Ch 8: Graphical Models

ifding

Abstract

Bayesian Networks, Conditional Independence, Inference in Graphical Models

All of the probabilistic inference and learning manipulations, no matter how complex, amount to repeated application of two equations: the sum rule and the product rule.

In a *probabilistic graphical model*, each node represents a random variable (or group of random variables), and the edges express probabilistic relationships between these variables. *Bayesian networks* are known as directed graphical models, and *Markov random fields* are known as undirected graphical models. Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are better suited to expressing soft constraints between random variables.

1. Bayesian Networks

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

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a) \quad (1)$$

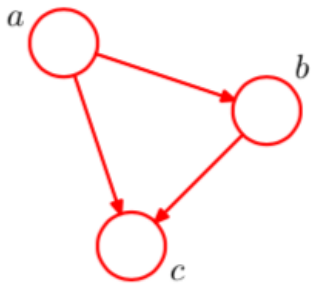

Figure 1: A directed graphical model.

As shown in Figure 1, 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. For a graph with K nodes, the joint distribution is given by

$$p(\mathbf{x}) = \prod_{k=1}^{K} p\left(x_k|\mathrm{pa}_k\right) \quad (2)$$

where pa_k denotes the set of parents of x_k, and $\mathbf{x} = \{x_1, \ldots, x_K\}$. This key equation expresses the *factorization* properties of joint distribution for directed graphical model.

1.1. Example: Polynomial regression

Consider the Bayesian polynomial regression model, the random variables are the vector of polynomial coefficients $\mathbf{w}$ and the observed data $\mathbf{t} = (t_1, \ldots, t_N)^{\mathrm{T}}$, the noise variance σ^2, and the hyperparameter α representing the precision of the Gaussian prior over $\mathbf{w}$. The joint distribution is given by the product of the prior $p(\mathbf{w})$ and N conditional distributions $p(t_n|\mathbf{w})$ for $n = 1, \ldots, N$

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p\left(t_n|\mathbf{w}\right) \quad (3)$$

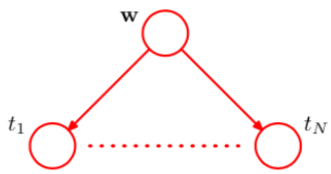

Figure 2: Directed graphical model representing the joint distribution corresponding to the Bayesian polynomial regression model.

We shall find it helpful to make the parameters of a model, as well as its stochastic variables, explicit.

$$p\left(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2\right) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p\left(t_n|\mathbf{w}, x_n, \sigma^2\right) \quad (4)$$

We can make $\mathbf{x}$ and α explicit in the graphical representation. Random variables will be denoted by option circles, and deterministic parameters will be denoted by smaller solid circles. A *plate* (the box labelled N) is introduced to represent N nodes of which only a single example t_n is shown explicitly.

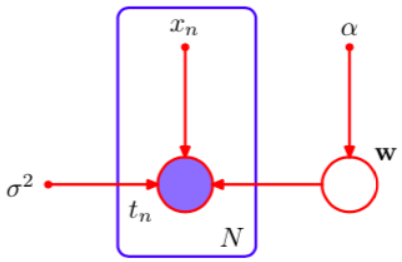

Figure 3: The deterministic parameters shown explicitly by the smaller solid nodes, observed variables $\{t_n\}$ are denoted by shading the corresponding nodes, the value of $\mathbf{w}$ is not observed, and so $\mathbf{w}$ is an example of a *latent* variable.

1.2. Discrete variables

The probability distribution $p(\mathbf{x}|\boldsymbol{\mu})$ for a single discrete variable $\mathbf{x}$ having K possible states (using the 1-of-K representation) is given by

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{5}$$

and is governed by the parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$. Due to the constraint $\sum_k \mu_k = 1$, only $K-1$ values for μ_k need to be specified in order to define the distribution.

2. Conditional Independence

Consider three variables a, b and c,

$$p(a|b,c) = p(a|c) \tag{6}$$

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

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned} \tag{7}$$

The variables a and b are statistically independent, given c. Conditional independence properties play an important role in using probabilistic by simplifying both the structure of a model ans the computations needed to perform inference and learning under that model.

2.1. Three example graphs

The first of the three examples:

$$p(a,b,c) = p(a|c)p(b|c)p(c) \tag{8}$$

If none of the variables are observed, then we can investigate whether a and b are independent by marginalizing both sides with respect to c

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c) \tag{9}$$

Now suppose we condition on the variable c, as represented by the graph of Figure 4, the conditional distribution of a and b, given c, can be written in the form

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned} \tag{10}$$

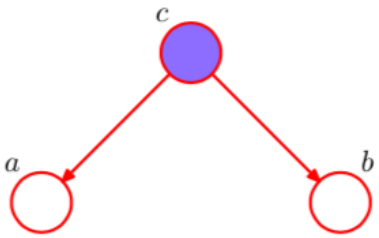

Figure 4: The variables a and b of conditional independence on the value of variable c.

The second example is shown in Figure 5.

$$p(a,b,c) = p(a)p(c|a)p(b|c) \tag{11}$$

First of all, suppose that none of the variables are observed. Again, we can test to see if a and b are independent by marginalizing over c to give

$$p(a,b) = p(a)\sum_c p(c|a)p(b|c) = p(a)p(b|a) \tag{12}$$

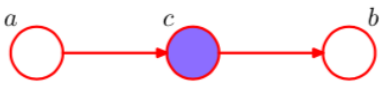

Figure 5: 3-node graphs conditioning on node c.

Now suppose we condition on node c, as shown in Figure 5. Using Bayes' theorem,

$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned} \tag{13}$$

Finally, we consider the third of our 3-node examples, shown by the graph in Figure 6. The joint distribution can again be written down

$$p(a, b, c) = p(a)p(b)p(c|a, b) \tag{14}$$

Consider first the case where none of the variables are observed. Marginalizing both sides over c

$$p(a, b) = p(a)p(b) \tag{15}$$

and so a and b are independent with no variables observed. The conditional distribution of a and b is then given by

$$\begin{aligned} p(a, b|c) &= \frac{p(a, b, c)}{p(c)} \\ &= \frac{p(a)p(b)p(c|a, b)}{p(c)} \end{aligned} \tag{16}$$

which in general does not factorize into the product $p(a)p(b)$.

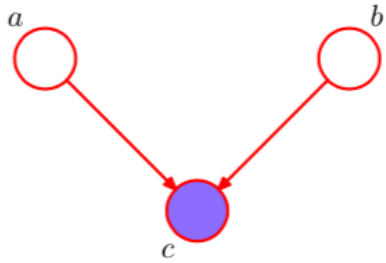

Figure 6: 3-node graphs conditioning on node c. The act of conditioning induces a dependence between a and b.

It's worth spending a moment to understand further the unusual behaviour of the graph of Figure 6. Consider a particular instance of such a graph corresponding to a problem with three binary random variables relating to the fuel system on a car, as shown in Figure 7.

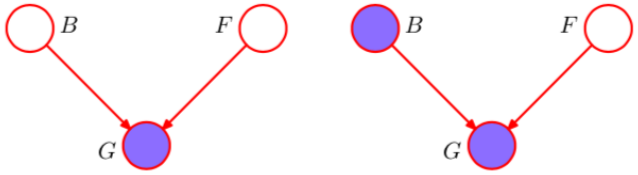

Figure 7: The three nodes represent the state of the battery(B), the state of the fuel tank (F) and the reading on the electric fuel gauge (G).

The B variables represent the state of a battery that is either charged (B=1) or flat (B=0), F represent the state of the fuel tank that is either full (F=1) or empty (F=0), and G indicates either full (G=1) or empty (G=0). The prior probabilities

$$p(B = 1) = 0.9, \;\; p(F = 1) = 0.9 \tag{17}$$

Given B and F, the G reads full with probabilities given by

$$\begin{aligned} p(G = 1|B = 1, F = 1) &= 0.8 \\ p(G = 1|B = 1, F = 0) &= 0.2 \\ p(G = 1|B = 0, F = 1) &= 0.2 \\ p(G = 1|B = 0, F = 0) &= 0.1 \end{aligned} \tag{18}$$

The prior probability of fuel tank being empty is $p(F = 0) = 1 - 0.9 = 0.1$. Now suppose that we observe the fuel gauge and discover that it reads empty, i.e., $G = 0$, corresponding to the left-hand graph in Figure 7.

$$\begin{aligned} p(G = 0) &= \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G = 0|B, F)p(B)p(F) \\ &= (1 - 0.8) * 0.9 * 0.9 + (1 - 0.2) * 0.9 * (1 - 0.9) + \\ &(1 - 0.2) * (1 - 0.9) * 0.9 + (1 - 0.1) * (1 - 0.9) * (1 - 0.9) \\ &= 0.315 \end{aligned} \tag{19}$$

and similarly we evaluate

$$p(G = 0|F = 0) = \sum_{B\in\{0,1\}} p(G = 0|B, F = 0)p(B) = 0.81 \tag{20}$$

and using these results we have

$$p(F = 0|G = 0) = \frac{p(G = 0|F = 0)p(F = 0)}{p(G = 0)} \simeq 0.257 \tag{21}$$

and so $p(F = 0|G = 0) > p(F = 0)$. Thus observing that the gauge reads empty makes it more likely that the tank is indeed empty.

We have now observed the states of the battery and find that it is flat, i.e., $B = 0$. We have now observed the states of both the fuel gauge and the battery, as shown by the right-hand graph in Figure 7.

$$\begin{aligned} &p(F = 0|G = 0, B = 0) \\ &= \frac{p(G = 0|B = 0, F = 0)p(F = 0)p(B = 0)}{\sum_{F\in\{0,1\}} p(G = 0|B = 0, F)p(F)p(B = 0)} \\ &= \frac{p(G = 0|B = 0, F = 0)p(F = 0)}{\sum_{F\in\{0,1\}} p(G = 0|B = 0, F)p(F)} \simeq 0.111 \end{aligned} \tag{22}$$

Thus the probability that the tank is empty has decreased (from 0.257 to 0.111) as a result of the observation of the state of the battery.

3. Inference in Graphical Models

To start with, let us consider the graphical interpretation of the joint distribution $p(x, y) = p(x)p(y|x)$. This can

be represented by the directed graph shown in Figure 8(a). Now suppose we observe the value of y, as indicated by the shaded node in Figure 8(b).

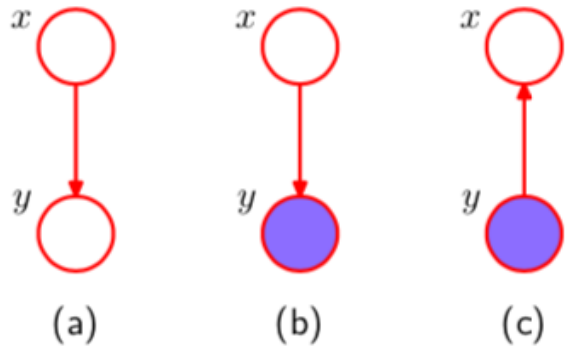

Figure 8: A graphical representation of Bayes' theorem.

We can view the marginal distribution $p(x)$ as a prior over the latent variable x, and our goal is to infer the corresponding posterior distribution over x.

$$p(y) = \sum_{x'} p\left(y|x'\right) p\left(x'\right) \tag{23}$$

which can then be used in Bayes' theorem to calculate

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{24}$$

The direction of the arrow is reversed, as shown in Figure 8(c).

References

[1] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.