SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

01.112 Machine Learning, Fall 2019
Lecture Notes for Week 11

19. Bayesian Networks (II)

Last update: Wednesday 27th November, 2019 18:34

Bayesian networks: graph and independence

We have already emphasized that the graph structure in Bayesian networks represents useful qualitative properties about the variables. Specifically, the graph encodes independence statements about how the variables relate to each other. We will need a criterion for reading such independence properties from the graph without consulting the underlying probability distribution. The subtlety here is that we cannot just pick any criterion we like. The probability distribution we will associate with the graph must be consistent with all the properties we can derive from the graph. Otherwise the graph would "lie" and wouldn't be useful to us.

We had previously introduced the graph as specifying the parents of each node $i = 1, \ldots, d$. In the graph, we call nodes j which start directed edges or arcs to node i as parents of i. Formally, we say that $j \in pa_i$. The choice of these parents, *i.e.*, the choice of parent sets pa_i, is unconstrained except for the fact that we cannot introduce directed cycles. In other words, the graph must be a *directed acyclic graph* (DAG). In terms of variables $X_1, \ldots, X_d$, we motivated the graph as specifying which other variables each X_i directly depends on, *i.e.*, variables whose values we would have to know prior to drawing a value for X_i. We write the set of parents as variables using notation $X_{pa_i} = \{X_j\}_{j \in pa_i}$. Once we know the parents, we can write (factor) the probability distribution over all the variables as

$$P(X_1 = x_1, \ldots, X_d = x_d) = \prod_{i=1}^{d} P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i}) \tag{1}$$

where, for some nodes, $pa_i = \{\}$ (the empty set) and $P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i})$ reduces to $P(X_i = x_i)$ (no other variable need to be consulted prior to sampling a value for X_i). You should convince yourself that any directed acyclic graph must have at least one node without any parents.

The above factorization resembles an application of the chain rule. First, we write the variables in some order such that the parents of each variable always come before the variable itself in the ordering. This is always possible for a directed acyclic graph (in fact, there are often a large number of such "consistent" orderings). Then we simply "drop" dependences in the conditional probabilities to get the above factorization. These simplifications represent independence assumptions about the variables. To simplify the notation, let's assume that the simple lexicographic ordering

works for our graph. In other words, assume that $pa_i \subset \{1, \ldots, i-1\}, i = 1, \ldots, d$ Then, by applying the chain rule, we could always write (without any assumptions)

$$P(X_1 = x_1, \ldots, X_d = x_d) = \prod_{i=1}^{d} P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) \tag{2}$$

When we construct the distribution for a particular graph, we are assuming that

$$P(X_i = x_i | X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) = P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i}) \tag{3}$$

where $pa_i \subset \{1, \ldots, i-1\}$. This is an independence assumption. Specifically, define $npa_i = \{1, \ldots, i-1\} \backslash pa_i$ as the "preceding non-parents". By factoring the joint distribution according to Eq. (1), *i.e.*, dropping dependences except for the parents, we make the assumption that X_i is independent of $\mathbf{X}_{npa_i}$ given values for the parents $\mathbf{X}_{pa_i}$. These independence statements, one statement per variable, would suffice to specify the directed graph (dropping all arcs from preceding non-parents). But there are many other independence statements that are implied by these. How do we read all of these off the graph directly?

Independence from the graph: D-separation

As an example, consider a slightly extended version of the alarm model we had discussed before. The model is given in Figure 1, with an additional binary variable L. This could be whether we "leave work" as a result of hearing/learning about the alarm. We will now define a procedure for answering questions such as: are R and B independent of each other? Are R and B independent of each other if we know (given) A? And so on.

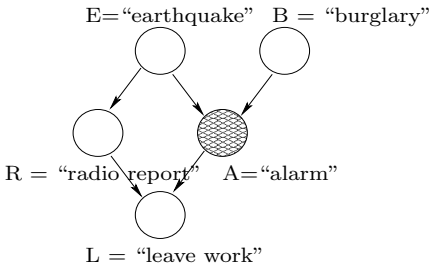
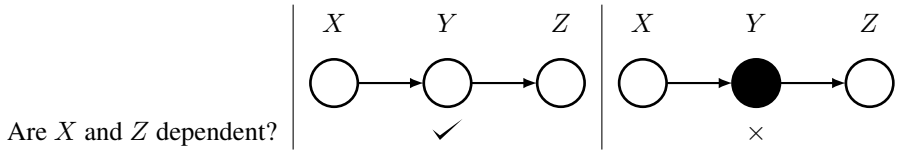


Figure 1: An example Bayesian network.

Let us first look at some simpler cases. We consider three variables X, Y and Z, and we are interested in finding the dependence/independence information between X and Z under different conditions (whether Y is given/observed or not). Three basic conditions are considered, where all three variables are connected to one another in some way.

Chain



From the graph, we have:

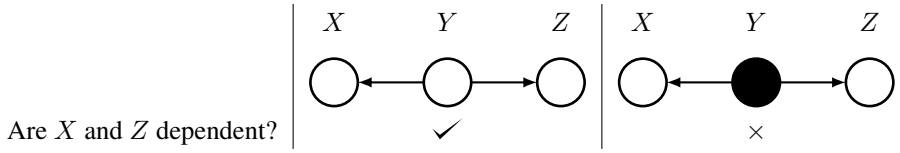
$$P(X,Y,Z) = P(X)P(Y|X)P(Z|Y) \quad (4)$$
$$P(X,Y) = P(X)P(Y|X) \quad (5)$$

Now let us look at the following conditional probability:

$$P(Z|X,Y) = \frac{P(X,Y,Z)}{P(X,Y)} = \frac{P(X)P(Y|X)P(Z|Y)}{P(X)P(Y|X)} = P(Z|Y) \quad (6)$$

This means Z and X are *independent* given Y. On the other hand, without knowing Y, the variables Z and X are *dependent*.

Common Cause



The joint probability for these three variables is defined as:

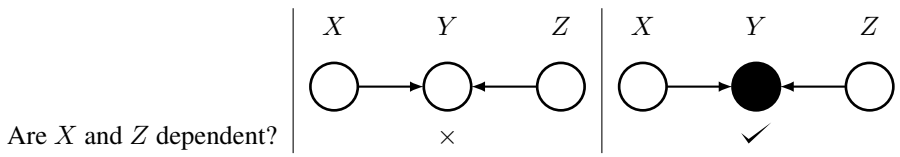
$$P(X,Y,Z) = P(Y)P(X|Y)P(Z|Y) \quad (7)$$
$$P(X,Y) = P(Y)P(X|Y) \quad (8)$$

Now we look at the following conditional probability:

$$P(Z|X,Y) = \frac{P(X,Y,Z)}{P(X,Y)} = \frac{P(Y)P(X|Y)P(Z|Y)}{P(Y)P(X|Y)} = P(Z|Y) \quad (9)$$

This means Z and X are *independent* given Y. On the other hand, without knowing Y, the variables Z and X are *dependent*.

Explaining Away



The joint probability for these three variables is defined as:

$$P(X,Y,Z) = P(X)P(Z)P(Y|X,Z) \quad (10)$$

Discussions Can we prove that $P(Z|X,Y) = P(Z|Y)$?

Okay, now let us instead look at the following *joint* probability:

$$P(X,Z) = \sum_Y P(X)P(Z)P(Y|X,Z) = P(X)P(Z)\sum_Y P(Y|X,Z) = P(X)P(Z) \quad (11)$$

This means Z and X are *independent* without knowing Y. On the other hand, Z and X are *dependent* given Y, as we have seen in the previous class when we discussed explaining away.

The notation of *d-separation* (or *directed-separation*) refers to the connectivities of two distinct sets of variables (given another set of variables) in a Bayesian network. We have so far seen the above basic cases, and next let us discuss a more general algorithm for determining *d-separation* between variables for a general Bayesian network.

Bayes' Ball Algorithm

The above independence/dependence information can be read off from the graphs directly when there are three variables. It can be shown that such independence/dependence information can be read off from a Bayesian network that involves more variables. The resulting algorithm is called the *Bayes' ball algorithm*, and it relies on the above simpler facts as building blocks.

We would like to find whether $\mathbf{x}_A$ and $\mathbf{x}_B$ are independent given $\mathbf{x}_C$, where $\mathbf{x}_A, \mathbf{x}_B$ and $\mathbf{x}_C$ are disjoint sets of random variables in a given Bayesian network.

We need to find for each variable in $\mathbf{x}_A$ and each variable in $\mathbf{x}_B$ whether they are all conditionally independent given $\mathbf{x}_C$. If so, then we conclude that $\mathbf{x}_A$ and $\mathbf{x}_B$ are conditionally independent given $\mathbf{x}_C$.

To check whether two variables X and Z are conditionally independent, we can use depth-first search to find all the paths from X to Z. Each path consists of the above-mentioned basic patterns. Imaging there is a ball (Bayes' ball) that is rolling on the path. We check if any of the path is open (*i.e.*, the path consists of patterns with a $\checkmark$ symbol below) so that the ball can reach Z from X by following that path. If there exists a path that is open, we say the two variables X and Z are *dependent*. Otherwise they are *independent*.

Markov Blanket

Assume we are interested in finding the following conditional probability:

$$p(X|\mathbf{V}_{-X}) \quad (12)$$

where $\mathbf{V}_{-X}$ is the set of variables in the Bayesian network, excluding the variable X.

Using what we have learned today, we can see that conditioning on the set of all other variables except X is equivalent to conditioning on a few variables surrounding the variable X only. This set of nodes is called the Markov blanket of the variable X.

Discussions What are the nodes that the Markov blanket consists of?

Answer: the Markov blanket of the variable X consists of X's _____, X's _____ as well as the parents of X's _____.

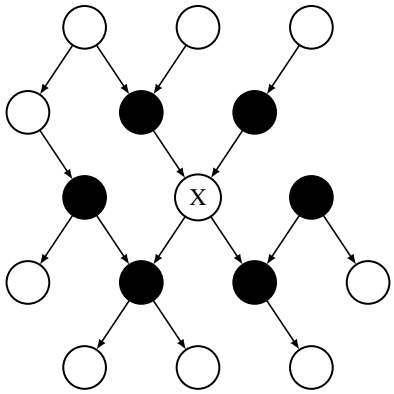


Figure 2: The conditional probability of X given all other variables is equivalent to $P(X|\mathbf{m}(X))$, where $\mathbf{m}(X)$ is the Markov blanket of X.

This observation allows us to derive the conditional probabilities easily. In fact this is a crucial observation that is useful for the Gibbs sampling algorithm, an important *approximate inference* algorithm used for inference in general Bayesian networks.

(**Optional**) The Gibbs sampling algorithm is used for generating samples $\mathbf{y} = \langle y_1, y_2, \ldots, y_n \rangle$ from a distribution given some evidence $P(\mathbf{y}|\mathbf{x})$, where $\mathbf{x}$ and $\mathbf{y}$ are sets of variables. The procedure is as follows:

1. Randomly initialize $\mathbf{y}^{(0)} = \left\langle y_1^{(0)}, y_2^{(0)}, \ldots, y_n^{(0)} \right\rangle$.
2. For $t = 1, \ldots, T$ do
   (a) For $k = 1, \ldots, n$ do
   $y_k^{(t)} \sim P(y_k|y_1^{(t)}, \ldots, y_{k-1}^{(t)}, y_{k+1}^{(t-1)}, y_{k+2}^{(t-1)}, \ldots, y_n^{(t-1)}, \mathbf{x})$ (This conditional probability can be simplified using Markov blanket.)
   (b) Collect the t-th sample as $\left\langle y_1^{(t)}, y_2^{(t)}, \ldots, y_n^{(t)} \right\rangle$
3. Return the collection of samples.

This algorithm can be used to perform approximate inference in Bayesian networks.

Learning Model Parameters

Suppose now that we have d discrete variables, $X_1, \ldots, X_d$ where $X_i \in \{1, \ldots, r_i\}$ and m *complete* observations $D = \{(x_1^{(t)}, \ldots, x_d^{(t)}), t = 1, \ldots, m\}$. In other words, each observation contains a value assignment to all the variables in our model. This is a simplification and models in practice (*e.g.*, mixture models, HMMs) are intended to be estimated from incomplete data. We will also assume that the conditional probabilities we need to specify for a Bayesian network are fully parameterized. This means, *e.g.*, that in $P(X_1 = x_1|X_2 = x_2)$ we can select the probability distribution over X_1 separately (without additional constraints) for each possible value of the parent x_2.

Models used in practice often have parametric constraints so that, for example, "similar" values of X_2 would lead to "similar" distributions over X_1. In our case here, the model interprets the value of each variable just as a symbol without any such constraints.

Given an acyclic graph G over d variables, we already know that we can write down the associated joint distribution as

$$P(X_1 = x_1, \ldots, X_d = x_d) = \prod_{i=1}^{d} P(X_i = x_i | \mathbf{X}_{pa_i} = \mathbf{x}_{pa_i}) = \prod_{i=1}^{d} \theta_i(x_i | \mathbf{x}_{pa_i}) \tag{13}$$

where $\theta_i(x_i|\mathbf{x}_{pa_i})$ are the probability tables that we must estimate. For example, if $X_3 \in \{1, \ldots, r_3\}$ has two parents, X_1 and X_2, each taking values in $\{1, \ldots, r_1\}$ and $\{1, \ldots, r_2\}$, respectively, then the table $\theta_3(x_3|x_1, x_2)$ is given by

$X_3|X_1, X_2$:

X_1, X_2	1,	...,	r_3
1,1	$\theta_3(1\|1,1)$,	...	$\theta_3(r_3\|1,1)$
1,2	$\theta_3(1\|1,2)$,	...	$\theta_3(r_3\|1,2)$
...	...,	...	...
$r_1, 1$	$\theta_3(1\|r_1,1)$,	...	$\theta_3(r_3\|r_1,1)$
$r_1, 2$	$\theta_3(1\|r_1,2)$,	...	$\theta_3(r_3\|r_1,2)$
...	...,	...	...
r_1, r_2	$\theta_3(1\|r_1,r_2)$,	...	$\theta_3(r_3\|r_1,r_2)$

(14)

Discussions How many parameters are in the table? What is the *effective* number of parameters (or the number of *free/independent* parameters)?

We can now write down the log-likelihood of observed data $D = \{(x_1^{(t)}, \ldots, x_d^{(t)}), t = 1, \ldots, m\}$ for any particular Bayesian network structure, *i.e.*, for any particular graph G. It is given by

$$l(D; \theta; G) = \sum_{t=1}^{m} \log \left[\prod_{i=1}^{d} \theta_i(x_i^{(t)} | \mathbf{x}_{pa_i}^{(t)}) \right] = \sum_{t=1}^{m} \sum_{i=1}^{d} \log \theta_i(x_i^{(t)} | \mathbf{x}_{pa_i}^{(t)}) = \sum_{i=1}^{d} \left[\sum_{t=1}^{m} \log \theta_i(x_i^{(t)} | \mathbf{x}_{pa_i}^{(t)}) \right] \tag{15}$$

where we grouped the terms by variable (given their parents) in order to highlight the fact that the associated parameters can be set separately from those pertaining to other variables. The *maximum likelihood estimation* of the model parameters then reduces to the problem of estimating individual tables such as the one in the table above (*i.e.*, (14)).

The table $\theta_i(x_i|\mathbf{x}_{pa_i})$ specifies a multinomial distribution over x_i for each setting of the parent variables $\mathbf{x}_{pa_i}$. As a result, we can maximize the likelihood analogously to estimating multinomial parameters we have seen before. Indeed,

$$\sum_{t=1}^{m} \log \theta_i(x_i^{(t)} | \mathbf{x}_{pa_i}^{(t)}) = \sum_{x_i, \mathbf{x}_{pa_i}} \text{Count}\Big((x_i, \mathbf{x}_{pa_i}) \text{ in } D \Big) \log \theta_i(x_i | \mathbf{x}_{pa_i}) \tag{16}$$

where $\text{Count}\Big((x_i, \mathbf{x}_{pa_i}) \text{ in } D \Big)$ gives the number of observations in data D for which $X_i = x_i$ and $\mathbf{X}_{pa_i} = \mathbf{x}_{pa_i}$ So, if we fix $\mathbf{x}_{pa_i}$, then $\text{Count}\Big((\cdot, \mathbf{x}_{pa_i}) \text{ in } D \Big)$ specifies the counts for a multinomial

$\theta_i(\cdot|\mathbf{x}_{pa_i})$. The corresponding maximum likelihood parameter estimate is simply (analogously to a single multinomial)

$$\hat{\theta}_i(x_i|\mathbf{x}_{pa_i}) = \frac{\text{Count}\Big((x_i, \mathbf{x}_{pa_i}) \text{ in } D\Big)}{\text{Count}\Big((\mathbf{x}_{pa_i}) \text{ in } D\Big)}, x_i \in \{1, \ldots, r_1\} \tag{17}$$

where $\text{Count}\Big((\mathbf{x}_{pa_i}) \text{ in } D\Big)$ is the number of times we see the pattern $\mathbf{x}_{pa_i}$ in D:

$$\text{Count}\Big((\mathbf{x}_{pa_i}) \text{ in } D\Big) = \sum_{x_i'} \text{Count}\Big((x_i', \mathbf{x}_{pa_i}) \text{ in } D\Big)$$

Repeating the procedure for each setting of $\mathbf{x}_{pa_i}$, and for different variables, yields the maximum likelihood parameter estimates $\hat{\theta}_i(x_i|\mathbf{x}_{pa_i}), i = 1, \ldots, d$.

Learning Objectives

You need to know:

1. How does the Bayesian network capture dependence information between different variables using a DAG representation.
2. How to read off the dependence and independence information between different variables given evidences from the DAG directly by using the Bayes' ball algorithm.
3. What is the Markov blanket of a variable in a Bayesian network.
4. How to count the effective number of parameters for a given Bayesian network
5. How to learn the parameters of Bayesian networks of a certain structure from complete data based on a collection of samples