Deep Learning - Foundations and Concepts

Chapter 11. Structured Distributions

nonlineark@github

March 14, 2025

Outline

Graphical models

The framework of probabilistic graphical models allows structured probability distributions to be expressed in graphical form:

- They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
- Insights into the properties of the model, including conditional independence properties, can be obtained by inspecting the graph.
- The complex computations required to perform inference and learning in sophisticated models can be expressed in terms of graphical operations.

Directed graphs

- In a probabilistic graphical model, each node represents a random variable, and the links express probabilistic relationships between these variables.
- Directed graphical models (Bayesian networks, or Bayes nets): The graphs have a particular direction indicated by arrows, useful for expressing causal relationships between random variables (the focus of this chapter).
- Undirected graphical models (Markov random fields): The links do not carry arrows and have no directional significance, useful for expressing soft constraints between random variables.

Factorization

Consider a joint distribution $p(a, b, c)$ over three variables a, b and c. We can write the joint distribution in the form:

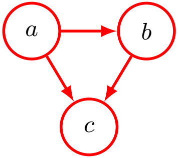
$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

which can be represented in terms of a simple graphical model as follows:

- Introduce a node for each of the random variables a, b and c.
- If a random variable y is conditioned on another random variable x, then add a directed link from x to y. We say that x is the parent of y, and y is the child of x.

Factorization

Figure: A directed graphical model representing the decomposition $p(a, b, c) = p(c|a, b)p(b|a)p(a)$



Factorization

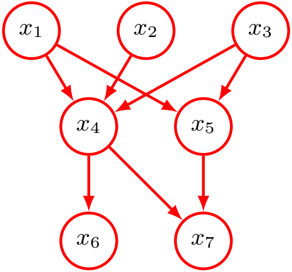
A directed graph also defines a joint distribution given by the product, over all of the nodes of the graph, of a conditional distribution for each node conditioned on the variables corresponding to the parents of that node in the graph. Thus for a graph with K nodes, the joint distribution is given by:

$$p(x_1, \ldots, x_K) = \prod_{k=1}^{K} p(x_k | \mathrm{pa}(k))$$

where $\mathrm{pa}(k)$ denotes the set of parents of x_k.

Factorization

Figure: This directed graph represents the joint distribution
$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$



Discrete variables

Dropping links in the graph reduces the number of independent parameters in a model. Consider two discrete variables x^1 and x^2, each of which has K states. The joint distribution can be written:

$$p(x_1, x_2; \mu) = \prod_{k=1}^{K} \prod_{k'=1}^{K} \mu_{kk'}^{x_k^1 x_{k'}^2}$$

- If there is a link from x^1 to x^2, we need $K^2 - 1$ parameters.
- If x^1 and x^2 are independent, we only need $2(K-1)$ paramters.
- In general, when there are M variables:
  - If their joint distribution is fully connected, we need $K^M - 1$ parameters.
  - If they are independent, we only need $M(K-1)$ parameters.

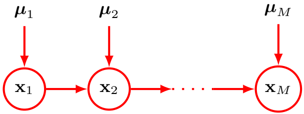
Discrete variables

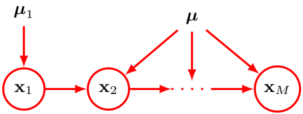
Figure: By dropping the link from x^1 to x^2, the number of parameters needed dropped from $K^2 - 1$ to $2(K-1)$

Discrete variables

An alternative way to reduce the number of independent parameters in a model is by sharing parameters:

- For the graphical model on the left, we need $K - 1 + (M-1)K(K-1)$ parameters.
- For the graphical model on the right, we only need $K - 1 + K(K-1) = K^2 - 1$ paramters.

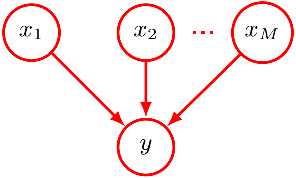




Discrete variables

Another way to reduce the number of independent parameters in a model is by using parameterized representations for the conditional distributions instead of complete tables of conditional probability values. For the example graph, assuming x_m are binary variables:

- If using complete tables, we need 2^M parameters.
- If using parameterized representation $p(y = 1|x_1, \ldots, x_M) = \sigma(w_0 + \sum_{m=1}^{M} w_m x_m)$, we only need $M + 1$ parameters.



Gaussian variables

For graphical models in which the nodes represent continuous variables having Gaussian distributions, we consider linear Gaussian models:

$$p(x_i|\mathrm{pa}(i)) = \mathcal{N}(x_i; \sum_{j\in\mathrm{pa}(i)} w_{ij}x_j + b_i, v_i)$$

where w_{ij} and b_i are parameters governing the mean and v_i is the variance of the conditional distribution for x_i. It's easy to see that the joint distribution is a multivariate Gaussian:

$$\begin{aligned}
-\log p(x_1, \ldots, x_D) &= -\log \prod_{i=1}^{D} p(x_i|\mathrm{pa}(i)) \\
&= \frac{1}{2}\sum_{i=1}^{D} \frac{1}{v_i}(x_i - \sum_{j\in\mathrm{pa}(i)} w_{ij}x_j - b_i)^2 + \frac{1}{2}\sum_{i=1}^{D} \log v_i + \frac{D}{2}\log 2\pi
\end{aligned}$$

Gaussian variables

Let's calculate $E(x_i)$ and $\mathrm{cov}(x_i, x_j)$:

$$
\begin{aligned}
E(x_i) &= \int x_i p(x) \mathrm{d}x = \int x_i \prod_{k=1}^{D} p(x_k|\mathrm{pa}(k)) \mathrm{d}x \\
&= \int \prod_{k=1}^{i-1} p(x_k|\mathrm{pa}(k)) (\int x_i p(x_i|\mathrm{pa}(i)) \mathrm{d}x_i) \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \\
&= \int (\sum_{j \in \mathrm{pa}(i)} w_{ij} x_j + b_i) \prod_{k=1}^{i-1} p(x_k|\mathrm{pa}(k)) \mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \\
&= \int (\sum_{j \in \mathrm{pa}(i)} w_{ij} x_j + b_i) p(x) \mathrm{d}x \\
&= \sum_{j \in \mathrm{pa}(i)} w_{ij} E(x_j) + b_i
\end{aligned}
$$

Gaussian variables

For $i < j$:

$$
\begin{aligned}
E(x_i x_j) &= \int x_i x_j p(x) \mathrm{d}x = \int x_i x_j \prod_{l=1}^{D} p(x_l|\mathrm{pa}(l)) \mathrm{d}x \\
&= \int x_i \prod_{l=1}^{j-1} p(x_l|\mathrm{pa}(l)) (\int x_j p(x_j|\mathrm{pa}(j)) \mathrm{d}x_j) \mathrm{d}x_1 \cdots \mathrm{d}x_{j-1} \\
&= \int (\sum_{k \in \mathrm{pa}(j)} w_{jk} x_k + b_j) x_i \prod_{l=1}^{j-1} p(x_l|\mathrm{pa}(l)) \mathrm{d}x_1 \cdots \mathrm{d}x_{j-1} \\
&= \int (\sum_{k \in \mathrm{pa}(j)} w_{jk} x_k + b_j) x_i p(x) \mathrm{d}x \\
&= \sum_{k \in \mathrm{pa}(j)} w_{jk} E(x_i x_k) + b_j E(x_i)
\end{aligned}
$$

Gaussian variables

$$
\begin{aligned}
E(x_i^2) &= \int x_i^2 p(x)\mathrm{d}x = \int x_i^2 \prod_{l=1}^{D} p(x_l|\mathrm{pa}(l))\mathrm{d}x \\
&= \int \prod_{l=1}^{i-1} p(x_l|\mathrm{pa}(l))(\int x_i^2 p(x_i|\mathrm{pa}(i))\mathrm{d}x_i)\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \\
&= \int ((\sum_{k\in\mathrm{pa}(i)} w_{ik}x_k + b_i)^2 + v_i) \prod_{l=1}^{i-1} p(x_l|\mathrm{pa}(l))\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1} \\
&= \int ((\sum_{k\in\mathrm{pa}(i)} w_{ik}x_k + b_i)^2 + v_i) p(x)\mathrm{d}x \\
&= \sum_{j,k\in\mathrm{pa}(i)} w_{ij}w_{ik}E(x_j x_k) + 2b_i \sum_{k\in\mathrm{pa}(i)} w_{ik}E(x_k) + b_i^2 + v_i
\end{aligned}
$$

Gaussian variables

Finally, for $i \neq j$ we have:

$$\mathrm{cov}(x_i, x_j) = E(x_i x_j) - E(x_i)E(x_j) = \sum_{k \in \mathrm{pa}(j)} w_{jk} \mathrm{cov}(x_i, x_k)$$

$$\begin{aligned} \mathrm{cov}(x_i, x_i) &= E(x_i^2) - (E(x_i))^2 \\ &= \sum_{j,k \in \mathrm{pa}(i)} w_{ij} w_{ik} \mathrm{cov}(x_j, x_k) + v_i \\ &= \sum_{k \in \mathrm{pa}(i)} w_{ik} \mathrm{cov}(x_i, x_k) + v_i \end{aligned}$$

We can calculate $E(x_i)$ and $\mathrm{cov}(x_i, x_j)$ by starting at the lowest numbered node and working recursively through the graph.

Binary classifier

Suppose a binary classifier model has probability distributions of the form:

$$p(t_1, \dots, t_N, w|x^1, \dots, x^N; \lambda) = p(w; \lambda) \prod_{n=1}^{N} p(t_n|x^n; w)$$

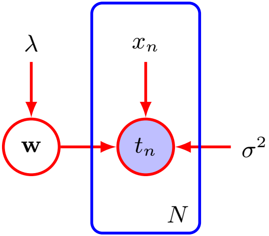
$$p(w; \lambda) = \mathcal{N}(w; 0, \lambda I)$$

Figure: Directed graphical model representing the binary classifier model and its more compact version

Parameters and observations

There are three kinds of variables in a directed graphical model:

- Unobserved (also called latent, or hidden) stochastic variables are denoted by open red circles.
- When stochastic variables are observed, so that they are set to specific values, they are denoted by red circles shaded with blue.
- Non-stochastic parameters are denoted by floating variables.



Bayes' theorem

Figure: A graphical representation of Bayes' theorem