from patients with other diseases with high sensitivity (0.95) and specificity (0.965). It was used for many years in the clinic.

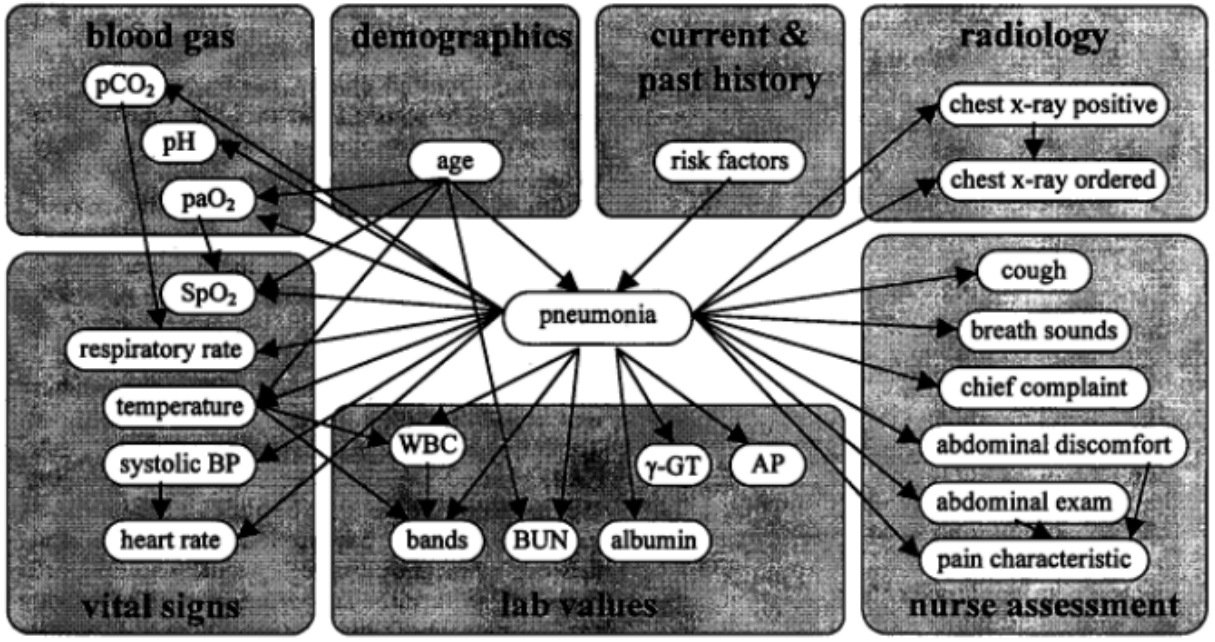


Figure 2: Structure of the Bayesian network. All variables are available in the HELP system during a patient's encounter in the emergency room with the exception of the chest x-ray information ("chest x-ray positive").

sensitivity: proportion of positives that are correctly identified; specificity: proportion of negatives that are correctly identified.

From the joint distribution to the graph

The graph captures the way in which the joint distribution over all the random variables can be decomposed into a product of factors, each depending only on a subset of the variables.

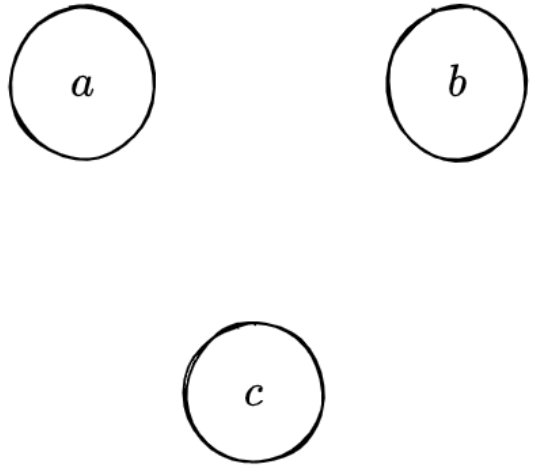
Consider first an arbitrary joint distribution $p(a, b, c)$ over three variables a, b, and c.

By application of the product rule (also called chain rule) of probability, we can write the joint distribution in the following form, without making any assumption:

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a)$$

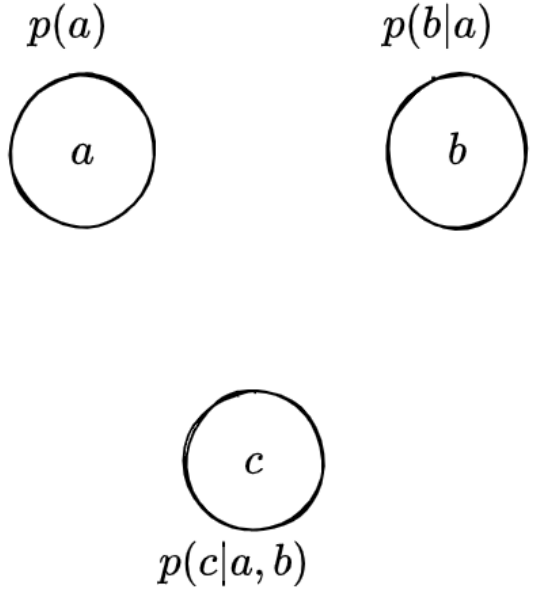
The factorization $p(a, b, c) = p(c|a, b)p(b|a)p(a)$ can be represented as a Bayesian network.

- introduce a node for each of the random variables;



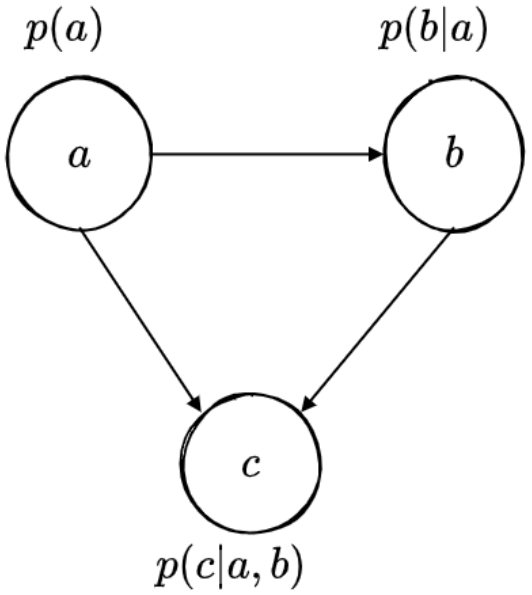
The factorization $p(a, b, c) = p(c|a, b)p(b|a)p(a)$ can be represented as a Bayesian network.

- introduce a node for each of the random variables;
- associate each node with the corresponding conditional distribution;



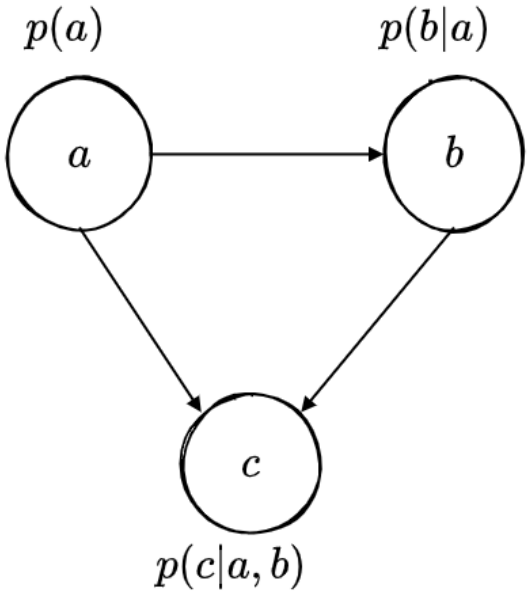
The factorization $p(a, b, c) = p(c|a, b)p(b|a)p(a)$ can be represented as a Bayesian network.

- introduce a node for each of the random variables;
- associate each node with the corresponding conditional distribution;
- for each conditional distribution, add directed links (arrows) from the nodes corresponding to the variables on which the distribution is conditioned.



The factorization $p(a, b, c) = p(c|a, b)p(b|a)p(a)$ can be represented as a Bayesian network.

- introduce a node for each of the random variables;
- associate each node with the corresponding conditional distribution;
- for each conditional distribution, add directed links (arrows) from the nodes corresponding to the variables on which the distribution is conditioned.



If there is a link going from a node a to a node b, then we say that node a is the parent of node b, and we say that node b is the child of node a.

This same principle applies for the joint distribution of an arbitrary number K of variables:

$$p(x_1, x_2, ..., x_K) = p(x_K|x_1, x_2, ..., x_{K-1}) ... p(x_2|x_1)p(x_1).$$

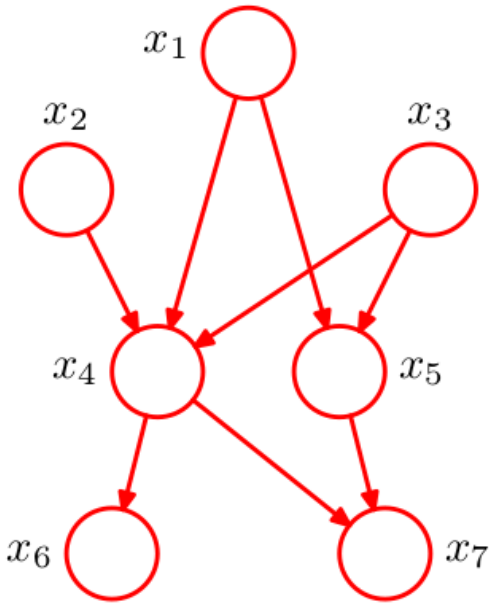
We can again represent this as a directed graph having K nodes, one for each conditional distribution on the right-hand side of the above equation. Each node has incoming links from all lower umbered nodes.

We say that this graph is fully connected because there is a link between every pair of nodes.

From the graph to the joint distribution

Consider the following Bayesian network, which is not fully connected as, for instance, there is no link from x_1 to x_2 or from x_3 to x_7.

The joint distribution of all the variables in this graph can be written as a product of conditional distributions, where each variable is conditioned only on the parents of the corresponding node.

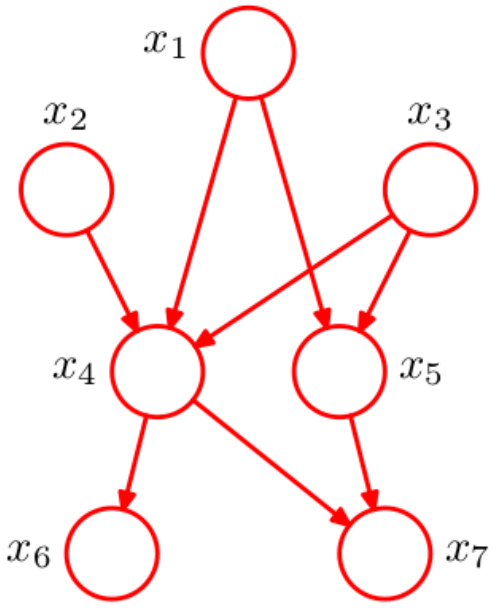


$$p(x_1, x_2, \ldots x_7) = p(x_7|x_{1:6})p(x_6|x_{1:5})p(x_5|x_{1:4})p(x_4|x_{1:3})p(x_3|x_{1:2})p(x_2|x_1)p(x_1)$$
$$= $$

From the graph to the joint distribution

Consider the following Bayesian network, which is not fully connected as, for instance, there is no link from x_1 to x_2 or from x_3 to x_7.

The joint distribution of all the variables in this graph can be written as a product of conditional distributions, where each variable is conditioned only on the parents of the corresponding node.



$$
\begin{aligned}
p(x_1, x_2, \ldots x_7) &= p(x_7|x_{1:6})p(x_6|x_{1:5})p(x_5|x_{1:4})p(x_4|x_{1:3})p(x_3|x_{1:2})p(x_2|x_1)p(x_1) \\
&= p(x_7|x_4, x_5)p(x_6|x_4)p(x_5|x_1, x_3)p(x_4|x_1, x_2, x_3)p(x_3)p(x_2)p(x_1)
\end{aligned}
$$

Factorizing the joint distribution in a Bayesian network

The joint distribution defined by a Bayesian network is given by the product, over all the nodes of the graph, of a conditional distribution for each node conditioned on the variables corresponding to the parents of that node.

Thus, for a graph with K nodes, the joint distribution is given by:

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k),$$

where pa_k denotes the set of parents of x_k and $\mathbf{x} = \{x_1, x_2, \ldots, x_K\}$.

This key equation expresses the factorization properties of the joint distribution for a Bayesian network.

Although we have considered each node to correspond to a single variable, we can equally well associate sets of variables and vector-valued variables with the nodes of a graph.

Important restriction

This is valid as long as there are no directed cycles, i.e. there are no closed paths within the graph such that we can move from node to node along links following the direction of the arrows and end up back at the starting node.

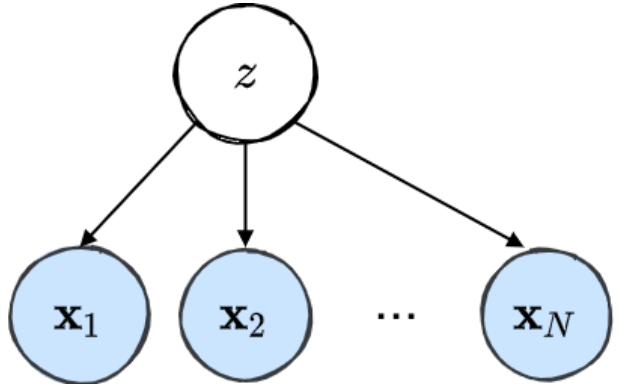
Bayesian networks are also called directed acyclic graphs, or DAGs.

Formal definition of a Bayesian network

A Bayesian network is a directed graph $G = (V, E)$ together with:

- A random variable x_k for each note $k \in V$
- A conditional probability distribution $p(x_k | \mathrm{pa}_k)$ for each node, defining the probability distribution of x_k conditioned on its parents.

Example (from lecture 2)

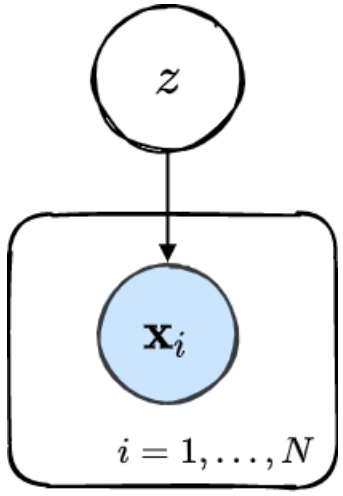


Latent variables are represented with empty circles, observations with filled circles.

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, z) = p(z) \prod_{i=1}^{N} p(\mathbf{x}_i | z)$$

Example (from lecture 2)

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, z) = p(z) \prod_{i=1}^{N} p(\mathbf{x}_i|z)$$



Latent variables are represented with empty circles, observations with filled circles.

The rectangle corresponds to the plate notation: the sub-graph contained in a rectangle is repeated according to the indicated indices. Any link that crosses a plate boundary is also replicated.

Generative model and ancestral sampling

A Bayesian network captures the causal process by which the data were generated. For this reason, such models are often called generative models.

Generative model and ancestral sampling

A Bayesian network captures the causal process by which the data were generated. For this reason, such models are often called generative models.

We can generate data from the definition of the Bayesian network, by sampling successively from the individual conditional distributions.

This method is called ancestral sampling, each variable being sampled given its parents (ancestors).

For example, assume we want to sample from $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$ where we know each one of the conditional distributions.

Ancestral sampling:

- we first sample $\tilde{x}_1$ from $p(x_1)$
- then we sample $\tilde{x}_2$ from $p(x_2|\tilde{x}_1)$
- finally we sample $\tilde{x}_3$ from $p(x_3|\tilde{x}_1, \tilde{x}_2)$

We obtain a sample $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ from the joint distribution $p(x_1, x_2, x_3)$.

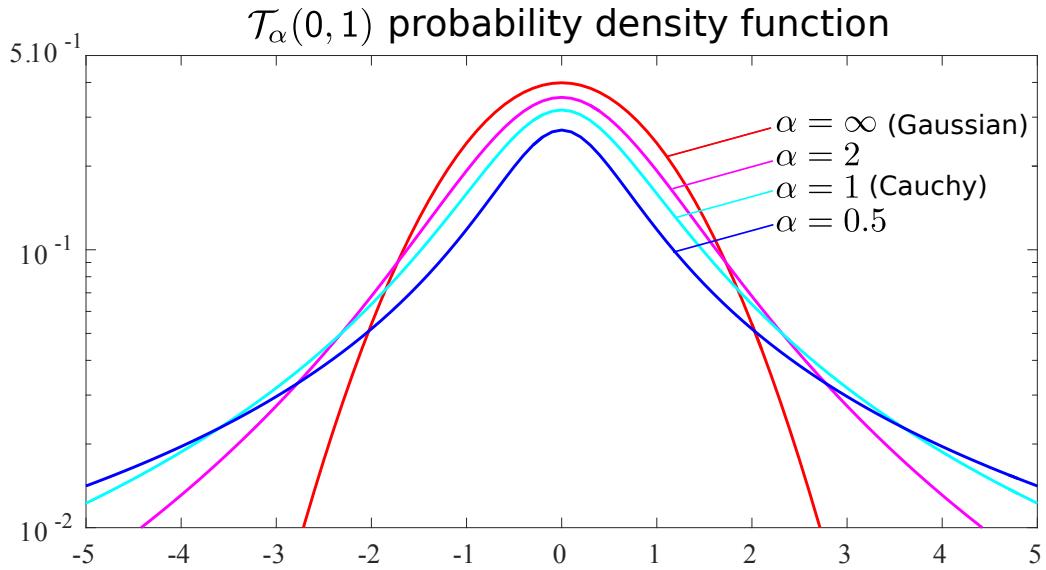
To obtain a sample from some marginal distribution corresponding to a subset of the variables, e.g. $p(x_2)$, we simply take the sampled values for the required nodes and ignore the sampled values for the remaining nodes, e.g. $\tilde{x}_2$.

Generative model with latent variables

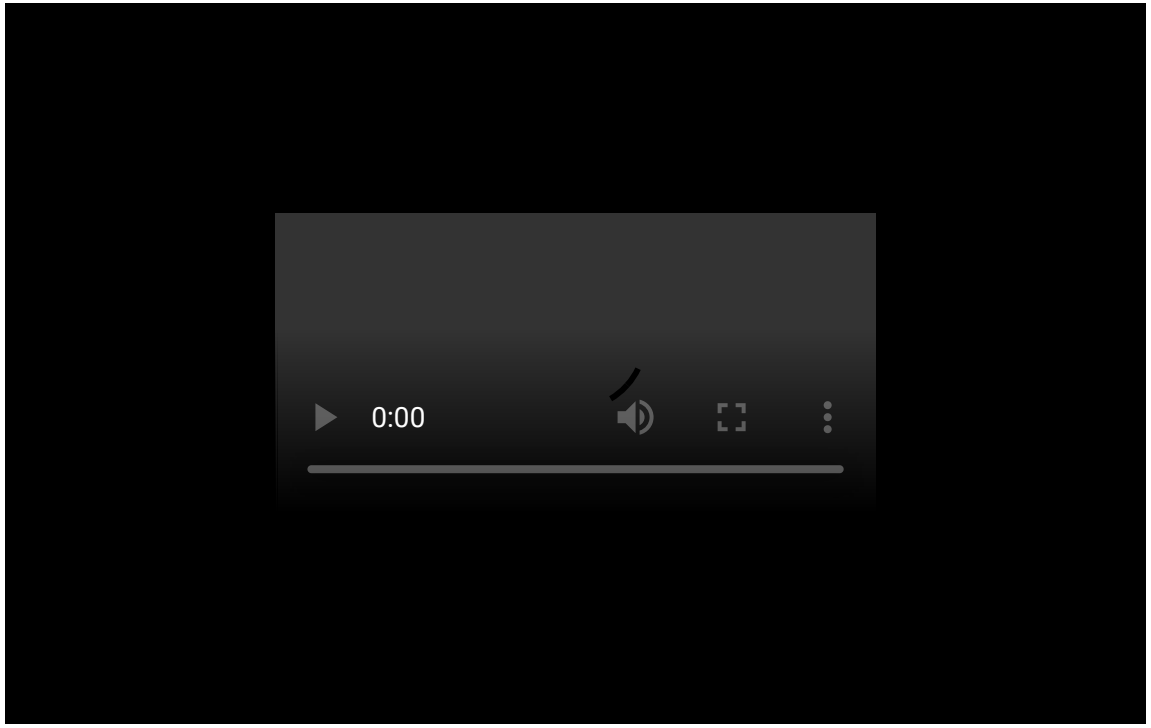
Hidden variables may be introduced to build a complex distribution from simpler ones.

The Student's t example (see Lecture 2):

$$\begin{cases} p(x|v) &= \mathcal{N}\left(x;\mu,v\right) \\ p(v) &= \mathcal{IG}\left(v;\frac{\alpha}{2},\frac{\alpha}{2}\lambda^2\right) \end{cases} \quad \Leftrightarrow \quad p(x) = \int_0^{+\infty} p(x|v)p(v)dv = \mathcal{T}_\alpha(x;\mu,\lambda)$$



Hidden variables may also have an explicit interpretation.



In either case, the technique of ancestral sampling applied to a generative model mimics the creation of the observed data.

D-Separation

Conditional independence

- Consider three variables a, b, and c, and suppose that the conditional distribution of a, given b and c, is such that it does not depend on b:

$$p(a|b,c) = p(a|c).$$

We say that a is conditionally independent of b given c.

Conditional independence

- Consider three variables a, b, and c, and suppose that the conditional distribution of a, given b and c, is such that it does not depend on b:

$$p(a|b,c) = p(a|c).$$

  We say that a is conditionally independent of b given c.

- This can be expressed in a slightly different way if we consider the joint distribution of a and b conditioned on c:

$$p(a,b|c) = p(a|b,c)p(b|c) = p(a|c)p(b|c).$$

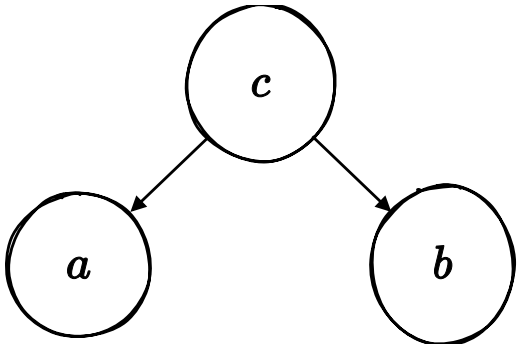
- This says that the variables a and b are statistically independent, given c.

- Conditional independence properties simplify both the structure of a model and the computations needed to perform inference and learning in Bayesian networks.
- An important and elegant feature of graphical models is that conditional independence properties of the joint distribution can be read directly from the graph without having to perform any analytical manipulations.
- The general framework for achieving this is called D-separation, where "D" stands for "directed".

To motivate and illustrate the concept of D-separation, let's start by looking at three simple Bayesian network structures with three nodes a, b and c.

"Tail-to-tail" or "common parent" structure

- None of the variables are observed.
- The node c is said to be tail-to-tail because it is connected to the tails of the two arrows.
- The joint distribution writes:



$$p(a, b, c) = p(c)p(a|c)p(b|c)$$

- Are a and b independent?

$$p(a, b) = \int p(a, b, c)dc = \int p(c)p(a|c)p(b|c)dc \neq p(a)p(b)$$

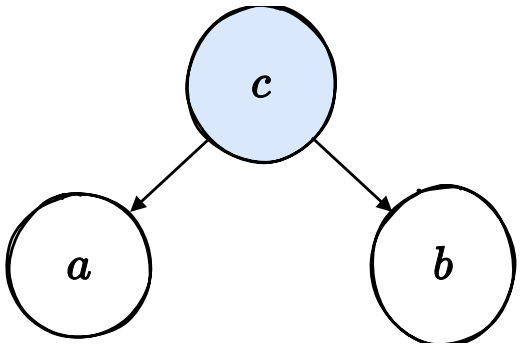
Intuitively, c connects a and b, making them dependent.

We assume continuous random variables. In the discrete case, the integral is simply replaced by a sum.

"Tail-to-tail" or "common parent" structure

- The variable c is now observed.
- The joint distribution writes:

$$p(a, b, c) = p(c)p(a|c)p(b|c)$$

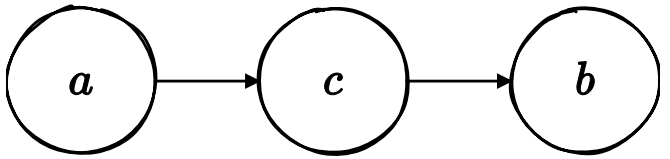


- Are a and b conditionally independent?

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(c)p(a|c)p(b|c)}{p(c)} = p(a|c)p(b|c)$$

c contains all the information that determines the outcomes of a and b. Once it is observed, there is nothing else that affects these variables' outcomes. In other words, $p(a, b|c) = p(a|b, c)p(b|c) = p(a|c)p(b|c)$.

"Head-to-tail" or "cascade" structure



- None of the variables are observed.
- The node c is said to be head-to-tail because it is connected to the head and tail of the left and right arrows, respectively.
- The joint distribution writes:

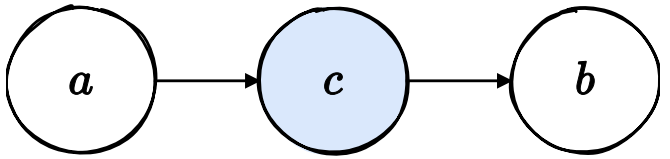
$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

- Are a and b independent?

$$p(a, b) = \int p(a, b, c)dc = \int p(a)p(c|a)p(b|c)dc \neq p(a)p(b)$$

c connects a and b, making them dependent.

"Head-to-tail" or "cascade" structure



- The variable c is now observed.
- The joint distribution writes:

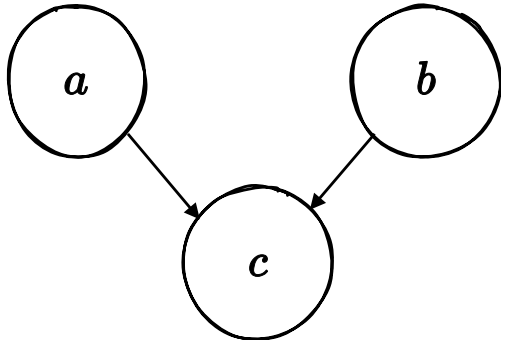
$$p(a, b, c) = p(a)p(c|a)p(b|c)$$

- Are a and b conditionally independent?

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(c|a)}{p(c)}p(b|c) = p(a|c)p(b|c)$$

c contains all the information that determines the outcomes of b, so once it is observed a has no influence on b anymore. In other words, $p(a, b|c) = p(b|a, c)p(a|c) = p(b|c)p(a|c)$.

"Head-to-head" or "V" structure



- None of the variables are observed.
- The node c is said to be head-to-head because it is connected to the heads of the two arrows.
- The joint distribution writes:

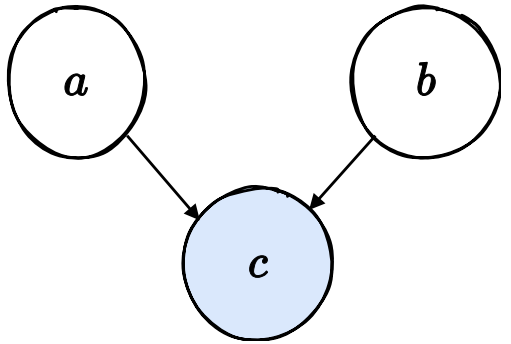
$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

- Are a and b independent?

$$p(a, b) = \int p(a, b, c)dc = \int p(a)p(b)p(c|a, b)dc = p(a)p(b)$$

a and b are two indepent factors that determine the outcome of c.

"Head-to-head" or "V" structure



- The variable c is now observed.
- The node c is said to be head-to-head because it is connected to the heads of the two arrows.
- The joint distribution writes:

$$p(a, b, c) = p(a)p(b)p(c|a, b)$$

- Are a and b conditionally independent?

$$p(a, b|c) = \frac{p(a, b, c)}{p(c)} = \frac{p(a)p(b)p(c|a, b)}{p(c)} \neq p(a|c)p(b|c)$$

Observing c makes a and b dependent as it is a common child of the two nodes. Suppose that $c = a + b$, if we know the value of c, a and b cannot vary independently.

This third example has the opposite behaviour from the first two.

Summary for a 3-variable graph

- A tail-to-tail (common parent) node makes the two other nodes conditionally independent when it is observed.
- A head-to-tail (cascade) node makes the two other nodes conditionally independent when it is observed.
- A head-to-head (V-structure) node makes the two other nodes conditionally dependent when it is observed.

Summary for a 3-variable graph

- A tail-to-tail (common parent) node makes the two other nodes conditionally independent when it is observed.
- A head-to-tail (cascade) node makes the two other nodes conditionally independent when it is observed.
- A head-to-head (V-structure) node makes the two other nodes conditionally dependent when it is observed, *and/or when at least one of its descendant is observed.*

Suppose that $c = a + b$ and $d = c + 2$, if we know the value of c and/or d, a and b cannot vary independently.

We can apply these 3 principles recursively to analyze larger Bayesian networks with arbitrary structure.

This is the notion of D-separation.

Definition of D-Separation

Consider a Bayesian network in which A, B, and C are arbitrary nonintersecting sets of nodes.

We say that A and B are D-separated given C if all possible paths that connect any node in A to any node in B are blocked given C.

Equivalently, A and B are D-separated given C if they are not connected by any path that is not blocked (i.e. that is active).

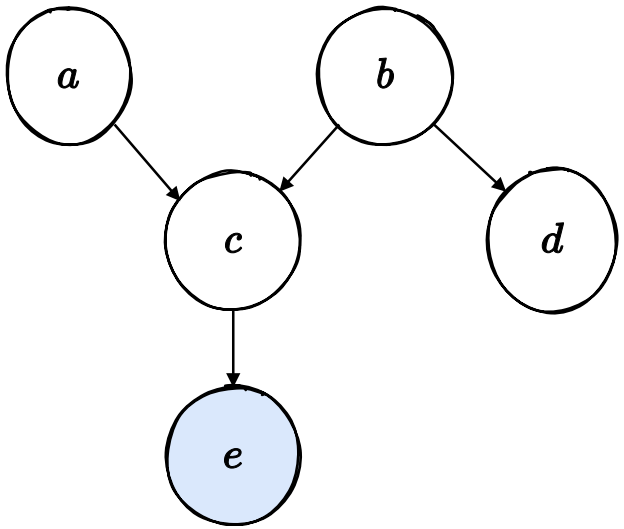
If A and B are D-separated given C, then $p(A, B|C) = p(A|C)p(B|C)$.

A path is said to be blocked given observed variables O if it includes a node Y such that either:

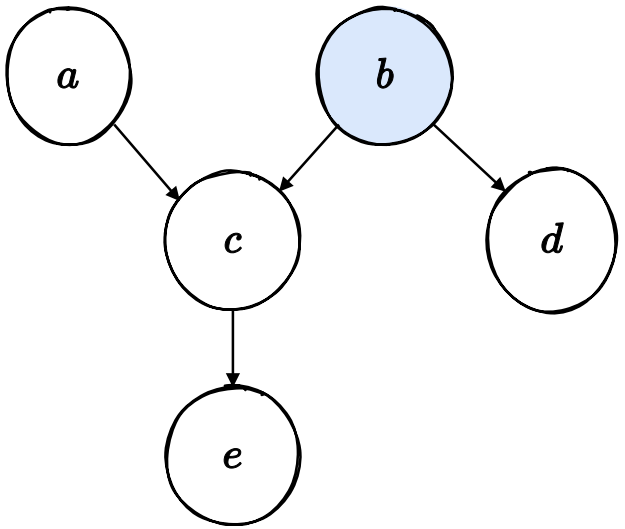
- Y is a head-to-tail node ($X \rightarrow Y \rightarrow Z$ or $X \leftarrow Y \leftarrow Z$) and Y is in O (i.e. observed), or
- Y is a tail-to-tail node ($X \leftarrow Y \rightarrow Z$) and Y is in O (i.e. observed), or
- Y is head-to-head node ($X \rightarrow Y \leftarrow Z$) and Y or any of its descendant is not in O (i.e. not observed).

Recipe for D-separation:

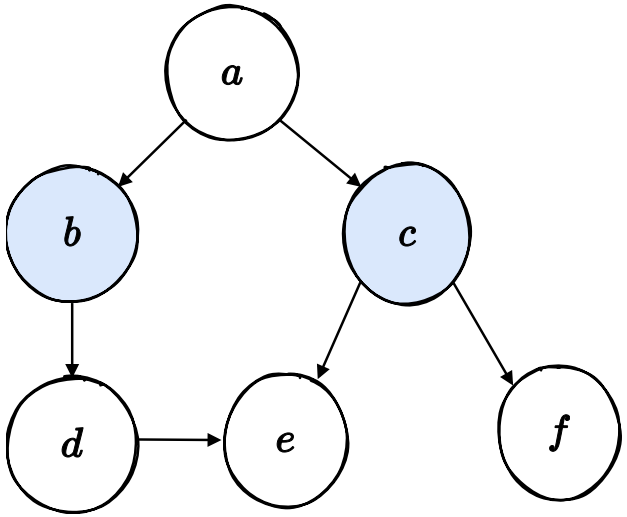
- List all the paths between any node in A and any node in B.
- If all paths are blocked, A and B are D-separated given C.
- Equivalently, if you can find one active path (i.e. not blocked), A and B are not D-separated given C.



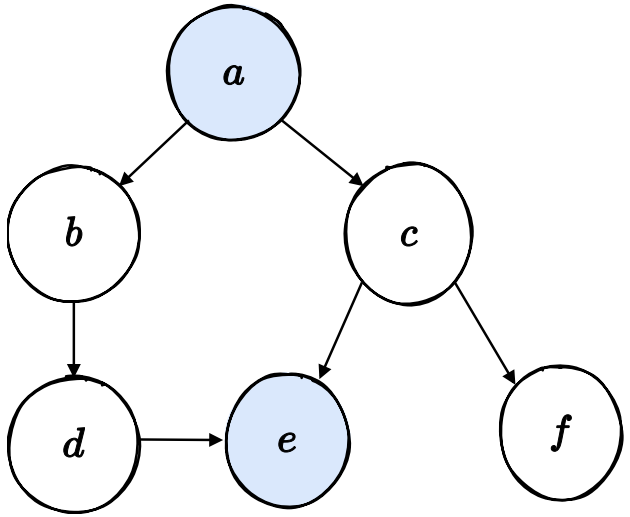
Are a and d D-separated given e?



Are a and d D-separated given b?



Are a and e D-separated given b and c?



Are b and c D-separated given a and e?

Markov blanket

Consider a joint distribution of an arbitrary number K of variables $p(x_1, x_2, \ldots, x_K)$ represented by a Bayesian network with K nodes. Consider the conditional distribution of a particular variable x_k given all the remaining ones $\{x_i\}_{i \neq k}$:

$$p\left(x_k | \{x_i\}_{i \neq k}\right) = \frac{p(x_1, x_2, \ldots, x_K)}{p(\{x_i\}_{i \neq k})} = \frac{p(x_1, x_2, \ldots, x_K)}{\int p(x_1, x_2, \ldots, x_K) dx_k} = \frac{\prod_{j=1}^{K} p(x_j | \mathrm{pa}_j)}{\int \prod_{j=1}^{K} p(x_j | \mathrm{pa}_j) dx_k}.$$

- Any factor $p(x_j | \mathrm{pa}_j)$ that does not have any functional dependence on x_k can be taken outside the integral and will therefore cancel between numerator and denominator.
- The only factors that remain will be the conditional distribution $p(x_k | \mathrm{pa}_k)$ for node x_k itself, together with the conditional distributions $p(x_j | \mathrm{pa}_j)$ where x_k is in pa_j.
- $p(x_k | \mathrm{pa}_k)$ will depend on the parents of x_k, whereas the remaining conditionals $p(x_j | \mathrm{pa}_j)$ will depend on the children of x_k, as well as its co-parents (the other parents of x_j).

The set of nodes comprising the parents, the children and the co-parents is called the Markov blanket.

In a Bayesian network, the conditional distribution of an arbitrary variable x_k given all the remaining variables in the graph only depends on the variables in its Markov blanket:

$$p\left(x_k|\{x_i\}_{i\neq k}\right) = p\left(x_k|\mathrm{MB}(x_k)\right).$$

Given its Markov blanket, x_k is conditionally independent of all the remaining variables in the graph. The Markov blanket contains all the information one needs to infer x_k.