IR2 – Data Mining 2002–2003

Exercises III

Peter Lucas
Institute for Computing and Information Sciences
University of Nijmegen

1 Introduction

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 $P(x) = 0.3$. The notation $P(X)$ also stands for the (marginal) probability distributions with regard to variable X.

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:

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

2 Bayesian networks: representation and inference

Exercise 1

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

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

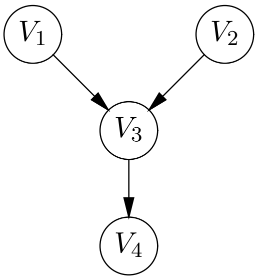

Figure 1: Bayesian belief network.

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 2

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 conditional probability distribution. Explain why computing $P(B \mid A_1, \ldots, A_n)$ may be computationally hard, and discuss possible solutions.

Exercise 3

Let X, Y and Z three different stochastic variables. Assume that X is conditionally independent of Y given Z, formally: $X \perp\!\!\!\perp Y \mid Z$.

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 $P(Y \mid X, Z) = P(Y \mid Z)$, or logically: $X \perp\!\!\!\perp Y \mid Z \rightarrow Y \perp\!\!\!\perp X \mid Z$ (symmetry axiom).

Exercise 4

Consider the Bayesian network $\mathcal{B} = (G, P)$ shown in Figure 1.

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

b. Argue why the following property holds:

$$P(V_1, V_2, V_3, V_4) = \prod_{i=1}^{4} P(V_i \mid \pi(V_i))$$

where $\pi(V_i)$ stands for the set of parents (direct ancestors) of vertex V_i.

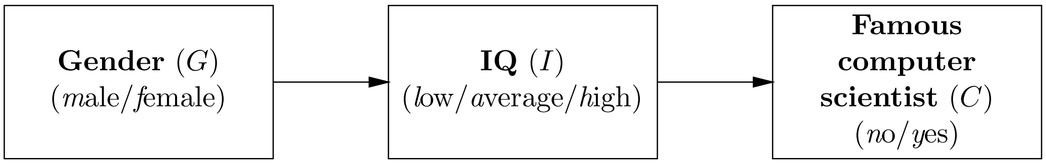

Figure 2: Example Bayesian network.

Exercise 5

Consider again the Bayesian network depicted in Figure 1.

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

b. Next, assume that we know that v_2 ($V_2 = true$) holds. What is the value of $P(v_2 \mid v_2)$ and $P(\neg v_2 \mid v_2)$? Compute $P(V_4 \mid v_2)$, i.e. $P(v_4 \mid v_2)$ and $P(\neg v_4 \mid v_2)$, and also $P(V_1 \mid v_2)$.

3 Bayesian networks: data mining

Exercise 6

Consider the Bayesian network shown in Figure 2, and the following (tiny) dataset D:

Student	**Gender**	**IQ**	**Famous computer scientist**
1	male	low	no
2	female	average	yes
3	male	high	yes
4	female	high	yes
5	female	low	no
6	male	high	no

Note that a probability can be taken as *relative frequency*:

$$P(X = x \mid Y = y) = \frac{N_{x \wedge y}}{N_y}$$

where N_x denotes the number of cases with x ($N_\top$: total number of cases).

a. Fill in the following tables:

$P(G)$	
m	f

	$P(I \mid G)$		
G	l	a	h
m			
f			

	$P(C \mid I)$	
I	n	y
l		
a		
h		

b. Professor D.K. has prior knowledge about the likelihood that a student becomes a famous computer scientist, given a particular IQ, based on having seen 200 students. We have

- $n_0 = 200$;

- Prof. D.K. says that 80% of students with a high IQ have become famous computer scientists.

Using the same dataset D as above and this pior knowledge, compute $P_D(C = y \mid I = h)$.

c. Discuss how this probability has changed in light of data with respect to prior knowledge.

Exercise 7

Now we change the topology of the Bayesian network shown in Figure 2 to that of a naive Bayesian network with C (famous computer scientist) as a class variable.

a. Compute the required (conditional) probability distributions using relative frequencies as done in Exercise 6.a.

b. Compare the resulting probabilities to those computed in Exercise 6.a.

c. Compute $P(C \mid \textit{female})$ for both probability distributions, and compare the results.

Exercise 8

Finally, change the topology of the Bayesian network shown in Figure 2 to that of a tree-augmented Bayesian network (TAN) with C (famous computer scientist) as a class variable.

a. Compute the required (conditional) probability distributions using relative frequencies as done in Exercises 6.a and 7.a.

b. Compare the resulting probabilities to those computed in Exercises 6.a and 7.a.

c. Compute $P(C \mid \textit{female})$ and compare the results to those of Exercises 6.c and 7.c.