1 Probability Theory for Bayesian Networks

Let us consider discrete random variables $x, y, z$ that can get discrete values $1, 2, \ldots, n$. We will write $P(x = 1)$ to mark the probability of the event that $x$ has the value 1. We will also write $P(x)$ to denote the probability distribution of $x$. The axioms of probability theory define that the probabilities are at least zero and the sum of them over the entire sample space is 1:

$$P(x = i) \geq 0 \forall i = 1, 2, \ldots, n \quad (1)$$

$$\sum_{i=1}^{n} P(x = i) = 1. \quad (2)$$

We will write $P(x, y)$ to denote the joint probability distribution of $x$ and $y$. For instance, $P(x = 1, y = 1)$ gives the probability of the event that both $x = 1$ and $y = 1$. When $P(x, y) = P(x)P(y)$, we say that $x$ and $y$ are independent, or $x \perp y$. Conditional probabilities are defined as follows:

$$P(x \mid y) = \frac{P(x, y)}{P(y)}. \quad (3)$$

This is read as “the probability of $x$ given $y$”. It means that assuming that we already know the value of $y$, what is the probability of $x$. Conditional probability is especially useful for causal thinking, that is, if $y$ might cause $x$, it is intuitive to think about $P(x \mid y)$.

From Eq. (3), we can derive two useful rules. First we notice we can divide joint probabilities into factors:

$$P(x, y, z) = P(x)\frac{P(x, y) P(x, y, z)}{P(x) P(x, y)} = P(x)P(y \mid x)P(z \mid x, y). \quad (4)$$

$$P(x, y, z) = P(x)P(y \mid x)P(z \mid x, y). \quad (5)$$
Then, we can prove the Bayes theorem by dividing $P(x, y)$ into factors both ways and by dividing both sides by $P(x)$:

$$P(x, y) = P(x \mid y)P(y) = P(y \mid x)P(x)$$  \hspace{1cm} (6)

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}.$$  \hspace{1cm} (7)

The Bayes theorem is useful to invert causality, that is, reasoning from the effects to the causes. For instance if $y$ causes $x$ but we want to know $P(y \mid x)$, we can use the Bayes theorem in Eq. (7). $P(y \mid x)$ is called the posterior of $y$, that is, what we know about $y$ after we observe $x$. $P(x \mid y)$ is called the likelihood. $P(y)$ is the prior of $y$, that is, what we know about $y$ before we observe $x$. $P(x)$ is called the evidence and often it is treated only as a normalization constant, because it is constant w.r.t. $y$.

Marginalization principle enables us to get rid of uninteresting variables. Say we are interested in $P(x)$ but we only know $P(x, y)$. We just take the sum over the possible values of $y$:

$$P(x) = \sum_y P(x, y) = \sum_y P(x \mid y)P(y).$$  \hspace{1cm} (8)

This is useful for instance to handle the evidence $P(x)$ in the Bayes theorem (Eq. 7).

What are Bayesian networks then? Recall from Equation (5) that we can always write the joint distribution of variables as a product of factors where one variable at a time is conditioned on previous ones. Then, we can make some independency assumptions by dropping out some of the conditioned variables. For instance if we assume that $y$ and $z$ are independent given $x$ (that is, $y \perp z \mid x$), we can write

$$P(x, y, z) = P(x)P(y \mid x)P(z \mid x, y)$$  \hspace{1cm} (9)

$$= P(x)P(y \mid x)P(z \mid x).$$  \hspace{1cm} (10)

By making enough independency assumptions, a Bayesian network becomes an efficient representation of the joint probability.