Introduction to Machine Learning

Probabilistic graphical models

Pascal Gepperth

January 25, 2021

Contents

1 What are graphical Models? 3

2 Bayesian Networks 3
2.1 Plates . . . 3
2.2 Deterministic parameters . . . 4
2.3 Observed variables . . . 5
2.4 Conditional Independence . . . 5
2.5 D-separation . . . 8

3 Markov random fields 9
3.1 Conditional independence . . . 10
3.2 Factorization properties . . . 10
3.3 Image de-noising . . . 11
3.4 Relation to directed graphs . . . 13

4 Advantages of probabilistic graphical models 15

1 What are graphical Models?

A (probabilistic) graphical model is a probabilistic model in which the conditional dependence structure between the random variables are given and visualized by a graph.

A graph contains *nodes* (or *vertices*) which are connected by *links* (or *edges*). These links can be directed or undirected.

Belonging to a probabilistic graphical model, the graph's nodes represent random variables. The links express a probabilistic relationship between these variables.

2 Bayesian Networks

Bayesian Networks are graphical models in which the links of the graph are directed. The corresponding graph has to be acyclic. Suppose there are two random variables a and b. If there is a link from a to b, the node a is called *parent* of node b.

The joint distribution given by a graph can be written as product of the conditional distribution for each node of the graph, conditioned on the random variables of all parents of the node.

For a graph with K nodes, the joint distribution of $\mathbf{x} = \{x_1, ..., x_K\}$ 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.

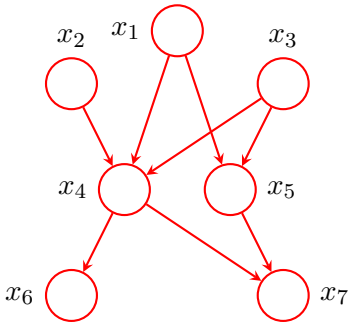

Figure 1 – Example of a directed acyclic graph.

The joint distribution corresponding to the graph in Figure 1 can be written as

$$p(x_1, ..., x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5).$$

2.1 Plates

In more complex models, it is very inconvenient to write out multiple nodes of the form $t_1, ..., t_N$ as shown in Figure 2.

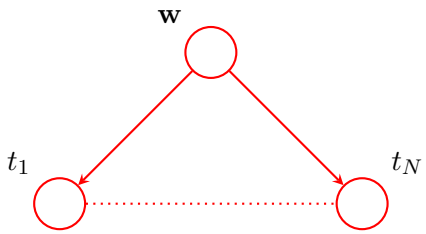

Figure 2 – Directed graph corresponding to the Bayesian polynomial regression model.

To write such multiple nodes more compactly, we can draw a single representative node t_n that is surrounded by a box labeled with N. This is called *plate*.

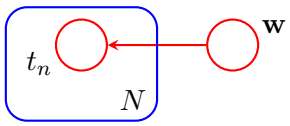

Figure 3 – The same graph as Figure 2 drawn more compactly using a plate.

So the plate tells us that there are N nodes like t_n with the same parents.

2.2 Deterministic parameters

In case of polynomial regression, our random variables are the vector of polynomial coefficients $\mathbf{w}$ and the observed target vector $\mathbf{t} = (t_1, ..., t_N)^T$. The joint distribution is given by

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}).$$

The target vector corresponds with the input vector $\mathbf{x} = (x_1, ..., x_N)^T$. We also include the noise variance σ^2 and the hyperparameter α, representing the precision of the Gaussian prior over $\mathbf{w}$. These parameters are called *deterministic parameters*.

For some problems it can be useful to make the parameters of a model explicit. Including this data as well, the joint distribution can be adjusted to

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2).$$

Deterministic parameters can also be visualized. While random variables are denoted by open circles, deterministic parameters will be denoted by a smaller solid circle, as shown in Figure 4.

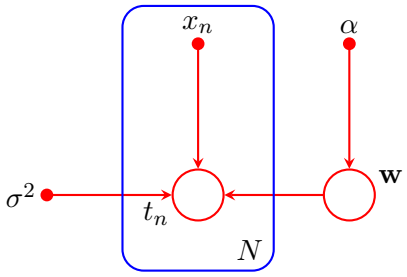

Figure 4 – The same graph as Figure 3 drawn with deterministic parameters.

2.3 Observed variables

Using a graphical model to solve a problem, some of the random variables are set to be observed. For example the variables $\{t_n\}$ are from the training set of a polynomial regression. These values are "known" and will not be changed by the model.

The corresponding nodes of observed variables will be shaded, shown in Figure 5. Not observed variables, as $\mathbf{w}$, are called *latent* or *hidden* variables.

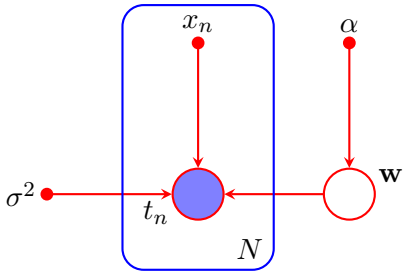

Figure 5 – The same graph as Figure 4 with $\{t_n\}$ observed.

2.4 Conditional Independence

Suppose we have three random variables a, b and c. If

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

a is called *conditionally independent* of b given c. The joint distribution of a and b conditioned on c factorizes into

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c). \end{aligned}$$

To shorten things up, we write

$$a \perp\!\!\!\perp b|c.$$

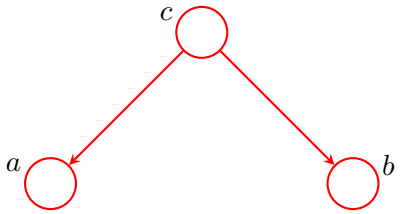

Figure 6 – A graph with a tail-to-tail node c.

Given the joint distribution

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

and the corresponding graph shown in Figure 6, the marginal distribution is given by

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

In general, this does not factorize into the product $p(a)p(b)$, so it is

$$a \not\perp\!\!\!\perp b|\emptyset.$$

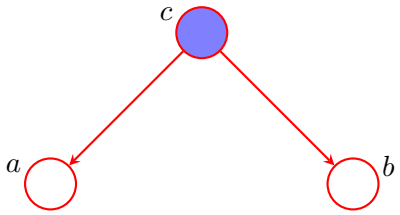

Figure 7 – The same graph as in Figure 6 with node c observed.

Conditioned on c (shown in Figure 7), the conditional distribution of a and b given c is

$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a|c)p(b|c)p(c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

and shows that

$$a \perp\!\!\!\perp b|c.$$

The node c is called *tail-to-tail*. If c is observed, it blocks the path from a to b, so a and b can become conditional independent.

Figure 8 – A graph with a head-to-tail node c.

Considering the graph shown in Figure 8 and the joint distribution corresponding to this graph

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

we see that

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

does not factorize into $p(a)p(b)$, so

$$a \not\perp\!\!\!\perp b|\emptyset.$$

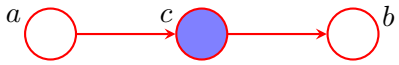

Figure 9 – The same graph as in Figure 8 with node c observed.

Observing the node c, the conditional distribution is given by

$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c). \end{aligned}$$

In this case, we obtain

$$a \perp\!\!\!\perp b|c.$$

The node c is called *head-to-tail*. If c is observed, it blocks the path from a to b, so they become conditional independent as before.

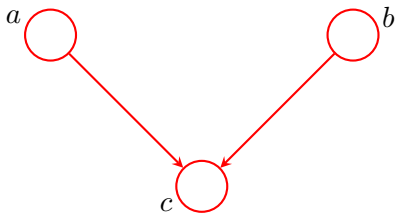

Figure 10 – A graph with a head-to-head node c.

The graph in Figure 10 corresponds to the joint distribution

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

None of the variables observed, the marginal distribution is given by

$$\begin{aligned} p(a, b) &= \sum_c p(a)p(b)p(c|a, b) \\ &= p(a)p(b) \sum_c p(c|a, b) \\ &= p(a)p(b), \end{aligned}$$

which shows

$$a \perp\!\!\!\perp b|\emptyset.$$

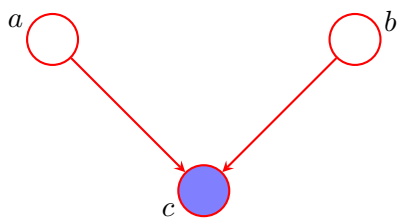

Figure 11 – The same graph as in Figure 10 with node c observed.

In case of the node c is observed, the conditional distribution

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

which in general does not factorize into $p(a)p(b)$, leads to

$$a \not\perp\!\!\!\perp b|c.$$

Similar things happens if one (or more) of c's descendants are observed.

The node c is called *head-to-head*. Sometimes, the whole structure is also called *v-structure*. If c is unobserved, it blocks the path from a to b, so they become conditional independent. When node c (and/or at least one of its descendants) is observed, the path becomes unblocked.

2.5 D-separation

Let A, B and C be a set of nodes. We consider all possible paths from any node in A to any node in B to be blocked. A path is said to be blocked if there is any node within the path, that either

- the node is head-to-tail or tail-to-tail, and the node is in the set C, or
- the node is head-to-head, and neither the node nor any of its descendants is in the set C.

A is said to be *d-separated* from B by C if all those paths are blocked, and it applies

$$A \perp\!\!\!\perp B|C$$

To show how this works, we take a look at Figure 12. There is just one path from a to b. In graph 12a the path is not blocked by e, because e is a head-to-head node and its descendant c is observed. The path is also not blocked by f, because f is a tail-to-tail node and it is not observed. So it is $a \not\perp\!\!\!\perp b|c$.

In graph 12b, the path is blocked by node e, because e and f are not observed. Further the path is also blocked by f, because f is observed. So $a \perp\!\!\!\perp b|f$ does follow from this.

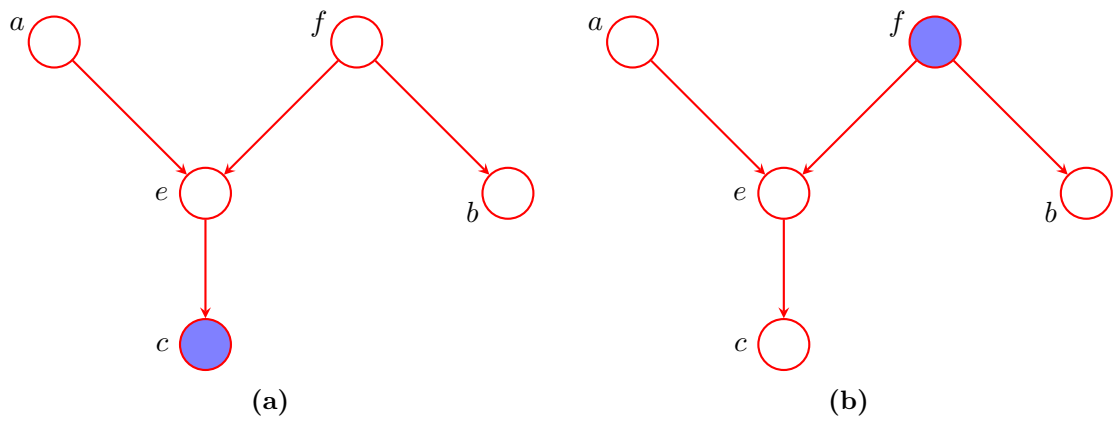

Figure 12 – Two types of graphical models.

Another important point is the concept of a *Markov blanket* or *Markov boundary*. Suppose we have a joint distribution $p(x_1, ..., x_D)$ over D nodes. Considering the conditional distribution of the node x_i conditioned on all other variables $x_{j\neq i}$, we get

$$p(x_i|x_{j\neq i}) = \frac{p(x_1, ..., x_D)}{\int p(x_1, ..., x_D)\ \mathrm{d}x_i}$$

$$= \frac{\prod_{k=1}^{D} p(x_k|\mathrm{pa}_k)}{\int \prod_{k=1}^{D} p(x_1, ..., x_D)\ \mathrm{d}x_i}$$

$$= \frac{\prod_{k\notin \mathrm{MB}} p(x_k|\mathrm{pa}_k) \prod_{k\in \mathrm{MB}} p(x_k|\mathrm{pa}_k)}{\prod_{k\notin \mathrm{MB}} p(x_k|\mathrm{pa}_k) \int \prod_{k\in \mathrm{MB}} p(x_k|\mathrm{pa}_k)\ \mathrm{d}x_i}$$

$$= \frac{\prod_{k\in \mathrm{MB}} p(x_k|\mathrm{pa}_k)}{\int \prod_{k\in \mathrm{MB}} p(x_k|\mathrm{pa}_k)\ \mathrm{d}x_i}.$$

The Markov Blanket of a node x_i is the set of all parents, children and co-parents. Conditioned on these, x_i is independent of the rest of the graph.

3 Markov random fields

A *Markov random field*, *Markov network* or *undirected graphical model* is a probabilistic graphical model in which the links of the graph are undirected.

3.1 Conditional independence

As in Bayesian networks, paths between nodes can also be blocked in Markov random fields.

Let A, B and C be sets of nodes. Given C, a path from any node in A to any node in B is blocked if the path passes at least one node in C. If all paths from A to B are blocked, it is

$$A \perp\!\!\!\perp B|C.$$

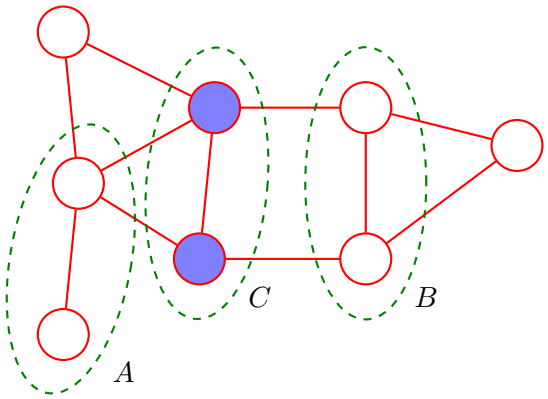

Figure 13 – Example of a Markov random field. A is conditionally independent of B given C because all paths from A to B passes at least one node in C.

The Markov blanket of a node x_i within an undirected graph is the set of all neighbors of x_i. Conditioned on these, x_i is conditionally independent of all remaining nodes of the graph.

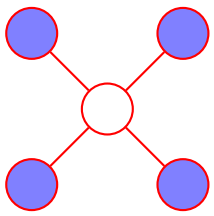

Figure 14 – Markov blanket of a node within an undirected graph.

3.2 Factorization properties

Let x_i and x_j be two unconnected nodes. Conditioned on all other nodes, x_i and x_j are conditionally independent.

$$p(x_i, x_j|\mathbf{x}_{\setminus\{i,j\}}) = p(x_i|\mathbf{x}_{\setminus\{i,j\}})p(x_j|\mathbf{x}_{\setminus\{i,j\}})$$

To factorize the joint distribution of an undirected graph, x_i and x_j cannot appear in the same factor.

To do so, we can use *cliques*. A clique is a subset of nodes, in which all nodes are connected to each other. A fully connected set of nodes is called *maximal clique* if no other node can be included, so that the new subset is still a clique.

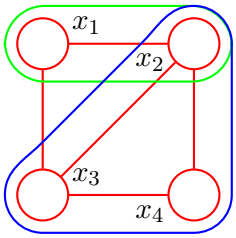

Figure 15 – Example of factorization (undirected) using cliques.

Figure 15 shows a graph with five cliques of two nodes (e.g. $\{x_1, x_2\}$) and two (maximal) cliques of three nodes (e.g. $\{x_2, x_3, x_4\}$). The set $\{x_1, x_2, x_3, x_4\}$ is not a clique.

Using *potential functions* $\psi_C(\mathbf{x}_C)$, the joint distribution factorizes into the product

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_C(\mathbf{x}_C),$$

where C denotes the clique and $\mathbf{x}_C$ denotes the set of variables in that clique. Z is a normalization constant and given by

$$Z = \sum_{\mathbf{x}} \prod_C \psi_C(\mathbf{x}_C).$$

The potential functions must be non-negative to ensure that $p(\mathbf{x}) \geq 0$.

To make connection between factorization and conditional independence, we use potential functions $\psi_C(\mathbf{x}_C)$ that are strictly positive. Those can be expressed as exponential functions

$$\psi_C(\mathbf{x}_C) = \exp\{-E(\mathbf{x}_C)\},$$

where $E(\mathbf{x}_C)$ is called *energy function*. The exponential representation is called *Boltzmann distribution*. Referring to the joint distribution, the total energy can be obtained by summing up the energies of the maximal cliques. Potential functions do not have any specific probabilistic interpretation, so they can be chosen more flexible.

3.3 Image de-noising

To illustrate the use of Markov random fields, we can take a look at image de-noising. Figure 16a shows the original (noise-free) binary image. By randomly flipping 10% of the pixels, a noisy image can be created out of the original image, shown in Figure 16b.

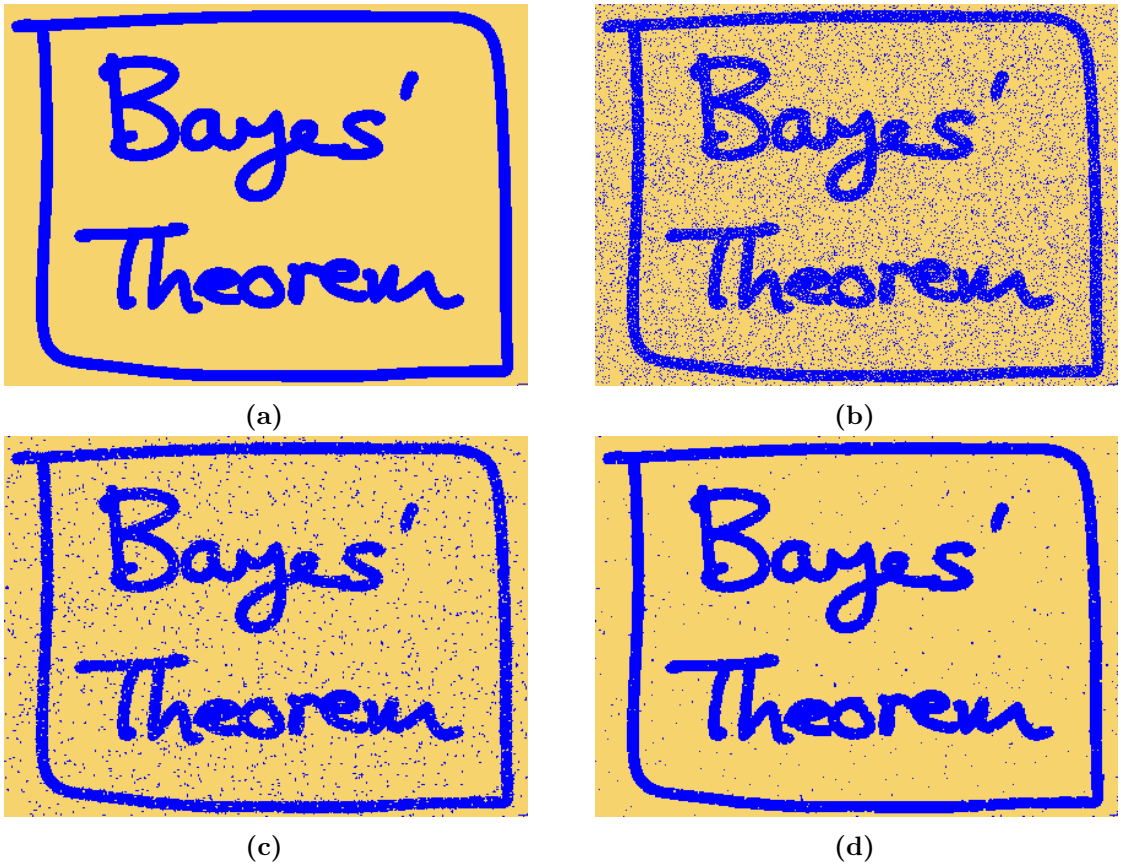

Figure 16 – Example of image de-noising. The original binary image (a) is use to generate a noisy image (b), by flipping randomly 10% of the pixels. Using iterated conditional models (ICM) (c), the image can be restored, where 96% of the pixels accords to the original image. Using the graph-cut algorithm (d), 99% of the pixels will accord to the original image.

Given the noisy image, we want to restore the original noise-free image.

Let us treat the images as D-dimensional vectors of binary pixels. The value of a pixel is -1 or $+1$. So $\mathbf{y} \in \{-1, +1\}^D$ represents the observed noisy image and $\mathbf{x} \in \{-1, +1\}^D$ represents the original (unknown) noise-free image.

In general, neighboring pixels x_i and x_j are strongly correlated. Also x_i and y_i are strongly correlated because the noise only flipped 10% of the pixels. This leads to two types of cliques: $\{x_i, y_i\}$ and $\{x_i, x_j\}$, where x_i and x_j are neighboring pixels, shown in Figure 17.

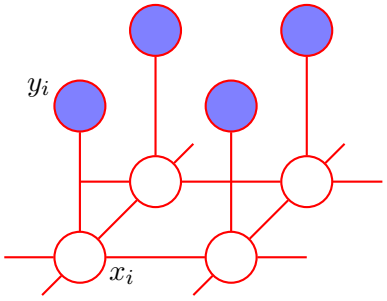

Figure 17 – Illustration of the Markov random field for image de-noising.

Now we can choose $-\eta x_i y_i$, where η is a positive constant, as energy function for the clique $\{x_i, y_i\}$ and $-\beta x_i x_j$, where β is a positive constant, for $\{x_i, x_j\}$. These energy functions are very simple and give a negative value if both pixels have the same sign. The property of the potential function allows us to add an extra term hx_i for each noise-free images' pixel i, which has the effect of a bias and prefers one particular sign.

The total energy is given by

$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

and the joint distribution by

$$p(\mathbf{xy}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}.$$

Observing $\mathbf{y}$ leads to the conditional distribution $p(\mathbf{x}|\mathbf{y})$. Further we initialize $\mathbf{x}$ by setting $\mathbf{x} = \mathbf{y}$.

Algorithm 1: ICM-iteration

```
for j=1,...,D,...1 do
    energyPosSign ← E(x_{x_j=+1}, y);
    energyNegSign ← E(x_{x_j=-1}, y);
    if energyPosSign < energyNegSign then
        x_j ← +1;
    end
    else
        x_j ← −1;
    end
end
```

This will be repeated until no changes in $\mathbf{x}$ appear anymore. This method leads to a local maximum but not necessarily to the global one.

This technique is called *iterated conditional models* (ICM).
Another (more efficient) technique is the *graph-cut*-algorithm, which results are shown in Figure 16d.

3.4 Relation to directed graphs

Considering the two graphs within Figure 18, we get the corresponding joint distributions

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)p(x_3|x_2)...p(x_N|x_{N-1})$$

from Figure 18a and

$$p(\mathbf{x})\frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)...\psi_{N-1,N}(x_{N-1}, x_N)$$

from Figure 18b. By identifying

$$\begin{aligned}
\psi_{1,2}(x_1, x_2) &= p(x_1)p(x_2|x_1) \\
\psi_{2,3}(x_2, x_3) &= p(x_3|x_2) \\
&\vdots \\
\psi_{N-1,N}(_{N-1}, x_N) &= p(x_N|x_{N-1})
\end{aligned}$$

and setting $Z = 1$, we see that undirected graphs and directed graphs can be equivalent.

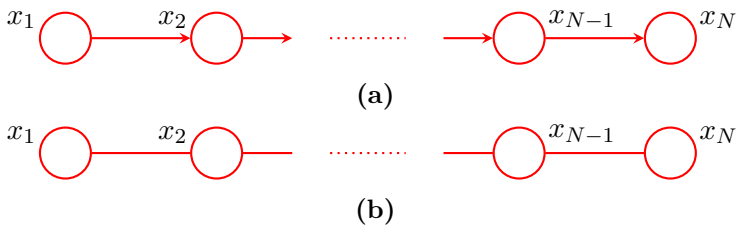

Figure 18 – Example of two equivalent graphs.

Taking a look at Figure 19a, we can write the joint distribution as

$$p(\mathbf{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3).$$

The last factor includes all four variables. To find a corresponding undirected graph, this last factor has to be identified by a potential function including all four variables, so they have to be within one clique.

This can be done by "marrying the parents", where we add additional links between all parents of x_4, shown in Figure 19b. This process is known as *moralization*, the corresponding graph is called *moral graph*. It is used for the "junction tree algorithm" which is used to eliminate cycles by clustering the nodes of the cycle into one single node.

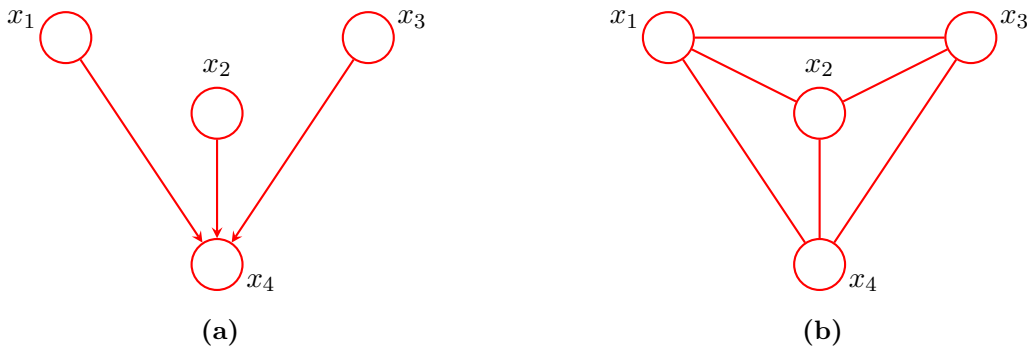

Figure 19 – Example of moralization.

This method of moralization discards conditional independence, so the original graph and the moral graph are not equivalent.

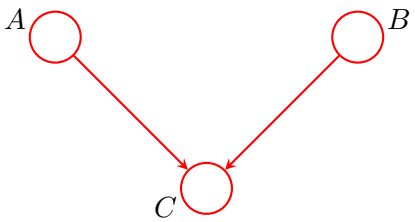

Figure 20 – Example of a directed graph, whose conditional independence properties cannot be shown as undirected graph.

Figure 20 shows a directed graph which corresponding distribution satisfies

$$A \perp\!\!\!\perp B|\emptyset \qquad \text{and} \qquad A \not\perp\!\!\!\perp B|C.$$

There is no equivalent undirected graph, that can do so.

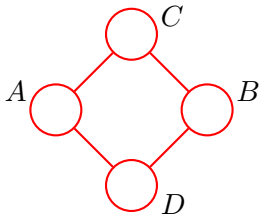

Figure 21 – Example of an undirected graph whose conditional independence properties cannot be shown as directed graph.

Figure 21 shows an undirected graph whose corresponding distribution satisfies

$$A \not\perp\!\!\!\perp B|\emptyset,$$

$$C \perp\!\!\!\perp D|A \cup B$$

and

$$A \not\perp\!\!\!\perp B|C \cup D.$$

As before, there is no equivalent directed graph that can do so.

4 Advantages of probabilistic graphical models

Now we know how probabilistic graphical models work. To sum things up, we take a look at the benefits:

1. The structure of a probabilistic model can be simply visualized, which makes it easy to design and motivate new models.
2. The properties of the model, like conditional independence, can be read out of the graph.
3. To perform inference and learning in sophisticated models, complex computations are needed, which can be expressed in terms of graphical manipulations, which carry underlying mathematical expressions along implicitly.

Probabilistic graphical models are therefore a powerful tool, that can be used in machine learning.