Bayesian Networks 2016–2017

Tutorial I – Basics of Probability Theory and Bayesian Networks

Peter Lucas
LIACS, Leiden University

Elementary probability theory

We adopt the following notations with respect to probability distributions and Boolean variables. Let X denote a variable; if X is a *binary* variable, e.g. taking either the value *true* or *false*, $X = \mathit{true}$ is also denoted by simply x; similarly, $X = \mathit{false}$ is also referred to by $\neg x$.

Furthermore, when it is not really important to know what value a variable takes, we just refer to the variable itself. For example, $P(X)$ either stands for $P(x)$ or $P(\neg x)$. Of course, a statement like $P(X) = 0.7$ would be silly, as in this case it does matter where X really stands for: if it would be x, then $P(x) = 0.7$ and $P(\neg x) = 0.3$, and if it would be $\neg x$, then $P(\neg x) = 0.7$ and $P(x) = 0.3$.

Finally, the expression $\sum_X P(X)$ means summing over all possible values of the variable X; if X is binary, then

$$\sum_X P(X) = P(x) + P(\neg x)$$

This notation can be generalised for joint probability distributions, e.g.

$$\sum_{X,Y} P(X, Y, a) = P(x, y, a) + P(\neg x, y, a) + P(x, \neg y, a) + P(\neg x, \neg y, a)$$

Of course, the same notation can also be used for conditional probability distributions, as those conform to the axioms of probability theory as well:

$$\sum_X P(X \mid Y) = P(x \mid Y) + P(\neg x \mid Y)$$

Exercise 1

Let P be a probability distribution defined on the set of Boolean expressions $\mathcal{B}$, i.e. the following axioms hold for P:

- $P(\top) = 1$;
- $P(\bot) = 0$;
- $P(x \vee y) = P(x) + P(y)$, if $x \wedge y = \bot$, i.e. x and y are disjoint, $x, y \in \mathcal{B}$.

a. Draw a Venn diagram with x, y and $x \wedge y$, for the case that $x \wedge y \neq \bot$.

b. Now prove that the following holds in general:

$$P(x \vee y) = P(x) + P(y) - P(x \wedge y)$$

Exercise 2

Let P be a *joint* probability distribution, defined as follows (note that $P(A, B)$ is a shorthand for $P(A \wedge B)$).

$P(a, b) = 0.3 \quad P(\neg a, b) = 0.2$
$P(a, \neg b) = 0.4 \quad P(\neg a, \neg b) = 0.1$

a. Compute $P(a)$ and $P(b)$.

b. Compute $P(a \mid b)$.

c. Using these last results, use Bayes' rule to compute $P(b \mid a)$.

Exercise 3

a. Prove that Bayes' rule holds, based on the definition of conditional probabilities:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

b. Why is this rule sometimes called the *arc reversal* rule?

c. Let $P(B \mid A_1, \ldots, A_n)$ be a family of conditional probability distributions. Explain why computing $P(B \mid A_1, \ldots, A_n)$ may be computationally hard when the variables are discrete (possibly binary), and discuss possible solutions.

Exercise 4

Consider joint probability distribution $P(X1, X_2, X_3, X_4)$:

$$\begin{aligned}
P(x_1, x_2, x_3, x_4) &= 0.1\\
P(x_1, \neg x_2, x_3, x_4) &= 0.04\\
P(x_1, x_2, \neg x_3, x_4) &= 0.03\\
P(x_1, x_2, x_3, \neg x_4) &= 0.1\\
P(\neg x_1, x_2, x_3, x_4) &= 0.0\\
P(\neg x_1, \neg x_2, x_3, x_4) &= 0.2\\
P(\neg x_1, x_2, \neg x_3, x_4) &= 0.08\\
P(\neg x_1, x_2, x_3, \neg x_4) &= 0.1\\
P(x_1, \neg x_2, \neg x_3, x_4) &= 0.015\\
P(x_1, \neg x_2, x_3, \neg x_4) &= 0.1\\
P(x_1, x_2, \neg x_3, \neg x_4) &= 0.004
\end{aligned}$$

$$\begin{aligned}
P(\neg x_1, \neg x_2, \neg x_3, x_4) &= 0.005\\
P(\neg x_1, \neg x_2, x_3, \neg x_4) &= 0.01\\
P(\neg x_1, x_2, \neg x_3, \neg x_4) &= 0.01\\
P(x_1, \neg x_2, \neg x_3, \neg x_4) &= 0.006\\
P(\neg x_1, \neg x_2, \neg x_3, \neg x_4) &= 0.2
\end{aligned}$$

a. Compute $P(x_2 \vee \neg x_3 \mid x_1 \wedge x_4)$.

b. Split up $P(X_1, X_2, X_3, X_4)$ into three factors, and compute the values of these factors for $X_j = \top$, $j = 1, \ldots, 4$.

Exercise 5

Let X, Y and Z three different stochastic variables. Assume that X is conditionally independent of Y given Z.

a. Try to define a probability distribution $P(X \mid Y, Z)$ (i.e. supply the numbers for that distribution) for which the conditional independence property holds.

b. Prove that from $P(X \mid Y, Z) = P(X \mid Z)$ it follows that $P(Y \mid X, Z) = P(Y \mid Z)$.

Exercise 6

a. Prove that if the sets of variables X and Y are conditionally independent given the set of variables Z, i.e.

$$P(X \mid Y, Z) = P(X \mid Z)$$

it follows that

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

b. Explain why the use of marginalisation is often combined with conditioning.

Exercise 7

The probability distributions of Bayesian networks are often based on the special probability distributions one finds described in textbooks of probability theory and statistics. Examples are:

- the *Bernoulli* distribution: $f(x; p) = p^x(1-p)^{1-x}$, where p is the parameter of the distribution (the probability of success of an experiment) and x is the outcome (0, i.e. failure, or 1, i.e. success). So, $f(0) = 1 - p$ and $f(1) = p$.

- the *binomial* distribution: $f(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}$, where p is again the probability of success; n represents the number of experiments and x the number of successes out of n (thus, there are $n - x$ failures). Note the relationship of this distribution with the Bernoulli distribution.

The notation '$f(\cdot;\cdot)$' says that the elements after the semicolon are *parameters*. We also have the notation '$f(\cdot,\ldots,\cdot)$' for joint probability distributions, where the parameters are uncertain.

Now, represent each of these distributions as two Bayesian networks, where for

- BN1: the parameters p and n are assumed to be fixed;
- BN2: the parameters p and n are assumed to be uncertain, assuming that $f(x,p) = P(x \mid p)g(p)$ for Bernoulli and $f(x,p,n) = P(x \mid p,n)g(p)h(n)$ for the binomial distribution (here g and h are probability density – the continuous case – or mass functions – the discrete case).

Bayesian networks and naive probabilistic reasoning

Recall that a Bayesian network $\mathcal{B}$ is defined as a pair $\mathcal{B} = (G,P)$, where $G = (V(G), A(G))$ is an acyclic directed graph with set of vertices (or nodes) $V(G) = \{X_1,\ldots,X_n\}$ and arcs $A(G) \subseteq V(G) \times V(G)$, and a joint probability distribution P defined on the variables corresponding to the vertices $V(G)$, as follows:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i))$$

where $\pi(X_i)$ stands for the set of parents (direct ancestors) of X_i. Another, more precise version of the definition of the joint probability distribution of a Bayesian network, is to say that $X_{V(G)}$ is the set of random variables corresponding to the vertices of G, where $V(G) = \{v_1,\ldots,v_n\}$ are the associated vertices. The joint probability distribution P is then defined as follows:

$$P(X_{V(G)}) = \prod_{v \in V(G)} P(X_v \mid X_{\pi(v)})$$

where $X_{\pi(v)}$ are the random variables that correspond to the parents of vertex $v \in V(G)$. However, although mathematically more beautiful, the latter definition is also slightly more difficult to understand. It is often used in mathematically oriented book, but less often in computer-science books.

Exercise 8

Consider the Naive Bayes network given in Figure 1. Given the evidence: $\mathcal{E} = \{temp > 37.5\}$, compute the probability of *flu* using Bayes' rule, i. e., $P(flu \mid temp > 37.5)$.

Exercise 9

Consider Figure 2, which displays a Bayesian network in which the three vertices A, B and C interact to cause effect D through a noisy-AND. The intermediate variables I_A, I_B and I_C are note indicated in the figure, but it is assumed that for computational purposes these intermediate variables are implicitly present. The following probabilities have been specified by the designer of the Bayesian network:

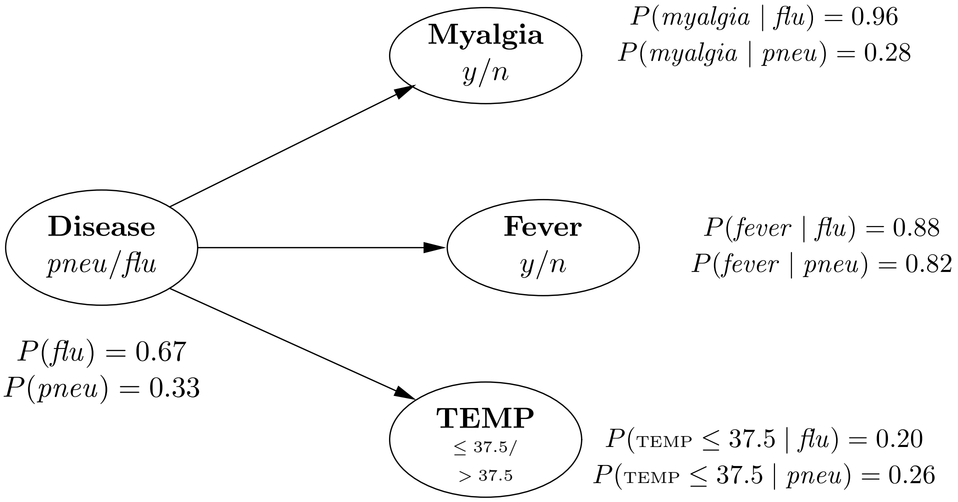
Myalgia
y/n
$P(myalgia \mid flu) = 0.96$
$P(myalgia \mid pneu) = 0.28$
Disease
$pneu/flu$
Fever
y/n
$P(fever \mid flu) = 0.88$
$P(fever \mid pneu) = 0.82$
$P(flu) = 0.67$
$P(pneu) = 0.33$
TEMP
$\leq 37.5/$
> 37.5
$P(\textsc{temp} \leq 37.5 \mid flu) = 0.20$
$P(\textsc{temp} \leq 37.5 \mid pneu) = 0.26$

Figure 1: Naive Bayesian network.

$$\begin{array}{ll}
P(i_A \mid a) = 0.7 & P(i_A \mid \neg a) = 0.9 \\
P(i_B \mid b) = 0.4 & P(i_B \mid \neg b) = 0.8 \\
P(i_C \mid c) = 0.3 & P(i_C \mid \neg c) = 0.3
\end{array}$$

$$\begin{array}{ll}
P(a) = 0.4 & P(b) = 0.7 \\
P(c) = 0.8 &
\end{array}$$

$$\begin{array}{ll}
P(e \mid d) = 0.2 & P(e \mid \neg d) = 0.6
\end{array}$$

a. Compute $P^*(e) = P(e \mid a, b, c)$, i.e. the marginal probability of e given that $A = B = C = true$.

b. Compute $P^*(e) = P(e \mid a, b)$.

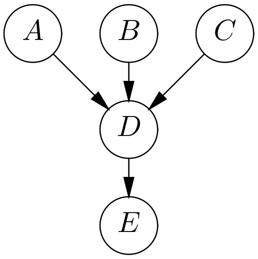

Figure 2: Bayesian network: noisy-AND.

Exercise 10

Note: this exercise is similar to the exercise at the end of Lecture 2.

Consider Figure 3.

a. Define a probability distribution P for this Bayesian network $\mathcal{B}$.

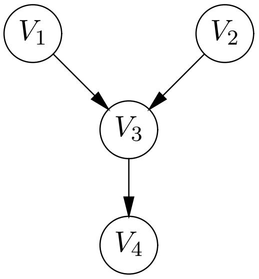

Figure 3: Bayesian network.

b. Compute the marginal probability distribution $P(V_4)$.

c. Next, assume that that we know that v_2 ($V_2 = \textit{true}$. What is the value of $P^*(v_2)$ and $P^*(\neg v_2)$? Compute $P^*(V_4)$, i.e. $P^*(v_4)$ and $P^*(\neg v_4)$, and also $P^*(V_1)$.

d. Finally, assume that $V_4 = \textit{true}$ (hence, V_2 is again unknown). Compute the marginal probability distribution $P^*(V_2)$.

e. Based on the probability distribution P defined in 3a, for the network $\mathcal{B} = (G, P)$ shown in Figure 3, compute the joint probability distribution of V_1, V_2, V_3 and V_4 for a fixed set of values; for example $P(v_1, \neg v_2, v_3, v_4)$.

Exercise 11

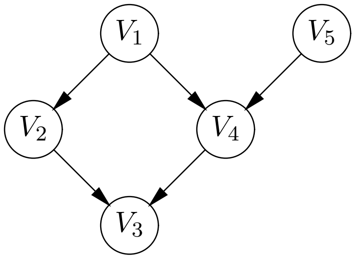

Figure 4: Bayesian network of Exercise 11.

Consider the Bayesian network $\mathcal{B} = (G, P)$ shown in Figure 4, where $G = (V(G), A(G))$ is the directed acyclic graph shown in the figure, and P is a probability distribution defined on the variables corresponding to the vertices $V(G) = \{V_1, V_2, V_3, V_4, V_5\}$. The following (local) probability distributions are defined for P:

$P(v_1) = 0.2$
$P(v_5) = 0.8$
$P(v_2 \mid v_1) = 0.5$ $\quad P(v_2 \mid \neg v_1) = 0.4$
$P(v_3 \mid v_2, v_4) = 0.4$ $\quad P(v_3 \mid \neg v_2, v_4) = 0.7$
$P(v_3 \mid v_2, \neg v_4) = 0.3$ $\quad P(v_3 \mid \neg v_2, \neg v_4) = 0.6$
$P(v_4 \mid v_1, v_5) = 0.6$ $\quad P(v_4 \mid \neg v_1, v_5) = 0.2$
$P(v_4 \mid v_1, \neg v_5) = 0$ $\quad P(v_4 \mid \neg v_1, \neg v_5) = 1$

a. Compute the probability $P(\neg v_5 \mid v_3)$, i.e. for V_5 equal to false given that V_3 is equal to true, if it is known that $P(v_3) \approx 0.5$.

b. What are the consequences of the fact that this Bayesian network contains a cycle in the underlying (undirected) graph?

Exercise 12

Consider the Bayesian network $\mathcal{B} = (G, P)$ shown in Figure 3.

a. Enumerate all conditional independence assumptions $U \perp\!\!\!\perp V \mid W$ for this Bayesian network.

b. Enumerate all the unconditional independence assumptions.

c. Enumerate the dependencies $U \not\!\perp\!\!\!\perp V|W$.