

Lecture 3

Cezar Ionescu

Bayesian networks

Joint probability distributions

The mathematical model of probability consists of a triple $(\Omega, \text{Event}, p)$ as in the previous lecture. Very frequently, however, the set of possible outcomes Ω has the structure of a *cartesian product*, i.e., a tuple of possible results. For instance, the diagnostic problem we discussed last time could also have been modelled by taking $\Omega = (H, R)$, i.e., the set of tuples $\{(\text{healthy}, \ominus), (\text{healthy}, \oplus), (\text{ill}, \ominus), (\text{ill}, \oplus)\}$. The event “healthy person” would then be described by $\text{Healthy} = \{(\text{healthy}, \ominus), (\text{healthy}, \oplus)\}$, etc.

Exercise: Define the other events of interest: “ill”, “tested positive”, “tested negative”.

We can now construct new probability spaces on each component of Ω . For example:

$$\begin{aligned} p_h : \mathbb{P}(H) &\rightarrow [0, 1] \\ p_h(\text{healthy}) &= p(\{(\text{healthy}, \ominus), (\text{healthy}, \oplus)\}) \\ &= p(\text{healthy}, \ominus) + p(\text{healthy}, \oplus) \end{aligned}$$

Remark: in general, if $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_n)$ and $\omega_i \in \Omega_i$, we will use $p(\omega_i)$ to denote

$$p(\omega_i) = p(\{(x_1, x_2, \dots, x_n) \mid x_1 \in \Omega_1, \dots, x_n \in \Omega_n, x_i = \omega_i\})$$

Remark: recall our convention that for $x \in \Omega$ $p(x)$ is an abbreviation for $p(\{x\})$. Above, we have taken this convention one step further, by using $p(x, y)$ to denote $p(\{(x, y)\})$, and $p(x)$ to denote $p(\{(x, y) \mid y \in Y\})$. We shall continue using this form of abbreviation throughout, as is standard usage.

The “complete” probability measure p is called the **joint probability distribution**, the probability measures over the individual components (or over tuples of individual components) are called **marginal probability distributions**.

Remark: it is trivial to derive the marginal probability distributions, given the joint probability distribution. However, even if we have all the marginal distributions, we cannot derive the joint distribution from them.

By definition, the joint distribution is sufficient to answer any query about the probabilities of events in Ω . The problem is that it requires exponential storage space with respect to the

number of components. For example, if Ω consists of a tuple of n boolean components, then the joint probability distribution can be represented as a table with 2^n rows.

Let $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_n)$ and $(\omega_1, \omega_2, \dots, \omega_n) \in \Omega$. Then

$$\begin{aligned} p(\omega_1, \omega_2, \dots, \omega_n) &= p(\omega_1 \wedge \omega_2 \wedge \dots \wedge \omega_n) && \text{-- recall convention above!} \\ &= p(\omega_1 \mid \omega_2 \wedge \dots \wedge \omega_n) * p(\omega_2 \wedge \dots \wedge \omega_n) && \text{-- Bayes' theorem} \end{aligned}$$

It is easy to see that, therefore:

$$\begin{aligned} p(\omega_1, \omega_2, \dots, \omega_n) &= p(\omega_1 \mid \omega_2 \wedge \dots \wedge \omega_n) * \\ &\quad p(\omega_2 \mid \omega_3 \wedge \dots \wedge \omega_n) * \dots \\ &\quad p(\omega_n) \end{aligned}$$

Independence, conditional independence

Definition: Two events $X, Y \subseteq \Omega$ are called **independent** if

$$p(X \wedge Y) = p(X) * p(Y)$$

Example: Consider a fair standard die, and the events $\text{even} = \{2, 4, 6\}$, $\text{big} = \{5, 6\}$. Then even and big are independent.

Exercise: If X, Y are independent, then so are $X, \neg Y$.

Example: The events big and $\text{divBy3} = \{3, 6\}$ are **not** independent.

Definition: Consider events $X, Y, Z \subseteq \Omega$ such that $p(Z) \neq 0$. Events X, Y are called **conditionally independent given Z** if

$$p(X \wedge Y \mid Z) = p(X \mid Z) * p(Y \mid Z)$$

Example:

- The events even, big are conditionally dependent given divBy3 . Intuition: knowing the result is divBy3 means that knowledge of it being big informs knowledge of it being even .
- The events $\text{divBy3}, \text{big}$ are conditionally independent given $\{2, 3, 5, 6\}$. Intuition: knowing that the result is in $\{2, 3, 5, 6\}$ means that knowledge of it being big doesn't tell me anything more about it being divBy3 .

Bayesian networks

A Bayesian network is a graphical device for recording conditional independence information for a finite $\Omega = (\Omega_1, \Omega_2, \dots, \Omega_n)$, in order to make the computation of entries of the joint probability distribution more efficient. The network is a *directed acyclic graph* (DAG), with one node for each component Ω_i . For example:

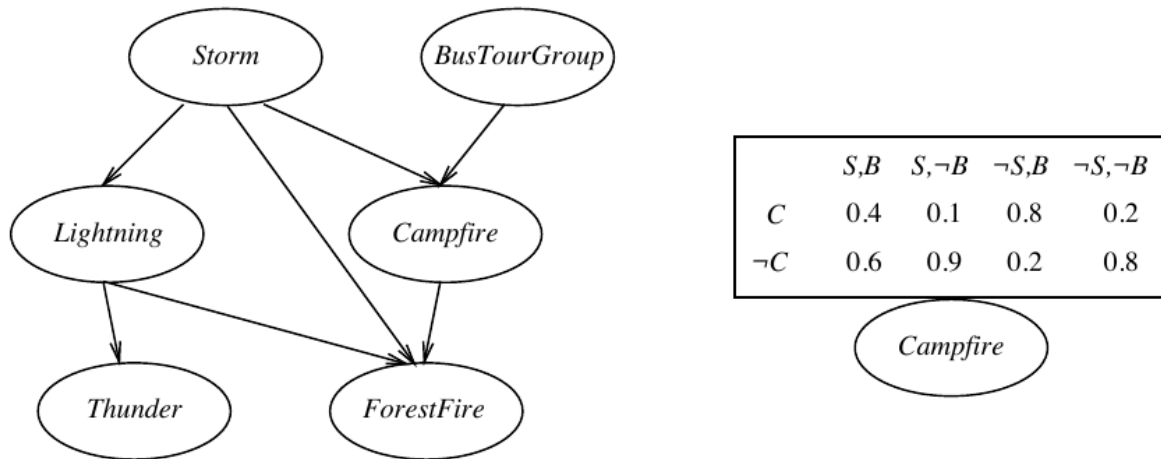


Figure 1: Mitchell, page 186

The network is assumed to be constructed in such a way that the probability of an ω_i is conditionally independent of any non-descendants given its parents.

Example: In the network above, we have

$$p(\text{thunder} \mid \text{lightning} \wedge \text{bustourgroup}) = p(\text{thunder} \mid \text{lightning})$$

but in general

$$p(\text{lightning} \mid \text{thunder} \wedge \text{storm}) \neq p(\text{lightning} \mid \text{storm})$$

In this way, the Bayesian network allows us to compute any entry in the joint distribution table. For example:

$$\begin{aligned}
 & p(\text{thunder}, \text{forestfire}, \text{campfire}, \text{lightning}, \text{storm}, \text{bustourgroup}) \\
 = & \\
 & p(\text{thunder} \mid \text{forestfire} \wedge \text{campfire} \wedge \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) * \\
 & p(\text{forestfire} \wedge \text{campfire} \wedge \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) \\
 = & \\
 & p(\text{thunder} \mid \text{lightning}) * \\
 & p(\text{forestfire} \wedge \text{campfire} \wedge \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) \\
 = & \\
 & p(\text{thunder} \mid \text{lightning}) * \\
 & p(\text{forestfire} \mid \text{campfire} \wedge \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) * \\
 & p(\text{campfire} \wedge \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) \\
 = & \\
 & p(\text{thunder} \mid \text{lightning}) * \\
 & p(\text{forestfire} \mid \text{campfire} \wedge \text{lightning}) * \\
 & p(\text{campfire} \wedge \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) \\
 = & \\
 & p(\text{thunder} \mid \text{lightning}) *
 \end{aligned}$$

$$\begin{aligned}
& p(\text{forestfire} \mid \text{campfire} \wedge \text{lightning}) * \\
& p(\text{campfire} \mid \text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) * \\
& p(\text{lightning} \wedge \text{storm} \wedge \text{bustourgroup}) \\
= & \\
& p(\text{thunder} \mid \text{lightning}) * \\
& p(\text{forestfire} \mid \text{campfire} \wedge \text{lightning}) * \\
& p(\text{campfire} \mid \text{storm} \wedge \text{bustourgroup}) * \\
& p(\text{lightning} \mid \text{storm}) * \\
& p(\text{storm}) * p(\text{bustourgroup})
\end{aligned}$$

Assuming all the Ω_i are boolean, we need two tables 1x2 tables to store the probability values for **Storm**, **BusTourGroup**, two 2x2 tables to store the conditional table for **Lightning** | **Storm** and **Thunder** | **Lightning**, and two 4x2 tables to store the conditional table for **ForestFire** | **Campfire** \wedge **Lightning** and **Campfire** | **Storm** \wedge **BusTourGroup**, altogether 28 values. If we stored the full joint distribution table, we would need $2^6 = 64$ values.

Homework: (based on Nilsson, 19.4, page 340) Consider the following Bayesian network:

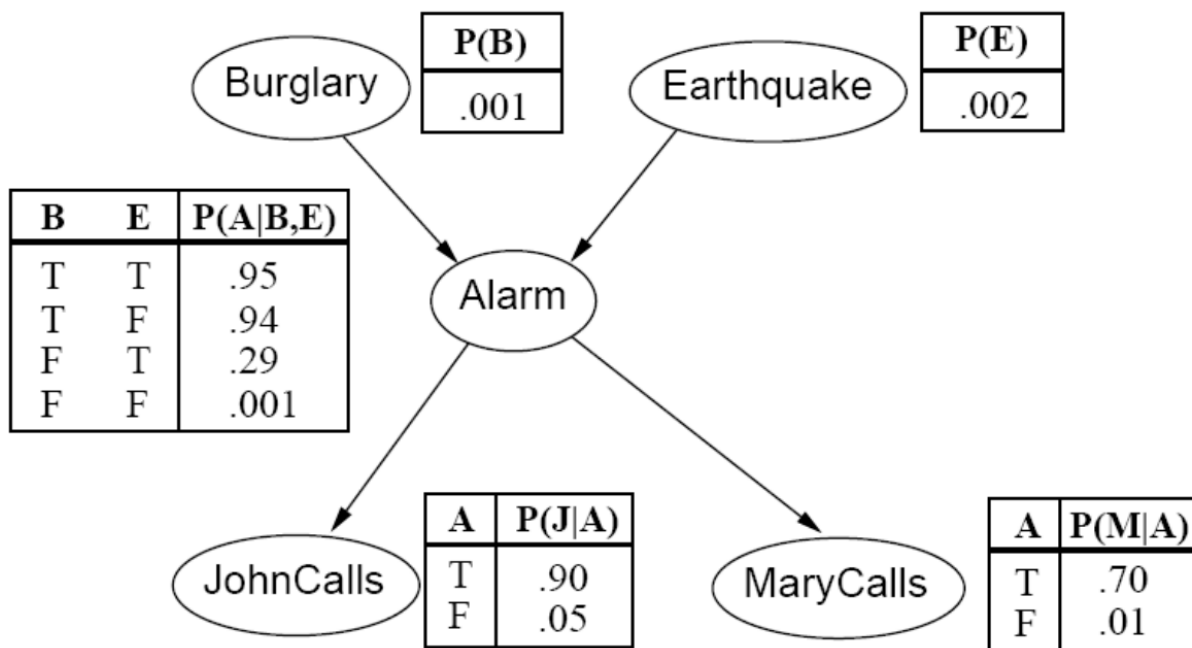


Figure 2: Nielsson, page 341

Compute $p(\neg J, \neg M, A, B, E)$. This is the probability that there is both an earthquake and a burglary, the alarm rings, but neither John nor Mary call.

Exercise: Compute $p(\neg J, \neg M, B, E)$ (this is exercise 19.4 in Nielsson).

Minimum description length (MDL) principle

$$h_{map} = \operatorname{argmax}_h p(d \mid h) * p(h)$$

iff
$$h_{map} = \operatorname{argmax}_h \log_2 (p(d \mid h) * p(h))$$

iff
$$h_{map} = \operatorname{argmax}_h \log_2 p(d \mid h) + \log_2 p(h)$$

iff
$$h_{map} = \operatorname{argmin}_h -\log_2 p(d \mid h) - \log_2 p(h)$$

iff
$$h_{map} = \operatorname{argmin}_h \text{length}(\text{encode}(d \mid h)) + \text{length}(\text{encode}(h))$$