Probabilistic Graphical Models

Wilker Aziz

February 19, 2018

Here I briefly describe some of the important results about graphical models that are necessary to follow my bachelor's course on natural language models and interfaces. For now, I only cover directed graphical models.

Notation We use capital Roman letters (e.g. X) for random variables (rvs) and lowercase letters for assignments (e.g. x). We use X_1^n as a shorthand for $X_1, \ldots, X_n$ and similarly with x_1^n. We write P_X for probability distributions, and $P_X(X = x)$ for probability values—where we sometimes omit one or both occurrences of X, e.g. $P_X(x)$, $P(X = x)$, or $P(x)$, if no ambiguity is possible.

I focus on discrete probability distributions with finite support throughout. We start with the class of *directed graphical models*, also known as *Bayesian networks*.

1 Directed graphical models

At the highest level our goal is to efficiently represent a joint distribution $P_{X_1,\ldots,X_n}$ over some collection of random variables $X_1, \ldots, X_n$. Before we talk about *efficient* ways of representing it, let's have a look at what it means to represent it *inefficiently*. If no simplifying assumptions can be made, the only way to represent the complete joint distribution is to use a table to store each probability value in its support. The support of the distribution is the set of all possible values its random variables may take on.

For example, suppose that X_i is a binary random variable—each X_i can only take on 1 of 2 values. This means that altogether the collection $X_1, \ldots, X_n$ can take on 2^n different values. If we represent the joint distribution by storing the probability value associated with each and every joint outcome of $X_1, \ldots, X_n$, then we will need 2^n probability values.[1] This representation is called *tabular*, since we can think of the distribution as a big table as in Table 1.

Joint assignments			Probability values
X_1	X_2	X_3	P_{X_1,X_2,X_3}
0	0	0	$P_{X_1,X_2,X_3}(0,0,0)$
0	0	1	$P_{X_1,X_2,X_3}(0,0,1)$
0	1	0	$P_{X_1,X_2,X_3}(0,1,0)$
1	0	0	$P_{X_1,X_2,X_3}(1,0,0)$
0	1	1	$P_{X_1,X_2,X_3}(0,1,1)$
1	1	0	$P_{X_1,X_2,X_3}(1,1,0)$
1	0	1	$P_{X_1,X_2,X_3}(1,0,1)$
1	1	1	$P_{X_1,X_2,X_3}(1,1,1)$

Table 1: Example of tabular joint distribution over 3 binary random variables

In the general case, where $\mathcal{X}_i$ is the support of the ith random variable, and $|\mathcal{X}_i|$ its size, the space of joint assignments will grow proportional to $\prod_{i=1}^{n} |\mathcal{X}_i|$, and that's the cost of our representation—one real number per joint assignment.

Exercise 1

Suppose A may take on 1 out of n values, B may take on 1 out of nm values, and C is a binary variable. What's the representation cost of the joint distribution $P_{A,B,C}$ in the general case?

[1] More precisely, it takes $2^n - 1$ independent probability values, this is because probability values are positive and sum to 1, thus we can determine the last value by computing 1 minus the sum of probabilities of all other values. If asked about the representation cost of this example, I take any of the following as correct: $O(2^n)$, 2^n, or $2^n - 1$.

A key concept in deriving compact representations of joint distributions is that of *conditional independence*. Conditional independence relates 3 collections of random variables, e.g. $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ where we use boldface to denote a collection. If $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$ —which we read as $\mathbf{X}$ *is conditionally independent of* $\mathbf{Y}$ *given* $\mathbf{Z}$ — then it holds that

$$P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Y}|\mathbf{Z}) \tag{1}$$

and by extension it also holds that

$$P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}) = P(\mathbf{X}|\mathbf{Z}) \ . \tag{2}$$

Exercise 2

Prove that conditional independence (1) implies (2).

A directed graphical model, or a Bayesian network (BN), is a directed acyclic graph (DAG) $\mathfrak{G}$, where nodes represent random variables and edges correspond to *direct influence* of one node on another. You may think of a directed graphical model as a recipe to compactly represent a complex joint distribution. The compression in representation is due to the graphical model specifying a *factorisation* of the joint distribution, that is, a way of building it *a factor at a time*.

Let $\mathfrak{G}$ be a directed graphical model over the variables $X_1, \ldots, X_n$, then each random variable X_i (and therefore each node of the graph) has an associated *conditional probability distribution* (cpd). A cpd fully specify the dependency of a random variable on other variables. In particular, the cpd for the ith variable X_i represents the probability of assignments to X_i given assignments to parents of X_i in $\mathfrak{G}$. The set of *parents* of the ith variable is the set of all random variables that are directly connected to X_i—where the arrow points from the parent to X_i. We denote the set of parents by Pa_{X_i}.

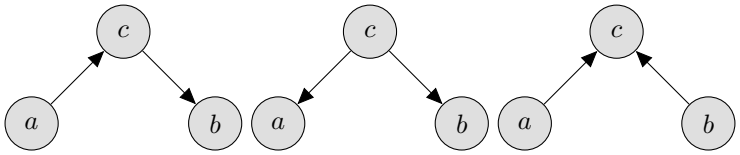

Figure 1: Examples of directed graphical models

Figure 1 illustrates three different BNs involving the same three random variables. For the first BN, on the left,

- A has no parents, i.e., $\mathrm{Pa}_A = \emptyset$;
- C has a single parent, i.e., $\mathrm{Pa}_C = \{A\}$;
- B too has a single parent, i.e., $\mathrm{Pa}_B = \{C\}$.

For the second BN, in the middle,

- C has no parents, i.e. $\text{Pa}_C = \emptyset$;
- A and B both have the same (single) parent C, i.e., $\text{Pa}_A = \text{Pa}_B = \{C\}$.

Finally, for the last BN, on the right,

- A and B have no parents, i.e. $\text{Pa}_A = \text{Pa}_B = \emptyset$;
- C has two parents, i.e., $\text{Pa}_C = \{A, B\}$.

A cpd is a *local probabilistic model*, that is, it is a probabilistic graphical model on its own right, but a small one that only explains a single variable. There are several ways to represent cpds, but for now we are only interested in the *tabular* representation. The idea is exactly the same as we saw earlier, we will use a table of probability values per assignment in order to represent the conditional distribution.

Consider an example where students' grades from 0 to 10 are mapped to a 3-class outcome (e.g. fail, pass, pass with honours), and consider whether or not I give the student a recommendation letter (a binary decision).[2] Table 2 (right) illustrates the set of cpds associated with the variable letter L given the variable grade G. Note that to each outcome $G = g$ we associate a conditional probability distribution over L, that is, we have a probability value associated with each outcome in the support of L and they sum to 1. In other words, every row in the tabular cpd is itself a distribution and we have as many cpds as there are values in the support $\mathcal{G}$ of G—exactly 3 in this case. Each cpd is itself a distribution over the support of L and, in this example, contains 2 probability values. This means that it takes 3×2 probability values (or $3 \times (2-1)$ independent probability values) to represent the complete set of cpds $P_{L|G}$. Table 2 (left) shows the cpd for P_G—you can think of it as a cpd that does not condition on anything (or conditions on the empty set)—note that the cpd sums to 1 as it should.

Grade (G)	P_G
1	0.2
2	0.7
3	0.1

Conditioning context		Letter (L)		
Grade	G	0	1	$\sum_{l=0}^{1} P_{L\|G}(l\|g)$
$[0, 6)$	1	0.8	0.2	1.0
$[6, 8)$	2	0.6	0.4	1.0
$[8, 10]$	3	0.1	0.9	1.0

Table 2: CPDs for P_G (left) and $P_{L|G}$ (right).

Recall that we can use chain rule to recover the complete joint distribution P_{GL} over assignments to G and L. That is, from chain rule $P_{GL}(g, l) = P_G(g)P_L(l|g)$. Figure 2 (left) shows the complete joint distribution obtained by application of chain rule. Note that while P_G and $P_{L|G}$ are explicitly represented—they are stored in tables—the joint distribution P_{GL} is *inferred* by application

[2] Example adapted from Daphne Koller's course on probabilistic graphical models.

Joint assignments		
G	L	P_{GL}
1	0	0.16
2	0	0.42
3	0	0.01
1	1	0.04
2	2	0.28
3	3	0.09

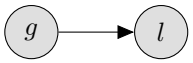

Figure 2: Joint distribution P_{GL}.

of chain rule. In other words, chain rule combines the *marginal* P_G with the conditionals $P_{L|G}$ to produce the joint P_{GL}.

The example just illustrated in fact corresponds to a very simple BN, shown in Figure 2 (right). The BN factorises the joint distribution using two factors—P_G and $P_{L|G}$—which correspond to 4 cpds (one for P_G, one for $P_{L|G=1}$, one for $P_{L|G=2}$, and one for $P_{L|G=3}$). Let's now think about the representation cost. P_G takes $3-1=2$ independent probability values, and $P_{L|G}$ takes $3\times(2-1)=3$ independent probability values, thus our representation of the joint distribution—indirectly via chain rule—takes $2+3=5$ independent probability values. It does not look like we achieved much, since without assuming a particular factorisation we would expect P_{GL} to take $3\times 2-1=5$ independent probability values anyway. The benefits, however, will become clear as we add more and more random variables.

The next result, known as the *BN chain rule*, is the most important result concerning BNs. Given a BN $\mathfrak{G}$, the joint probability $P_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$ factorises as a product

$$P_{X_1,\ldots,X_n}(x_1,\ldots,x_n)=\prod_{i=1}^{n}P(X=x_i|\mathrm{Pa}_{X_i}=\mathrm{pa}_{X_i}) \tag{3}$$

where Pa_{X_i} denotes the set of parents of X_i in $\mathfrak{G}$ and pa_{X_i} are assignments to those parents. This result also implies that $\mathbf{X}\perp\mathrm{NonDescendants}_{\mathbf{X}}\mid\mathrm{Pa}_{\mathbf{X}}$ where $\mathbf{X}$ is any subset of the variables in the BN, $\mathrm{Pa}_{\mathbf{X}}$ the union of their parents, and $\mathrm{Pa}_{\mathbf{X}}$ the union of variables that do not descend from variables in $\mathbf{X}$.

Now with definition (3) in mind, let's get back to the BNs in Figure (1). The first BN (left) implies that the joint distribution factorises as

$$P_{ABC}(a,b,c)=P_A(a)P_{C|A}(c|a)P_{B|C}(b|c)\ . \tag{4}$$

The second BN (middle) implies the factorisation

$$P_{ABC}(a,b,c)=P_C(c)P_{A|C}(a|c)P_{B|C}(b|c)\ . \tag{5}$$

And finally, the third BN (right) factorises the joint distribution as

$$P_{ABC}(a,b,c)=P_A(a)P_B(b)P_{C|AB}(c|a,b)\ . \tag{6}$$

Table 3 shows the representation cost for the different cpds in Figure 1 assuming that A is "one of n", B is "one of m", and C is binary. Recall that a general joint distribution P_{ABC}, if no independence assumptions can be made, takes $O(n \times m)$ parameters to represent (because there are $2 \times n \times m$ unique joint assignments). With the conditional independence assumptions in each of the BNs in Figure 1, we achieve different representation costs. For example, (4) takes $O(n + m)$, (5) also takes $O(n + m)$, but (6) takes $O(n \times m)$. The reason why the last BN is not asymptotically more compact than the complete joint distribution is because it still correlates all 3 variables directly through $P_{C|AB}$. The other two BNs introduce stronger assumptions, for example by never correlating more than 2 variables at a time.

Distribution	# joint assignments	Representation cost
P_A	n	$O(n)$
P_B	m	$O(m)$
P_C	2	$O(1)$
$P_{A\|C}$	$2 \times n$	$O(n)$
$P_{C\|A}$	$n \times 2$	$O(n)$
$P_{B\|C}$	$2 \times m$	$O(m)$
$P_{C\|B}$	$m \times 2$	$O(m)$
$P_{C\|AB}$	$n \times m \times 2$	$O(n \times m)$

Table 3: Representation cost for cpds in Figure 1 assuming $\mathcal{A} = \{1, \ldots, n\}$, $\mathcal{B} = \{1, \ldots, m\}$, and $\mathcal{C} = \{0, 1\}$.

1.1 Inferences

Equation (3) shows us how to construct the joint distribution out of the cpds specified in a BN. However, we may be interested in other quantities as well, such as marginals and conditionals that are not directly stored in cds. For that, we start from the joint distribution—taking into account the particular factorisation—and proceed by applying standard rules of probability, such as, marginalisation, conditioning, and Bayes rule.

For example, the BN in Figure 1 (left) implies that we only have P_A, $P_{C|A}$, and $P_{B|C}$ stored as tabular cpds. However, we may be interested in computing the marginal distribution P_B, or the marginal P_C, or a conditional such as $P_{A|B}$ or $P_{B|A}$. Let's illustrate this by computing the marginal distribution P_B.

We start from the BN chain rule

$$P_{ABC}(a, b, c) = P_A(a)P_{C|A}(c|a)P_{B|C}(b|c) \tag{7a}$$

then we proceed to marginalise A

$$P_{BC}(b, c) = \sum_{a=1}^{n} P_A(a)P_{C|A}(c|a)P_{B|C}(b|c) \tag{7b}$$

then note that the last factor $P_{B|C}$ does not depend on a, thus we can factor it out (read "push it outside") of the sum

$$= P_{B|C}(b|c) \sum_{a=1}^{n} P_A(a) P_{C|A}(c|a) \tag{7c}$$

then we may proceed to marginalise C

$$P_B(b) = \sum_{c=0}^{1} P_{B|C}(b|c) \left(\sum_{a=1}^{n} P_A(a) P_{C|A}(c|a) \right) \tag{7d}$$

Note that every cpd in the final expression is a tabular cpd for which we can look probability values up directly. Also note that inferring the marginal distribution P_B requires performing the computation in (7d) m times, that is, once per valid assignment of B, and therefore the overall computation runs in time proportional to $O(m \times 2 \times n) = O(n \times m)$—assuming we can look up probability values in constant time—even though we reduced the representation cost of the underlying joint distribution to $O(n + m)$.

If inferring marginals and conditionals may still essentially take $O(n \times m)$, why do we bother? The answer has to do with *learning*. When we rely on parameter estimation derived from observations in order to choose the precise probability values that make up our various tabular cpds (i.e. P_A, $P_{C|A}$, and $P_{B|C}$), we need to be cautious with sparsity. If we do not make assumptions, we will be left with enormous tabular cpds, and we will simply not find enough data to estimate their parameters. Also, some assumptions are actually extremely reasonable and convey knowledge about the problem we are interested in. In fact, we can see probabilistic graphical models as a general framework to express our domain knowledge in terms of probability theory.

Exercise 3

For the BN in Figure 3, write down the joint distribution P_{ABC}, the conditional $P_{BC|A}$ and the conditional $P_{C|A}$. Your expressions must all be a function of the tabular cpds in the BN. Also state the representation cost of the joint distribution as a function of the size of the support of each variable: $n_a = |\mathcal{A}|$, $n_b = |\mathcal{B}|$, and $n_c = |\mathcal{C}|$.

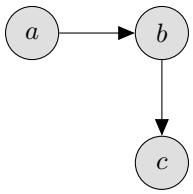

Figure 3

Exercise 4

For the BN in Figure 4, write down the joint distribution $P_{ABCDEFG}$ and the conditional $P_{FG|ABCDE}$. Also state the representation cost of the joint distribution as a function of the size of the support of each variable: $n_a = |\mathcal{A}|$, $n_b = |\mathcal{B}|$, ..., $n_g = |\mathcal{G}|$. Then suppose all variables are "1 of n" for some large n and compare the representation cost of the BN to that of a general joint distribution over 7 such random variables.

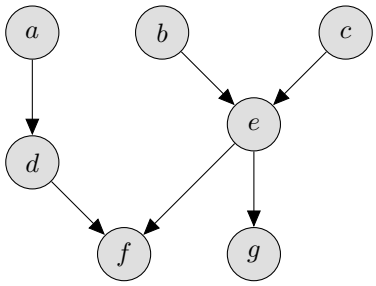

Figure 4

1.2 Answers of exercises

Answer of exercise 3

We have two categorical variables A and B, whose supports are the sets $\mathcal{A} = \{1, \ldots, n\}$ and $\mathcal{B} = \{1, \ldots, nm\}$, and we also have a binary variable, whose support is the set $\mathcal{C} = \{0, 1\}$. The space of joint assignments corresponds to $\mathcal{A} \times \mathcal{B} \times \mathcal{C}$ whose size is $n \times m \times 2$. Therefore, it takes $O(\times n \times m)$ probability values to represent the joint distribution $P_{A,B,C}$.

Answer of exercise 2

Proof.

From the definition of conditional independence (1) we know that

$$P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \underbrace{P(\mathbf{X}|\mathbf{Z})P(\mathbf{Y}|\mathbf{Z})}_{\text{conditional independence}} \tag{8a}$$

from chain rule, we know that

$$P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \underbrace{P(\mathbf{Y}|\mathbf{Z})P(\mathbf{X}|\mathbf{Z}, \mathbf{Y})}_{\text{chain rule}} \tag{8b}$$

if we then replace the left-hand side of the former by the right-hand side of the latter we get

$$P(\mathbf{Y}|\mathbf{Z})P(\mathbf{X}|\mathbf{Z},\mathbf{Y}) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Y}|\mathbf{Z}) \tag{8c}$$
$$P(\mathbf{X}|\mathbf{Z},\mathbf{Y}) = P(\mathbf{X}|\mathbf{Z}) \tag{8d}$$

□

Answer of exercise 3

The joint P_{ABC} factorises as below

$$P_{ABC}(a,b,c) = P_A(a)P_{B|A}(b|a)P_{C|B}(c|b) \tag{9}$$

and it takes $O(n_a + n_a \times n_b + n_b \times n_c)$ probability values to represent.

The conditional $P_{BC|A}$ is shown below

$$P_{BC|A}(b,c|a) = \frac{P_A(a)P_{B|A}(b|a)P_{C|B}(c|b)}{P_A(a)} \tag{10a}$$
$$= P_{B|A}(b|a)P_{C|B}(c|b) \tag{10b}$$

where we first applied the definition of conditional probability, and then cancelled the terms $P_A(a)$ that appear both in the numerator and in the denominator.

The conditional $P_{C|A}$ requires can be obtaining by marginalising B out of $P_{BC|A}$, thus we have

$$P_{C|A}(c|a) = P_{BC|A}(b,c|a) \tag{11a}$$
$$= \sum_{b=1}^{n_b} P_{B|A}(b|a)P_{C|B}(c|b) \tag{11b}$$

by reusing the result just obtained for $P_{BC|A}$. Note that since both terms inside of the sum depend on B (the variable we are marginalising), we cannot further simplify this expression.

Answer of exercise 4

The joint distribution factorises as

$$\begin{aligned} P_{ABCDEFG}(a,b,c,d,e,f,g) = {} & P_A(a)P_B(b)P_C(c) \\ & \times P_{D|A}(d|a) \\ & \times P_{E|BC}(e|b,c) \\ & \times P_{F|DE}(f|d,e) \\ & \times P_{G|E}(g|e) \end{aligned} \tag{12}$$

which takes $O(n_a + n_b + n_c + n_a \times n_d + n_c \times n_b \times n_e + n_e \times n_d \times n_f + n_e \times n_g)$ probability values to represent. Suppose the worst case where all variables are

"1 of n" for some large n, then this takes $O(n^3)$ rather than $O(n^7)$ which would the case had no assumptions been made.

For the conditional $P_{FG|ABCDE}$ it's sufficient to recall that the BN chain rule implies that $\mathbf{X} \perp \text{NonDescendants}_{\mathbf{X}} \mid \text{Pa}_{\mathbf{X}}$ for any subset of variables in the network. Thus, if we take the subset $\{F, G\}$, we see that $\text{Pa}_{\{F,G\}} = \{D, E\}$, $\text{NonDescendants}_{\{F,G\}} = \{A, B, C\}$, and then

$$P_{FG|ABCDE}(f, g|a, b, c, d, e) = P_{FG|DE}(f, g|d, e) \tag{13a}$$
$$= P_{G|E}(g|e)P_{F|DE}(f|d, e) \tag{13b}$$

where the first equality is due to independence on non-descendants A, B, C and the second is due to the factorisation of the BN.

2 Further reading

- Graphical models in a nutshell