Probabilistic graphical models

Kristian Wichmann

October 5, 2017

1 Bayesian networks

A group of graphical models are *Bayesian networks*, also know as *directed graphical models*.

1.1 An example

As an example, consider three random variables A, B, and C with a joint distribution $p(a, b, c)$. Then we can use the multiplication rule of probability to rewrite this as:

$$p(a, b, c) = p(c|a, b)p(a, b) \tag{1.1}$$

Using the same idea once more, this can be further rewritten:

$$p(a, b, c) = p(c|a, b)p(a, b) = p(c|a, b)p(b|a)p(a) \tag{1.2}$$

We can interpret this as c depending on a and b, and b depending on a. We will use directed *edges* to denote this dependency as shown in figure 1.

Do note, that we might as we might as well have split the the variables into dependency in another order! This graph is merely one way of viewing the joint distribution - in general a distribution can be represented by many different graphs.

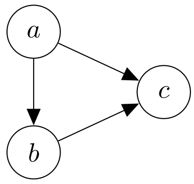

Figure 1: A Bayesian network with three nodes

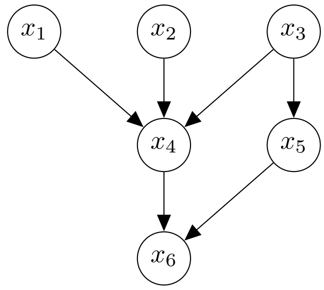

Figure 2: A less than fully connected Bayesian network

1.2 General case: Fully connected

Using the same scheme, we can decompose a joint distribution of K variables and split it into a form similar to equation 1.2 by applying the multiplication rule of probability $K-1$ times:

$$p(x_1, x_2, \cdots, x_K) = p(x_K|x_1, \cdots, x_{K-1}) \cdots p(x_2|x_1)p(x_1) \tag{1.3}$$

This gives rise to a *fully connected* graph, as there will be edges between any two nodes.

1.3 Less than fully connected

However, in general we will be more interested in cases where some pairs of nodes will not have an edge between them. We will still require the graph to be *acyclical*, i.e. there must be no loops when following the directed edges.

As an example, consider the six-node graph in figure 2. The corresponding decomposition would be:

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_3)p(x_6|x_4, x_5) \tag{1.4}$$

This is more generally expressed in the form:

$$p(\mathbf{x}) = \prod_i p(x_i|\mathrm{pa}_i) \tag{1.5}$$

Here $\mathbf{x}$ is a shorthand for all the variables, while pa_i is short for the list of *parent nodes* of x_i, i.e. nodes which have edges pointing to x_i. If there are no parent nodes, this is simply $p(x_i)$.

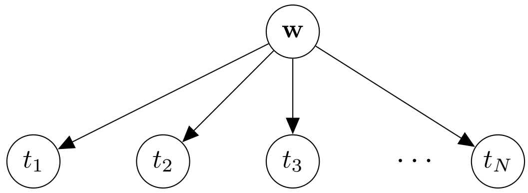

Figure 3: Bayesian regression model

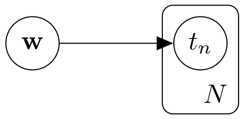

Figure 4: The same model with plate notation

1.4 Example: Polynomial regression

Let's take a look at a more practical model: Bayesian polynomial regression. Here, we have a set of N observation points:

$$\mathbf{x} = (x_1, \cdots, x_N)^t, \quad \mathbf{t} = (t_1, \cdots, t_N)^t \tag{1.6}$$

We think of the t's as being dependent on the x'es: $t = t(x)$. Note that is usually the case in regression models, only the t's are considered random.

The model consists of a series of weights, together denoted $\mathbf{w}$. Since the model is Bayesian, this is a random variable. Now, each t_i depends on x_i and $\mathbf{w}$:

$$t_i = t_i(x_i, \mathbf{w}) \tag{1.7}$$

But remember that x_i is not random. So, at its heart, the random variables of the model depend on each other as shown in the graph of figure 3

Often, having a shorthand for a collection of variables, such as the N t's is useful. This is done by using a *plate* as shown in figure 4.