Machine Learning: Probabilistic Graphical Models 2

Hao Tang

March 19, 2026

Recap

- **Definition.** A distribution p factorizes according to a directed graph G if

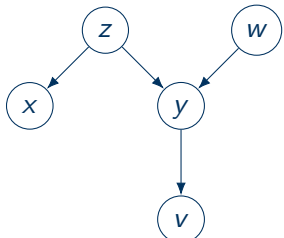
$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | \mathrm{Pa}_{x_i}) \tag{1}$$

  where Pa_x is the set of parent vertices of x in G.

- Obviously, the graph G needs to have the same number of vertices as there are in the distribution p.

- The vertex u is one of v's parent if there is an edge from u to v.

- What factorization can we read off from this graph?

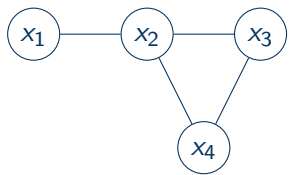


- The above graph represents the distribution

$$p(v, w, x, y, z) = p(v|y)p(x|z)p(y|z, w)p(z)p(w). \tag{2}$$

Factorization represented as undirected graphs

- There is another family of graphical models based on undirected graphs.
- The following graph



represents the factorization of the distribution

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4). \tag{3}$$

- In the distribution

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4), \tag{4}$$

the functions ϕ_1 and ϕ_2 are called potential functions, and the constant Z is called the partition function.

- In the distribution

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4), \tag{4}$$

the functions ϕ_1 and ϕ_2 are called potential functions, and the constant Z is called the partition function.

- Note that ϕ_1 and ϕ_2 can return any real value, and it is necessary to have

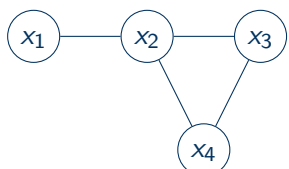
$$Z = \sum_{x_1, x_2, x_3, x_4} \phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4). \tag{5}$$

Otherwise, p is not a valid probability distribution.

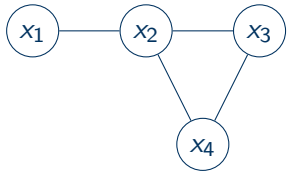
- The distribution

$$p(x_1, x_2, x_3, x_4) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4) \tag{6}$$

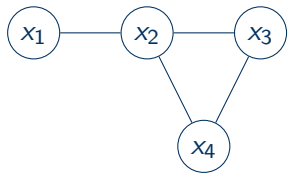
factorizes according to maximal cliques in the following graph.



- A clique is a subset of vertices where every pair is connected.
- A maximal clique is a clique that is not included in another clique.
- In the graph below,

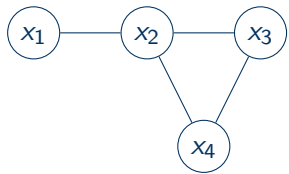


- A clique is a subset of vertices where every pair is connected.
- A maximal clique is a clique that is not included in another clique.
- In the graph below,



$\{x_1, x_2\}$ is a maximal clique, $\{x_2, x_3\}$ is a clique but is not maximal, and $\{x_2, x_3, x_4\}$ is a maximal clique.

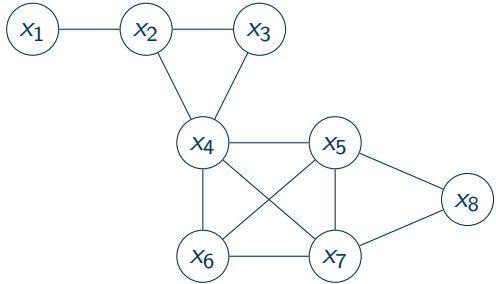
- A clique is a subset of vertices where every pair is connected.
- A maximal clique is a clique that is not included in another clique.
- In the graph below,



  $\{x_1, x_2\}$ is a maximal clique, $\{x_2, x_3\}$ is a clique but is not maximal, and $\{x_2, x_3, x_4\}$ is a maximal clique.

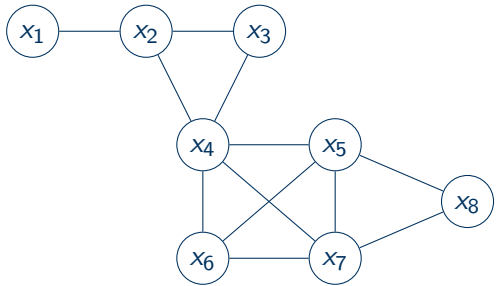
- Finding all maximal cliques in general is unfortunately NP-complete.

- The following graph



represents the factorization of the distribution

- The following graph



represents the factorization of the distribution

$$p(x_1, x_2, \ldots, x_8) = \frac{1}{Z}\phi_1(x_1, x_2)\phi_2(x_2, x_3, x_4)\phi_3(x_4, x_5, x_6, x_7)\phi_4(x_5, x_7, x_8). \quad (7)$$

- **Definition.** A distribution is said to factorize according to an undirected graph G if

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i=1}^{K} \phi_i(C_i), \tag{8}$$

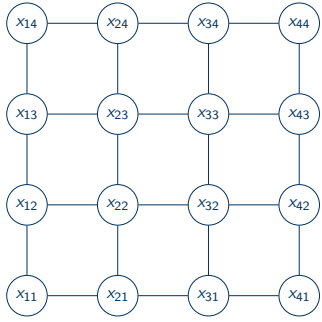
where $C_i \subseteq \{x_1, \ldots, x_n\}$ is a maximal clique in G and

$$Z = \sum_{x_1, \ldots, x_n} \prod_{i=1}^{K} \phi_i(C_i). \tag{9}$$

- An undirected graph paired with a distribution that factorizes accordingly is called a **Markov random field**.

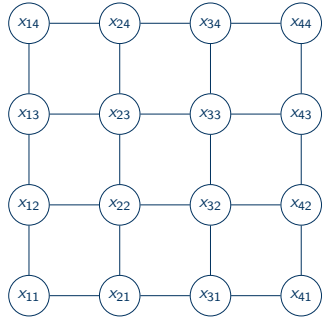
- Terms like random fields and partition function are taken from statistical mechanics.

- The Ising model is a graph



paired with the distribution

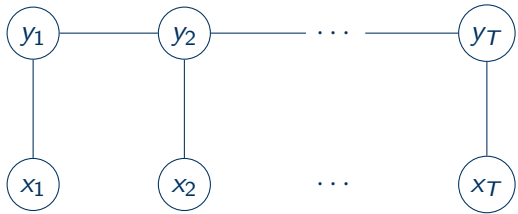
- The Ising model is a graph



paired with the distribution

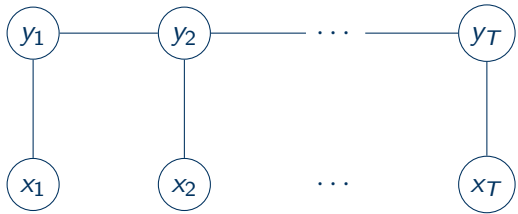
$$p(x_{11}, \ldots, x_{44}) = \frac{1}{Z} \prod_{(i,j) \in E} \phi(x_i, x_j). \tag{10}$$

- The linear-chain conditional random fields is a graph



paired with the distribution

- The linear-chain conditional random fields is a graph



paired with the distribution

$$p(y_1, \ldots, y_T | x_1, \ldots, x_T) = \frac{1}{Z(x_1, \ldots, x_T)} \phi(x_1, y_1) \prod_{t=2}^{T} \phi(y_{t-1}, y_t) \phi(x_t, y_t). \quad (11)$$

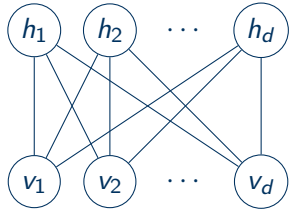
Reading off statistical independencies from an undirected graph

- There are no child-parent relationships.
- A path is blocked if any vertex on the path is given.
- Two variables are separated if all paths between the two variables are blocked.

- If a distribution matches all the independencies on a directed graph, then the distribution factorizes according to the graph.

- **Theorem (Hammersley–Clifford).** If a distribution matches all the independencies on an undirected graph and the distribution is strictly positive, then the distribution factorizes according to the graph.

- This graph



over binary random variables paired with the distribution

$$p(v, h) = \frac{1}{Z} \exp(-v^\top W h - a^\top h - b^\top v) \tag{12}$$

is called a restricted Boltzmann machine.

- We can see that $h_i \perp\!\!\!\perp h_j \mid v$ and $v_i \perp\!\!\!\perp v_j \mid h$.
- In fact,

$$p(v|h) = \frac{1}{Z(h)} \prod_{i=1}^{d} p(v_i|h) \tag{13}$$

and

$$p(h|v) = \frac{1}{Z(v)} \prod_{i=1}^{d} p(h_i|v). \tag{14}$$

Discussion

- The partition function is often difficult to compute.
- Sometimes computing $p(x)$ might be difficult for a Markov random field, but we can still compute $\operatorname{argmax}_x p(x)$.

Discussion

- The partition function is often difficult to compute.
- Sometimes computing $p(x)$ might be difficult for a Markov random field, but we can still compute $\operatorname{argmax}_x p(x)$.
- However, it is non-trivial to generate samples from p, also known as sampling.

- We know that a Gaussian is

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{15}$$

- How do we generate random samples from a Gaussian?

- We know that a Gaussian is

$$p(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{15}$$

- How do we generate random samples from a Gaussian?
- We can take $U_1 \sim U[0,1]$ and $U_2 \sim U[0,1]$, and perform

$$Z_1 = \sqrt{-2\log U_1}\cos(2\pi U_2) \tag{16}$$

$$Z_2 = \sqrt{-2\log U_1}\sin(2\pi U_2). \tag{17}$$

  The two samples Z_1 and Z_2 are two independent variables from standard Gaussian.

- The above is known as the Box–Muller transform.

- A Bayesian network is sometimes described as a "generative story" or a generative model.
- Nowadays, a generative model is a neural network f, such that $f(\epsilon)$ outputs a sample when $\epsilon \sim U[0, 1]$.
- A Bayesian network also does not describe causality.
- There is a different family of models called causal Bayesian network.