SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

01.112 Machine Learning, Spring 2018
Lecture Notes for Week 11

18. Bayesian Networks (I)

Last update: Friday 30th March, 2018 11:49

Overview

Bayesian networks are generative probabilistic models that were developed for representing and using probabilistic information. All generative models involve variables. For example, the choice of mixture component is a variable, the state in an HMM is a variable, and so on. How we select values for these variables is governed by a probability distribution. For example, mixture models specify a probability distribution over the selection of the mixture component as well as the (Gaussian) output variable. HMMs specify a distribution over the sequence of hidden states as well as the corresponding observation symbols. As generative models, Bayesian networks subsume mixture models, hidden Markov models, and many others. In fact, Bayesian networks provide a simple language for specifying generative probability models.

There are two parts to any Bayesian network model: 1) a directed graph over the variables and 2) the associated probability distribution. The graph represents qualitative information about the random variables (conditional independence properties), while the associated probability distribution, consistent with such properties, provides a quantitative description of how the variables relate to each other. If we already have the distribution, as we have for mixture models or HMMs, why do we need the graph? The graph structure serves two important functions. First, it explicates the properties about the underlying distribution that would be otherwise hard to extract from a given distribution. For example, it tells us whether two sets of variables are independent of each other, and in which scenarios (known values for some variables). The graph is a compact summary of such statements. Given that the graph constrains the distribution, it explicates how we can generate data. As a result, the graph structure can be learned from available data, *i.e.*, we can explicitly learn qualitative properties from data. Second, since the graph pertains to independence properties about the random variables, it is very useful for understanding how we can use the probability models efficiently to evaluate various marginal and conditional properties. This is exactly why we were able to carry out efficient computations in HMMs. The forward-backward algorithms relied on simple Markov properties which are independence properties, and these are generalized in Bayesian networks. We can make use of independence properties whenever they are explicit in the model.

Impt.

Bayesian networks: examples, properties

Let us start with a simple example model over three binary variables. We imagine that two people are flipping coins independently from each other. The resulting values of their unbiased coin flips are stored in binary (H/T) variables X_1 and X_2. Another person checks whether the coin flips resulted in the same value and the outcome of the comparison is a binary (T/F) variable $X_3 = [\![X_1 = X_2]\!]$ (logical true/false). We will first construct the distribution, then look at how we should represent it as a graph.

The two coin flips are governed by simple uniform probability distributions. For example, $P(X_1 = H) = 0.5$ and $P(X_1 = T) = 0.5$. We can represent these probabilities as tables

$$X_1 : \begin{array}{|cc|}\hline H & T \\ \hline 0.5 & 0.5 \\ \hline \end{array}, \; X_2 : \begin{array}{|cc|}\hline H & T \\ \hline 0.5 & 0.5 \\ \hline \end{array} \tag{1}$$

where each row in the table must sum to one. The value of X_3, on the other hand, depends on (in fact, is a function of) X_1 and X_2 and cannot be determined until we know which values X_1 and X_2 take. We must therefore specify a conditional distribution $P(X_3 = x_3 | X_1 = x_1, X_2 = x_2)$ for this variable. The conditional probability can also be represented as a table where we introduce a row for each possible setting of X_1 and X_2.

$$X_3|X_1, X_2 : \begin{array}{cc|cc}\hline X_1 & X_2 & T & F \\ \hline H & H & 1 & 0 \\ H & T & 0 & 1 \\ T & H & 0 & 1 \\ T & T & 1 & 0 \\ \hline \end{array} \tag{2}$$

Again, each row of the probability table sums to one. Note that the probability values are extreme valued (zero and one) because X_3 is a function of X_1 and X_2. So, for example, $P(X_3 = T | X_1 = H, X_2 = H) = 1$ while $P(X_3 = F | X_1 = H, X_2 = H) = 0$, the first row in the above table. Now, since the two coins are flipped independently of each other, we can write the joint distribution over the three variables as

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = P(X_1 = x_1) P(X_2 = x_2) P(X_3 = x_3 | X_1 = x_1, X_2 = x_2) \tag{3}$$

In order to represent this as a Bayesian network, we will use a directed graph over the variables X_1, X_2, and X_3 in addition to the distribution. The nodes in the graph represent variables while the directed edges specify dependences, *i.e.*, whether one variable directly *depends on* another. We know in this example that X_1 and X_2 do not directly depend on each other, while X_3 depends on both X_1 and X_2. As a result, the directed graph for this model is as given by Figure 1.

Typically, we would write down the distribution in response to the graph rather than the other way around. In fact, how the distribution factors is determined directly by the graph. We need a bit of terminology for this. In the graph, X_1 is a parent of X_3 since there's a directed edge from X_1 to

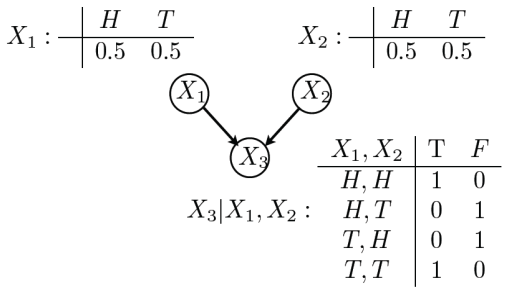


Figure 1: A directed graph for the coin toss example with the associated conditional probability tables

X_3 (the value of X_3 depends on X_1). Analogously, we can say that X_3 is a *child* of X_1. Now, X_2 is also a parent of X_3 so that the value of X_3 depends on both X_1 and X_2. We will discuss later what the graph means more formally (it captures independence properties). For now, we just note that Bayesian networks always define acyclic graphs (no directed cycles) and represent how values of the variables depend on their parents, *i.e.*, how we can generate values for the variables. Any joint distribution consistent with the graph, *i.e.*, any distribution we could imagine associating with the graph, has to be able to be written as a product of conditional probabilities of each variable given its parents. If a variable has no parents (as is the case with X_1) then we just write $P(X_1 = x_1)$. Eq. (3) is exactly a product of conditional probabilities of variables given their parents.

Marginal independence and induced dependence

Let us analyze the properties of the simple model a bit. For example, what is the marginal probability over X_1 and X_2? This is obtained from the joint simply by summing over the values of X_3:

$$P(X_1 = x_1, X_2 = x_2) = \sum_{x_3} P(X_1 = x_1)P(X_2 = x_2)P(X_3 = x_3|X_1 = x_1, X_2 = x_2) \quad (4)$$

$$= P(X_1 = x_1)P(X_2 = x_2) \sum_{x_3} P(X_3 = x_3|X_1 = x_1, X_2 = x_2) \quad (5)$$

$$= P(X_1 = x_1)P(X_2 = x_2) \quad (6)$$

Thus X_1 and X_2 are *marginally independent* of each other (a product distribution means that the variables are independent). In other words, if we don't know the value of X_3 then there's nothing that ties the coin flips together (they were, after all, flipped independently in the description). This is also a property we could have extracted directly from the graph. We will provide shortly a formal way of deriving this type of independence properties from the graph.

Another typical property of probabilistic models is *induced dependence*. Suppose now that the coins X_1 and X_2 were flipped independently but we don't know their outcomes. All we know that $X_3 = T$, *i.e.*, that the outcomes where identical. What do we know about X_1 and X_2 in this case? We know that either $X_1 = X_2 = H$ or $X_1 = X_2 = T$. So their values are clearly *dependent*. The

dependence was *induced by additional knowledge*, in this case observing the value of X_3. This is again a property we could have read off directly from the graph (explained below). Both marginal independence and induced dependence are typical properties of realistic models.

Explaining away

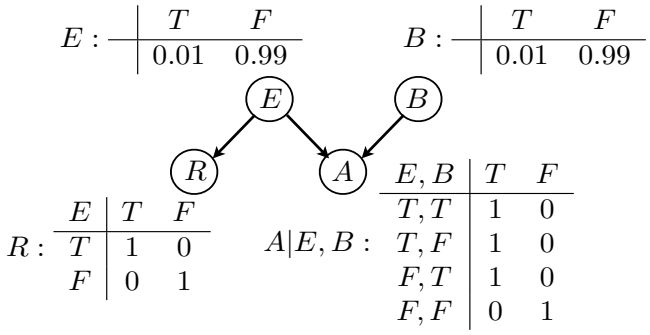


Figure 2: Alarm example with four variables, E, B, R, and A representing true/false values of earthquake, burglary, radio report, and alarm, respectively. The corresponding probability tables are given next to the variables.

Another typical phenomenon that probabilistic models can capture is explaining away. Consider the following typical example (Pearl 1988) in Figure 2. We have four variables A, B, E, and R capturing possible causes for why a burglary alarm went off. All the variables are binary (T/F) and, for example, $A = T$ means that the alarm went off. In our example here all the observed values are T (property is true). In general, observations in the graph would be represented by shaded nodes. We assume that earthquakes ($E = T$) and burglaries ($B = T$) are equally unlikely events $P(E = 1) = P(B = 1) = 0.01$. Alarm is likely to go off only if either $E = 1 or B = 1$ or both. Moreover, either event will trigger the alarm so that $P(A = T|E, B) = 1$ whenever either $E = T$ or $B = T$ or $E = B = T$, and $P(A = T|E, B) = 0$ when $E = B = F$. An earthquake ($E = T$) is likely to be followed by a radio report ($R = T$) where $P(R = T|E = T) = 1$, and we assume that the report never occurs unless an earthquake actually took place: $P(R = T|E = F) = 0$. Based on the graph, or based on how we constructed the distribution, we can write down the joint distribution over all the binary variables as

$$P(E = e, B = b, A = a, R = r) = P(E = e)P(B = b)P(A = a|E = e, B = b)P(R = r|E = e) \quad (7)$$

Note that it again factors as a product of "variable given its parents".

What do we believe about the values of the variables if we only observe that the alarm went off ($A = T$)? At least one of the potential causes $E = T$ or $B = T$ should have occurred. However, since both are unlikely to occur by themselves, we are basically left with either $E = T$ or $B = T$ but (most likely) not both. We therefore have two alternative or competing explanations for the observation and both explanations are equally likely. We can evaluate the posterior probability that

there was a burglary $P(B = T|A = T)$ as follows. Let us first evaluate the marginal probability over the variables we are interested in:

$$
\begin{aligned}
& P(B = b, A = T) && (8)\\
= & \sum_{e\in\{T,F\}} \sum_{r\in\{T,F\}} P(E = e)P(B = b)P(A = T|E = e, B = b)P(R = r|E = e) && (9)\\
= & \sum_{e\in\{T,F\}} P(E = e)P(B = b)P(A = T|E = e, B = b) \sum_{r\in\{T,F\}} P(R = r|E = e) && (10)\\
= & \sum_{e\in\{T,F\}} P(E = e)P(B = b)P(A = T|E = e, B = b) && (11)\\
= & \; P(B = b) \sum_{e\in\{T,F\}} P(E = e)P(A = T|E = e, B = b) && (12)
\end{aligned}
$$

Note how the radio report (R) dropped out since it is a variable downstream from E, and we did not observe its value. It represents an observation we could have made but didn't. Such "imagined" possibilities will not affect our calculations. Now,

$$P(B = T|A = T) = \frac{P(B = T, A = T)}{\sum_{b\in\{T,F\}} P(B = b, A = T)} \tag{13}$$

and evaluates just slightly above 0.5. Why not exactly 0.5? Because there's a slight chance that both $B = T$ and $E = T$, not just one or the other.

If we now hear, in addition, that there was a radio report about an earthquake, we believe that $E = T$ because $R = T$ only if $E = T$. As a result, $E = T$ perfectly explains the alarm $A = T$, removing any evidence about $B = T$. In other words, the additional observation about the radio report explained away the evidence for $B = T$. Thus, $P(B = T|A = T, R = T) = P(B = T) = 0.01$ (prior probability) whereas $P(E = T|A = T, R = T) = 1$.

Note that we have implicitly captured in our calculations here that R and B are *dependent* given $A = T$ (induced dependence). If they were not, we would not be able to learn anything about the value of B as a result of also observing $R = T$. Here the effect is drastic and the variables are strongly dependent. We could have, again, deduced this dependence from the graph directly. In the next lecture, we will look at independence a bit more formally.

Learning Objectives

You need to know:

1. How to parameterize a Bayesian network
2. How to compute marginalised conditional probabilities for simple Bayesian networks
3. What is "explaining away"

Recap of W10 02 (HMM)

- $P(x_1, \dots, x_n, y_0 = \text{START}, y_1, \dots, y_n, y_{n+1} = \text{STOP}) = \prod_{i=0}^{n} a_{y_i, y_{i+1}} \prod_{i=1}^{n} b_{y_i}(x_i)$

Supervised Learning : Given (X, Y). Find : $\theta = (a, b)$ - Estimate due to maximum likelihood.

MLE → log → ∂ → ~~path~~

→ $a_{u,v} = \frac{\text{count}(u,v)}{\text{count}(u)}$

$b_u(o) = \frac{\text{count}(u \to o)}{\text{count}(u)}$

Decode / Test / Ev : Given $\theta = (a, b)$, X. Find: Y via Viterbi Algorithm → path

① E-Step (Find membership)

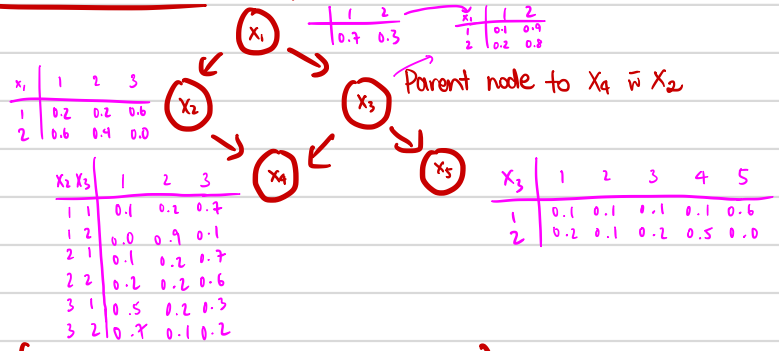
② M-Step (Re-estimate parameters) via Hard or Soft EM.

Inference : Given $\theta = (a, b)$. $P(y_j = u \mid x_1, \dots, x_n)$ ← inference · $x_1, \dots, x_n$: Evidence, Infer y_j when $= u$

$P(y_j = u, y_{j+1} = v \mid x_1, \dots, x_n)$

W11 01 (Bayesian Networks)

require picture & value of parameters → Describe WB Generative Models.



Properties

- Directed Graph.
- No loops / cycle in Generative Model.
  - ↳ Nodes generated from parent nodes / not future nodes.

→ Directed Acyclic Graph.

$P(x_1, x_2, x_3, x_4, x_5)$

$= P(x_5 \mid x_3) \cdot P(x_4 \mid x_2, x_3) \cdot P(x_3 \mid x_1) P(x_2, x_1) \cdot P(x_1)$

where

$P(x_1, \dots, x_m) = \prod_{j=1}^{m} P(x_j \mid Pa(x_i))$ where $Pa(x_j)$ represents the set of parents of x_j.

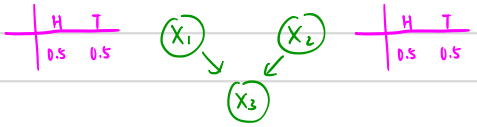
Eg : Iching

6 times.

opposes

Two head / tails

↓ Just 1 line



$X_1 X_2$	1	2 times
HH	1	0
HT	0	1
TH	0	1
TT	1	0

2 Independent Coins

X_1 and X_2 Independent

$P(X_1, X_2, X_3) = P(X_3 \mid X_1, X_2) \cdot P(X_1) \cdot P(X_2)$

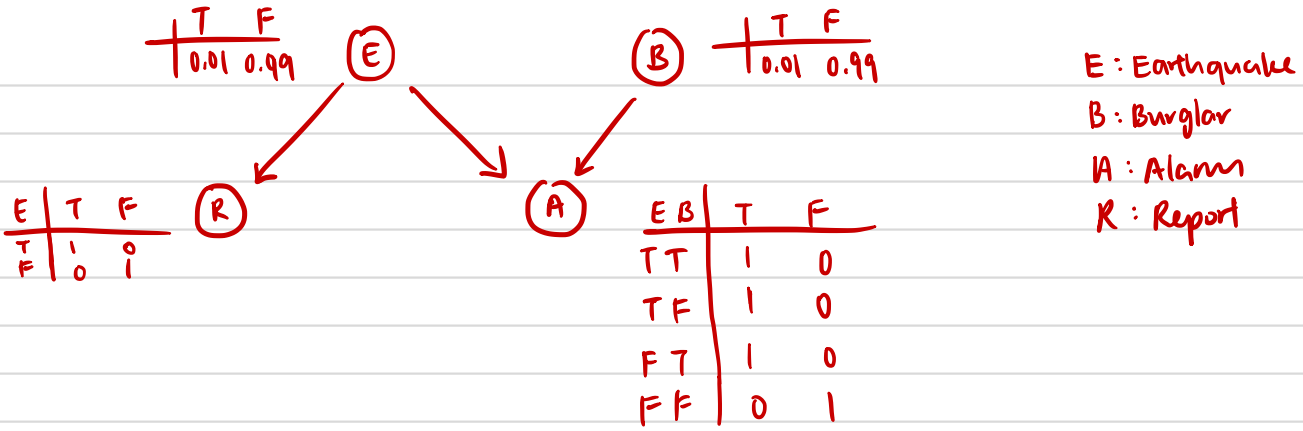
$P(X_1, X_2) = P(X_1) \cdot P(X_2)$

$P(X_1, X_2) = \sum_{X_3} P(X_1, X_2, X_3) = \sum_{X_3} P(X_1) P(X_2) P(X_3 \mid X_1, X_2)$

$= P(X_1) P(X_2) \sum_{X_3} (X_3 \mid X_1, X_2)$

$= P(X_1) P(X_2)$ where $\sum_Y (Y \mid X) = 1$ and $\sum_Y P(X, Y) = P(X)$

To know whether X_1 & X_2 are independent given X_3, first solve



$$P(E=e, B=b, R=r, A=a)$$
$$= P(E=e) \cdot P(B=b) \cdot P(R=r \mid E=e) \cdot P(A=a \mid E=e, B=b)$$

Qn. $P(B=T \mid A=T)$

$$= \frac{P(B=T, A=T)}{P(A=T)}$$

$$= \frac{P(B=T, A=T)}{\sum_b P(B=b, A=T)}$$

$$= \frac{P(B=T, A=T)}{[P(B=T)\sum_e P(E=e)\,P(A=T \mid E=e, B=T) + P(B=F)\sum_e P(E=e)\,P(A=T \mid E=e, B=F)]}$$

$$= \frac{0.01 \times \{0.01 \times 1 + 0.99 \times 1\}}{0.01 \times \{0.01 \times 1 + 0.99 \times 1\}} \qquad 9\{0.01 \times 1 + 0.99 \times 0\}$$

$$= \frac{0.01 \times 1}{0.01 \times (1.99)}$$

$$= \frac{0.01}{0.0199} > 0.5$$ (bcos 11 chance both occur)

$$P(B=b, A=T) = \sum_e \sum_r P(E=e, B=b, R=r, A=T)$$
$$= \sum_e \sum_r P(E=e) P(B=b) \cdot P(R=r \mid E=e) \cdot P(A=T \mid E=e, B=b)$$
$$= P(B=b) \sum_e P(E=e)\, P(A=T \mid E=e, B=b) \sum_r P(R=r \mid E=e)$$
$$= P(B=b) \sum_e P(E=e)\, P(A=T \mid E=e, B=b) \quad (1)$$

$P(B=T) \sum_e P(E=e)\, P(A=T \mid E=e, B=T)$

$$= 0.01 \times \{0.01 \times 1 + 0.99 \times 1\} \quad (1)$$

$P(B=F) \sum_e P(E=e)\, P(A=T \mid E=e, B=F)$ (1)

$$= 0.99 \times \{0.01 \times 1 + 0.99 \times 0\}$$

$$P(B=T \mid A=T, R=T)$$
$$= P(B=T \mid A=T, E=T)$$
$$= \frac{P(B=T, A=T, E=T)}{P(A=T, E=T)} \longrightarrow \text{Since } P(B=T, A=T) = 0.01$$
$$= \frac{0.01 \times 0.01}{0.01}$$
$$= 0.01$$

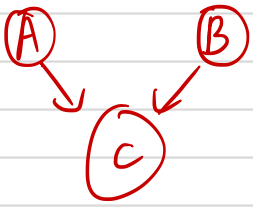
Since $P(B=T \mid A=T) \neq P(B=T \mid A=T, E=T)$

If they are not the same, they are not independent
↳ they are dependent.

Explaining away: Researcher interested in correlation between 2 diseases.
If history of A ⇒ getting B.

Even though A and B may not be dependent but mtd explains high occurrence of B given A.

Can explain Perkson's Paradox usg Explaing Away Mtd



but $A \not\Rightarrow B$

↳ explain independence / dependence