Introduction to Probabilistic Graphical Models

Christoph Lampert

IST Austria (Institute of Science and Technology Austria)
Markov Networks
Markov Networks

So far: write probability as a product of conditional distributions

\[ p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)) \]

- exactly one term per variable
- result is automatically non-negative and normalized
Markov Networks

So far: write probability as a product of conditional distributions

\[ p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)) \]

- exactly one term per variable
- result is automatically non-negative and normalized

More flexible: allow products of arbitrary factors

\[ p(x, y, z) \propto \phi(x, y)\phi(y, z) \]
Markov Networks

So far: write probability as a product of conditional distributions

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i))$$

- exactly one term per variable
- result is automatically non-negative and normalized

More flexible: allow products of arbitrary factors

$$p(x, y, z) \propto \phi(x, y)\phi(y, z)$$

- result is non-negative, if each factor is non-negative, but might not be normalized!
Markov Networks

So far: write probability as a product of conditional distributions

\[ p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)) \]

- exactly one term per variable
- result is automatically non-negative and normalized

More flexible: allow products of arbitrary factors

\[ p(x, y, z) = \phi(x, y)\phi(y, z) \]

- result is non-negative, if each factor is non-negative, but might not be normalized!
- normalization constant \( Z \) or \textit{partition function}

\[ Z = ? \]
Markov Networks

So far: write probability as a product of conditional distributions

\[ p(x_1, \ldots, x_D) = \prod_{i=1}^D p(x_i \mid pa(x_i)) \]

- exactly one term per variable
- result is automatically non-negative and normalized

More flexible: allow products of arbitrary factors

\[ p(x, y, z) = \phi(x, y)\phi(y, z) \]

\[ p(x, y, z) = \frac{1}{Z} \phi(x, y)\phi(y, z) \]

- result is non-negative, if each factor is non-negative, but might not be normalized!
- normalization constant \( Z \) or partition function

\[ Z = \sum_{x,y,z} \phi(x, y)\phi(y, z) \]
Markov Networks

So far: write probability as a product of conditional distributions

\[ p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i \mid pa(x_i)) \]

- exactly one term per variable
- result is automatically non-negative and normalized

More flexible: allow products of arbitrary factors

\[ p(x, y, z) = \phi(x, y)\phi(y, z) \]

\[ p(x, y, z) = \frac{1}{Z} \phi(x, y)\phi(y, z) \]

- result is non-negative, if each factor is non-negative, but might not be normalized!
- normalization constant \(Z\) or partition function

\[ Z = \sum_{x,y,z} \phi(x, y)\phi(y, z) \]

- convenience notation: \( p(x, y, z) \propto \phi(x, y)\phi(y, z) \) "proportional to"
Definitions

Potential

A potential \( \phi(x_1, \ldots, x_D) \) is a non-negative function of the set of variables.

- special case: conditional distributions \( \phi(x_1, \ldots, x_D) = p(x_1|x_2, \ldots, x_D) \) as in belief networks
For a set of variables $\mathcal{X} = \{x_1, \ldots, x_D\}$ a Markov network (or Markov random field) is defined as a product of potentials over the cliques $\mathcal{X}_c$ of the graph $G$

$$p(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

For example:

$$p(a, \ldots, e) \propto \phi_{abc}(a, b, c)\phi_{ab}(a, b)\phi_{cd}(c, d)\phi_c(c)\phi_e(e)$$
For a set of variables $\mathcal{X} = \{x_1, \ldots, x_D\}$ a Markov network (or Markov random field) is defined as a product of potentials over the cliques $\mathcal{X}_c$ of the graph $\mathcal{G}$

$$p(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

For example:

$$p(a, \ldots, e) \propto \phi_{abc}(a, b, c)\phi_{ab}(a, b)\phi_{cd}(c, d)\phi_c(c)\phi_e(e)$$

- Equivalent: use only maximal cliques (with different potentials)

$$p(a, \ldots, e) \propto \phi'_{abc}(a, b, c)\phi'_{cd}(c, d)\phi_e(e)$$
Markov Networks

Directed vs. Undirected

Factor Graphs

Markov Network

For a set of variables $\mathcal{X} = \{x_1, \ldots, x_D\}$ a Markov network (or Markov random field) is defined as a product of potentials over the cliques $\mathcal{X}_c$ of the graph $G$

$$p(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\mathcal{X}_c)$$

For example:

$$p(a, \ldots, e) \propto \phi_{abc}(a, b, c)\phi_{ab}(a, b)\phi_{cd}(c, d)\phi_{c}(c)\phi_{e}(e)$$

- Equivalent: use only maximal cliques (with different potentials)

  $$p(a, \ldots, e) \propto \phi'_{abc}(a, b, c)\phi'_{cd}(c, d)\phi_{e}(e)$$

- Special case: cliques of size 2 – pairwise Markov network
Properties of Markov Networks

\[ p(a, b, c) = \frac{1}{Z} \phi_{ac}(a, c) \phi_{bc}(b, c) \]

Variables are independent if they have no path between them. Otherwise they are usually dependent.

Check (by marginalising over \( c \)): \( p(a, b) \neq p(a)p(b) \).
Properties of Markov Networks

\[ p(a, b, c) = \frac{1}{Z} \phi_{ac}(a, c)\phi_{bc}(b, c) \]

Conditioning on \( c \) makes \( a \) and \( b \) independent. Check: \( p(a, b|c) = p(a|c)p(b|c) \).

Difference to directed model: there, conditioning could introduce dependency:

- for example, \( a \rightarrow c \rightarrow b \), \( a \perp \perp b \), but \( a \not\perp \perp b|c \)
Global Markov Property

Separation
A subset $S$ separates $A$ from $B$ if every path from a member of $A$ to any member of $B$ passes through $S$.

Example: $\{x_4\}$ separates $\{x_1, x_2, x_3\}$ from $\{x_5, x_6, x_7\}$.

Global Markov Property
For disjoint sets of variables $(A, B, S)$ where $S$ separates $A$ from $B$, then $A \perp \perp B \mid S$.

Example: $\{x_1, x_2, x_3, x_4\}$ are conditionally independent of $\{x_7\}$ conditioned on $\{x_5, x_6\}$.
Gibbs Distribution

A probability distribution that can be written in the form \( p(x) = \frac{1}{Z} e^{-E(x)} \) for a function \( E : \mathcal{X} \to \mathbb{R} \) is called \textit{Gibbs distribution}. \( E \) is called \textit{energy function}.

In particular, a Gibbs distribution can only have strictly positive values (i.e. no zero values).

Any Markov network that has only strictly positive potentials is a Gibbs distribution:

\[
p(x_1, \ldots, x_D) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(x_c) = \frac{1}{Z} e^{-E(x_1, \ldots, x_D)}
\]

with energy function

\[
E(x_1, \ldots, x_D) = \sum_c E_c(x_c) \quad \text{for} \quad E_c(x_c) = -\log \phi_c(x_c)
\]

Gibbs distributions are often also written as

\[
p(x_1, \ldots, x_D) = e^{-E(x_1, \ldots, x_D) - \log Z} = e^{-\sum_c \log \phi_c(x_c) - \log Z}
\]
Local Markov Property

For Markov networks that are Gibbs distributions, the so-called local Markov property holds

$$p(x \mid \mathcal{X} \setminus \{x\}) = p(x \mid ne(x))$$
For Markov networks that are Gibbs distributions, the so-called local Markov property holds.

The set of neighboring nodes $ne(x)$ is called the Markov blanket.

This also holds for sets of variables $\Rightarrow$ simple independence check by separation.
Local Markov Property – Example

- $p(x_4 \mid x_1, x_2, x_3, x_4, x_5, x_6, x_7) = p(x_4 \mid x_2, x_3, x_5, x_6)$
- In other words, $x_4 \perp \perp \{x_1, x_7\} \mid \{x_2, x_3, x_5, x_6\}$
Local Markov Property – Example

\[ p(x_4 \mid x_1, x_2, x_3, x_4, x_5, x_6, x_7) = p(x_4 \mid x_2, x_3, x_5, x_6) \]

- in other words \( x_4 \perp \perp \{x_1, x_7\} \mid \{x_2, x_3, x_5, x_6\} \)
- and others
The Hammersley-Clifford Theorem

We know:

- Every Gibbs distribution that is defined with respect to a graph $G$ has certain conditional independencies (the local Markov property).
The Hammersley-Clifford Theorem

We know:

- Every Gibbs distribution that is defined with respect to a graph $\mathcal{G}$ has certain conditional independencies (the local Markov property).

The opposite also holds!

**Hammersely-Clifford Theorem** [Hammersley, Clifford, 1971]

Every positive distribution that fulfills the local Markov property with respect to a graph $\mathcal{G}$ can be written as a Markov network over $\mathcal{G}$.
Directed vs Undirected who wins?
Bayes or Markov?

- So which one is better? Directed or Undirected?
- Both directed and undirected graphical models imply sets of conditional independences
Bayes or Markov?

- So which one is better? Directed or Undirected?
- Both directed and undirected graphical models imply sets of conditional independences
- Which one models more distributions? Or are they the same?
Bayes or Markov?

- So which one is better? Directed or Undirected?
- Both directed and undirected graphical models imply sets of conditional independences
- Which one models more distributions? Or are they the same?
- First introduce “canonical” representation
A graph is said to be a D map (dependency map) of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph.
A graph is said to be a **D map** (dependency map) of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph.

The graph on the right specifies one conditional independence relation: $x_1 \perp x_2 | x_3$.
The graph on the right specifies one conditional independence relation: $x_1 \independent x_2 \mid x_3$

$\Rightarrow$ it is a D map for every distribution that fulfills this independence or less (i.e. none)
D Map

A graph is said to be a D map (dependency map) of a distribution if every conditional independence statement satisfied by the distribution is reflected in the graph.

- The graph on the right specifies one conditional independence relation: \( x_1 \perp\!\!\!\perp x_2 | x_3 \)
- \( \Rightarrow \) it is a D map for every distribution that fulfills this independence or less (i.e. none)
- A completely disconnected graph contains all possible independence statements for its variables
- \( \Rightarrow \) it is a trivial D map for any distribution
A graph is said to be a I \textit{map} (independence map) of a distribution if every conditional independence implied by the graph is satisfied by the distribution.
I Map

A graph is said to be a I map (independence map) of a distribution if every conditional independence implied by the graph is satisfied by the distribution.

- The graph on the right specifies one conditional independence relation: $x_1 \perp \perp x_2 | x_3$
A graph is said to be a I map (independence map) of a distribution if every conditional independence implied by the graph is satisfied by the distribution.

The graph on the right specifies one conditional independence relation: $x_1 \perp \perp x_2 | x_3$

⇒ it is a I map for every distribution that fulfills this independence or more
A graph is said to be an I-map (independence map) of a distribution if every conditional independence implied by the graph is satisfied by the distribution. For example:

- The graph on the right specifies one conditional independence relation: $x_1 \perp \perp x_2 | x_3$
- It is an I-map for every distribution that fulfills this independence or more.
- A fully connected graph implies no independence statements.
- It is a trivial I-map for any distribution.
Relationship directed – undirected models: maps

**Perfect Map**

If every conditional independence property of the distribution is reflected in the graph, **and vice versa**, then the graph is said to be a **perfect map** for that distribution.
Relationship directed – undirected models: maps

Perfect Map
If every conditional independence property of the distribution is reflected in the graph, and vice versa, then the graph is said to be a perfect map for that distribution.

- A perfect map: Both I map and a D map of the distribution
Relationship directed – undirected GM

- $P$ – set of all distributions for a given set of variables
Relationship directed – undirected GM

- $P$ – set of all distributions for a given set of variables
- Distributions that can be represented as a perfect map
  - using undirected graph – $U$
  - using a directed graph – $D$
Middle: conditional independence properties cannot be expressed using an undirected graph over the same three variables
Middle: conditional independence properties cannot be expressed using an undirected graph over the same three variables

Right: conditional independence properties cannot be expressed using a directed graph over the same four variables
▶ How to form the smallest undirected model that is \textit{at least as powerful} as a)?
How to form the smallest undirected model that is at least as powerful as a)?

b) "Moralize" the graph, i.e. connect unconnected parents.
How to form the smallest undirected model that is at least as powerful as a)?

b) "Moralize" the graph, i.e. connect unconnected parents.

c) Remove arrows.
- How to form the smallest undirected model that is at least as powerful as a)?

b) ”Moralize” the graph, i.e. connect unconnected parents.

c) Remove arrows.

c) is the 'smallest' undirected model that can represent all distributed that a) can. There's many others, e.g. fully connected.
Factor Graphs
Consider \[ p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a) \]
Consider \( p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a) \)

What is the graph of the corresponding Markov network?
Consider \( p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a) \)

What is the graph of the corresponding Markov network?

How about this one? \( p(a, b, c) = \phi(a, b, c) \)
Consider $p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a)$

What is the graph of the corresponding Markov network?

▶ The same!

▶ How about this one? $p(a, b, c) = \phi(a, b, c)$

▶ The same!
Consider \( p(a, b, c) = \phi(a, b)\phi(b, c)\phi(c, a) \)

What is the graph of the corresponding Markov network?

How about this one? \( p(a, b, c) = \phi(a, b, c) \)

The same!

no one-to-one relation between the graph and the factorization of the potential functions!
Relationship Factorizations to Graphs

Why is this a problem?
- Many problems have only small (e.g. pairwise) interactions, e.g. "friendship" in Facebook
Many problems have only small (e.g. pairwise) interactions, e.g. "friendship" in Facebook

\[ p(x_1, \ldots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j) \quad \text{with} \quad x_i \in \{1, \ldots, L\} \]

\[ \binom{6}{2} = 15 \text{ factors of size 2} \quad \rightarrow \quad \text{distribution specified by } 15L^2 \text{ values} \]
Relationship Factorizations to Graphs

Why is this a problem?

- Many problems have only small (e.g. pairwise) interactions, e.g. "friendship" in Facebook

\[ p(x_1, \ldots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j) \quad \text{with } x_i \in \{1, \ldots, L\} \]

- \( \binom{6}{2} = 15 \) factors of size 2 \( \rightarrow \) distribution specified by \( 15L^2 \) values

- corresponding graph: fully connected

\[ \begin{array}{c}
\text{x}_1 \\
\text{x}_2 \\
\text{x}_3 \\
\text{x}_4 \\
\text{x}_5 \\
\text{x}_6 \\
\end{array} \]
Why is this a problem?

- Many problems have only small (e.g. pairwise) interactions, e.g. "friendship" in Facebook
  
  \[ p(x_1, \ldots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j) \quad \text{with } x_i \in \{1, \ldots, L\} \]

  \[ \binom{6}{2} = 15 \text{ factors of size 2} \quad \rightarrow \quad \text{distribution specified by } 15L^2 \text{ values} \]

- corresponding graph: fully connected

- also compatible with, e.g.,
  
  \[ p(x_1, \ldots, x_6) = \frac{1}{Z} \phi(x_1, x_2, x_3, x_4) \phi(x_1, x_2, x_5, x_6) \phi(x_3, x_4, x_5, x_6) \quad \rightarrow \quad 3L^4 \text{ values!} \]

- or even \( p(x_1, \ldots, x_6) = \frac{1}{Z} \phi(x_1, \ldots, x_6) \quad \rightarrow \quad L^6 \text{ values!} \)
Why is this a problem?

- Many problems have only small (e.g. pairwise) interactions, e.g. "friendship" in Facebook
- \[ p(x_1, \ldots, x_6) = \frac{1}{Z} \prod_{i \neq j} \phi_{ij}(x_i, x_j) \quad \text{with } x_i \in \{1, \ldots, L\} \]
- \( \binom{6}{2} = 15 \) factors of size 2 \( \rightarrow \) distribution specified by \( 15L^2 \) values
- corresponding graph: fully connected

- also compatible with, e.g.,
  \[ p(x_1, \ldots, x_6) = \frac{1}{Z} \phi(x_1, x_2, x_3, x_4)\phi(x_1, x_2, x_5, x_6)\phi(x_3, x_4, x_5, x_6) \rightarrow 3L^4 \text{ values!} \]
  \[ \text{or even } p(x_1, \ldots, x_6) = \frac{1}{Z} \phi(x_1, \ldots, x_6) \rightarrow L^6 \text{ values!} \]

The graph alone does not tell us if the model is tractable or not. So why bother with it???
We overcome his by augmenting the notation.
We introduce an extra node (a square) for each factor in the factorization. The square is connected to all nodes contributing to the factor.

\[(a), (b), (c)\]

(a): Markov Network graph
We overcome his by augmenting the notation. We introduce an extra node (a square) for each factor in the factorization. The square is connected to all nodes contributing to the factor.

- (a): Markov Network graph
- (b): Factor graph representation of $p(a, b, c) \propto \phi(a, b, c)$
We overcome his by augmenting the notation.

- We introduce an extra node (a square) for each factor in the factorization
  The square is connected to all nodes contributing to the factor.

(a): Markov Network graph
(b): Factor graph representation of $p(a, b, c) \propto \phi(a, b, c)$
(c): Factor graph representation of $p(a, b, c) \propto \phi(a, b)\phi(b, c)\phi(c, a)$
We overcome his by augmenting the notation.

We introduce an extra node (a square) for each factor in the factorization. The square is connected to all nodes contributing to the factor.

- (a): Markov Network graph
- (b): Factor graph representation of $p(a, b, c) \propto \phi(a, b, c)$
- (c): Factor graph representation of $p(a, b, c) \propto \phi(a, b)\phi(b, c)\phi(c, a)$

Different factor graphs can have the same Markov network $(b,c) \Rightarrow (a)$
Directed Factor Graphs

- This also works for directed graph / belief network.
- The structure of the factorization is retained:

- But doesn’t add much information, so typically not used.
Factor Graph Definition

Factor Graph

Given a function

\[ f(x_1, \ldots, x_n) = \prod_i \psi_i(x_i), \]

the factor graph (FG) has a node (represented by a square) for each factor \( \psi_i(x_i) \) and a variable node (represented by a circle) for each variable \( x_j \).
Factor Graph Definition

**Factor Graph**

Given a function

\[ f(x_1, \ldots, x_n) = \prod_i \psi_i(x_i), \]

the factor graph (FG) has a node (represented by a square) for each factor \( \psi_i(x_i) \) and a variable node (represented by a circle) for each variable \( x_j \). When used to represent a distribution

\[ p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_i \psi_i(x_i), \]

a normalization constant is assumed.
A bipartite graph is a graph whose vertices can be divided into two disjoint sets $U$ and $V$ such that every edge connects a vertex in $U$ to one in $V$.

Factor graphs are bipartite graphs. Edge are always between a variables node (circle) and a factor node (square).
Question: which distribution?

Answer:
Question: which distribution?

Answer:

$$p(x) = \frac{1}{Z} f_a(x_1, x_2)f_b(x_1, x_2)f_c(x_2, x_3)f_d(x_3)$$
factor graph: example 2

Question: Which factor graph?

\[ p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2) \]

Answer:
Question: Which factor graph?

\[ p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 | x_1, x_2) \]

Answer:
Example: A Factor Graph and Energy Function for Image Denoising

\[ p(x, y) = \frac{1}{Z} e^{-E(x, y)} \]

\[ E(x, y) = \sum_{i \in \{\text{pixels}\}} E_i(x_i, y_i) + \sum_{(i, j) \in \{\text{edges}\}} E_{ij}(y_i, y_j) \]

Pairwise Markov Random Field (MRF):

- \( E_i(x_i, y_i) = \alpha (x_i - y_i)^2 \) outputs are likely similar to inputs
- \( E_{ij}(y_i, y_j) = \beta |y_i - y_j| \) neighboring outputs are likely similar to each other \( \rightarrow \) smooth output
- \( \alpha \in \mathbb{R} \) and \( \beta \in \mathbb{R} \) can be adjusted per image
Example: A Factor Graph and Energy Function for Human Pose Estimation

\[ p(y|x) = \frac{1}{Z} e^{-E(y;x)} \]

\[ E(y; x) = \sum_{i \in \{\text{head, torso, \ldots} \}} E_i(y_i; x_i) + \sum_{(i,j)} E_{ij}(y_i, y_j) \]

- Unary factors (depend on one label): appearance
  - e.g. \( E_{\text{head}}(y; x) \) "Does location y in image x look like a head?"

- Pairwise factors (depend on two labels): geometry
  - e.g. \( E_{\text{head-torso}}(y_{\text{head}}, y_{\text{torso}}) \) "Is location \( y_{\text{head}} \) above location \( y_{\text{torso}} \)?"
Example: A Factor Graph and Energy Function for Image Segmentation

\[ p(y|x) = \frac{1}{Z} e^{-E(y;x)} \]

\[ E(y; x) = \sum_{i \in \{\text{pixels}\}} E_i(y_i; x_i) + \sum_{(i,j) \in \{\text{edges}\}} E_{ij}(y_i, y_j) \]

Energy function components ("Ising" model):

\[ E_i(y_i = 1, x_i) = \begin{cases} 
\text{low} & \text{if } x_i \text{ is the right color, e.g. brown} \\
\text{high} & \text{otherwise} 
\end{cases} \]

\[ E_i(y_i = 0, x_i) = -E_i(y_i = 1, x_i) \]

\[ E_i(y_i, y_j) = \begin{cases} 
\text{low} & \text{if } y_i = y_j \\
\text{high} & \text{otherwise} 
\end{cases} \]

Higher probability if neighbors have same labels → smooth labelings
The graphs of graphical models represent families of probability distributions graphically:

- **Bayesian networks**: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist

## Summary (so far)

The graphs of graphical models represent families of probability distributions graphically:

- **Bayesian networks**: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- Markov networks: undirected, local cliques of dependent variables
The graphs of graphical models represent families of probability distributions graphically:

- **Bayesian networks**: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist

- **Markov networks**: undirected, local cliques of dependent variables

- **Factor graphs**
  - makes the factorization explicit
  - not a larger class of distributions, “just” a different way of drawing the graph
The graphs of graphical models represent families of probability distributions graphically:

- Bayesian networks: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- Markov networks: undirected, local cliques of dependent variables
- Factor graphs
  - makes the factorization explicit
  - not a larger class of distributions, “just” a different way of drawing the graph
- for modeling undirected models, thinking in terms of factor graphs is very useful
Summary (so far)

The graphs of graphical models represent *families of probability distributions graphically*:

- **Bayesian networks**: directed acyclic graphs, product of conditional distribution
  - by default, arrows have no causal interpretation
  - but: causal Bayesian networks also exist
- **Markov networks**: undirected, local cliques of dependent variables
- **Factor graphs**
  - makes the factorization explicit
  - not a larger class of distributions, “just” a different way of drawing the graph
- for modeling undirected models, thinking in terms of factor graphs is very useful

To specify an actual distribution, we also have to provide:

- for directed models: the conditional tables
- for undirected models: the potentials

Often, these are learned from training data (while the graph structure is fixed manually).