Bayesian Networks 2016-2017

Assignment I – Building a Bayesian Network

1 Introduction

The aim of this assignment is to build up experience in developing Bayesian networks for *realistic problems*. The exercises included in this assignment learn you something about the relationship between consulting Bayesian networks, using tools such as SamIam (see below), and problem solving. *The problem domain for which you have to develop a Bayesian network you have to select yourself. You have* **to complete this assignment in pairs** *(just select your favourite student colleague to work on the assignment).* Chapter 9 (10 in the 2nd edition) in the book "Bayesian Artificial Intelligence" [2] describes methods for systematically building Bayesian networks, and can be used as a source of inspiration. Read this chapter first.

We start by describing the software tool for building Bayesian networks by hand, which is used in this assignment. This is followed by a summary of basic theory and simple examples in Section 3. Some exercises with real-world example Bayesian networks are included in Section 4. Problem solving using Bayesian networks is discussed in Section 5. Finally, *Section 6 describes what you have to do and what to submit in order to complete Assignment I.*

2 Software tools

SamIam is a software tool for the creation and consultation of Bayesian networks. Bayesian networks are graph-based formalisms for the representation and manipulation of uncertain knowledge, based on probability theory. The exercises described in this document are meant to increase your understanding of what Bayesian networks are, and to enhance your impression of what can be done with Bayesian networks.

The SamIam software package is java-based and runs on all operating systems. The instructions for downloading and setting up the software are provided below:

1. Download the software package in zip format from

   `http://reasoning.cs.ucla.edu/samiam/`

   and extract its content on your local machine. This results in the folder "samiam".

2. For Window users: download the batch file samiam2.bat from

   `http://cs.ru.nl/~peterl/teaching/CI/samiam2.bat`

   and save it in the folder "samiam" just created. For Unix-like OS (Linux, MacOS) users, there is a runsamiam script in the directory.

The book "Bayesian Artificial Intelligence" also mentions a number of alternative software packages. In the remainder of this practical it is assumed that you use SamIam.

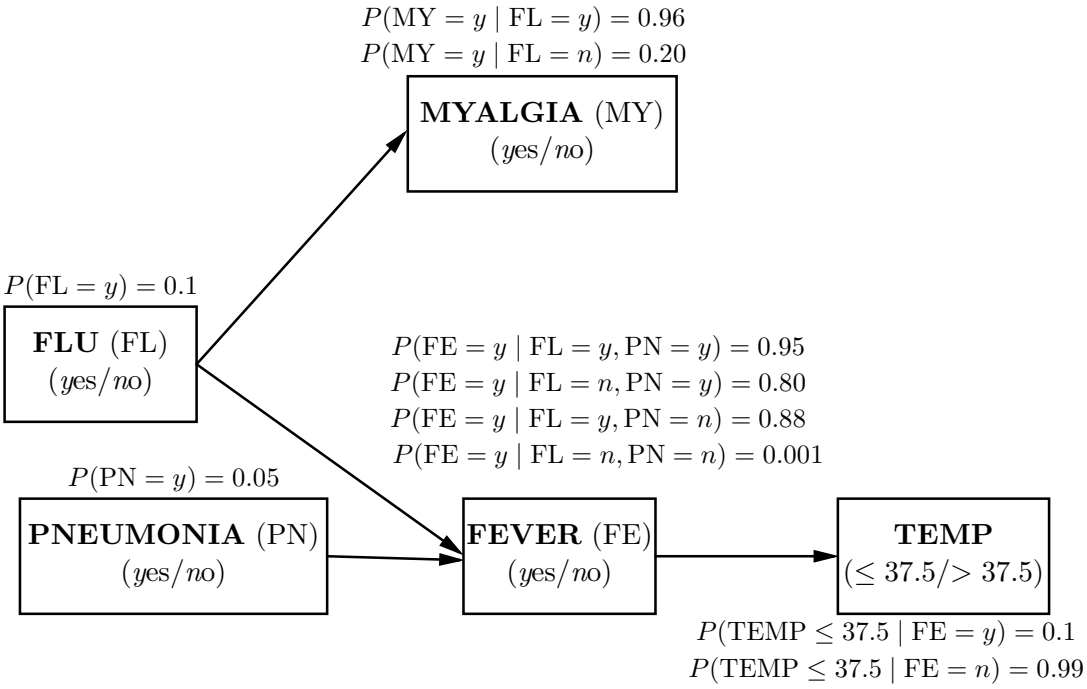


Figure 1: Bayesian network describing some signs and symptoms of flu and pneumonia; variable names have been abbreviated.

3 Basic aspects of Bayesian networks

In the first few exercises, we go through the basic aspects of Bayesian networks.

3.1 What is a Bayesian network?

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 a set of vertices (or nodes) $V(G) = \{X_1, X_2, \ldots, X_n\}$ and a set of arcs $A(G) \subseteq V(G) \times V(G)$, and where P is a joint probability distribution defined on the variables corresponding to the vertices $V(G)$ [5]. The basic property of a Bayesian network is that the joint probability distribution $P(X_1, X_2, \ldots, X_n)$ is equivalent to the product of the (conditional) probabilities which are specified for the network; formally:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \pi(X_i))$$

where $\pi(X_i)$ is the set of parents of the vertex corresponding to the variable X_i. Thus, $P(X_i \mid \pi(X_i))$ are the (conditional) probability distributions which are specified for the variable X_i, for $i = 1, \ldots, n$, in creating a Bayesian network.

Consider the Bayesian network in Figure 1, which concerns the signs and symptoms of flu and pneumonia. Its implementation for SamIam is in the file `flu.net` in the directory `http://cs.ru.nl/~peterl/teaching/CI/networks`. Figure 2 displays a screen-shot of SamIam after the network has been loaded.

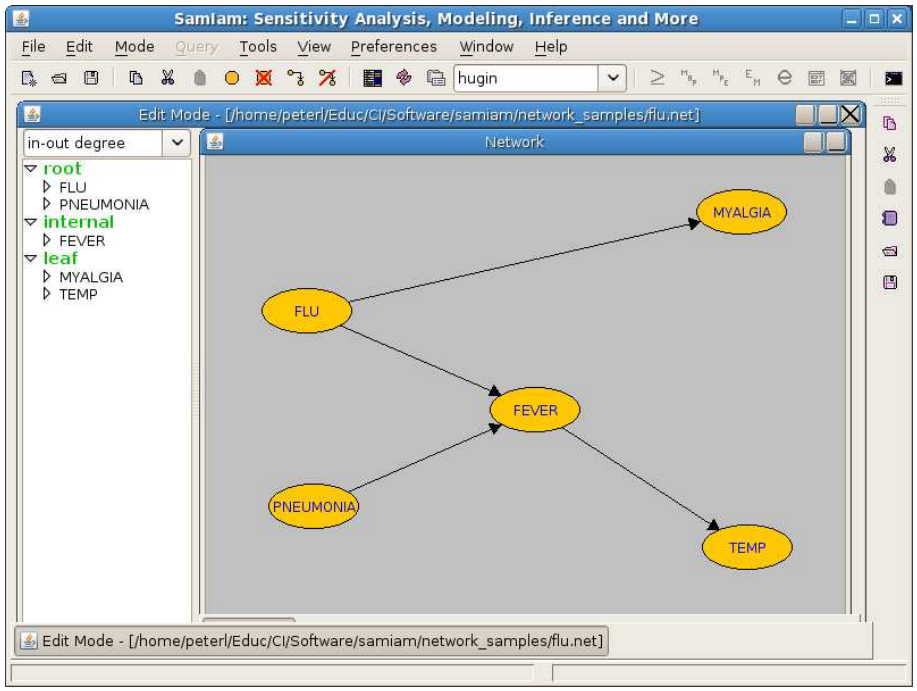


Figure 2: Screen-shot of SamIam with the flu example network.

- *Start* SamIam *and load the flu network. Check whether all probabilities are identical to those in Figure 1 by right clicking on the vertices and inspecting the 'properties', and determine the correspondence between the probabilities* $P(X_i \mid \pi(X_i))$*, e.g.* $P(\text{FE} = y \mid \text{FL} = y, \text{PN} = n)$*, and the numbers in the table.*

- *Select from the 'Mode' menu the 'Query mode'. Next, from the 'Query' menu select from the 'Show monitors' submenu the option 'Show All'. The marginal probabilities* $P^*(X_i)$ *will now be displayed in monitor windows. Instantiating variables, i.e. after assigning values to variables, can be done by left clicking into a probability bar. What are the probabilities* $P^*(\text{FEVER} = yes)$*,* $P^*(\text{TEMP} > 37.5)$*,* $P^*(\text{FLU} = yes)$*,* $P^*(\text{PNEUMONIA} = yes)$ *and* $P^*(\text{MYALGIA} = yes)$*?*

- *Left click on the green bar associated with the value '*> 37.5*' of the variable* TEMP *(which will turn red as a consequence); in this way you enter values for a variable. Note whether the probabilities* $P^*(\text{FEVER} = yes)$*,* $P^*(\text{TEMP} > 37.5)$*,* $P^*(\text{FLU} = yes)$*,* $P^*(\text{PNEUMONIA} = yes)$ *and* $P^*(\text{MYALGIA} = yes)$ *have changed after you entered the evidence into the network.*

The *marginal probability distribution* $P^*(X_i)$ for each variable X_i in a Bayesian network is computed using the probabilistic information P that is supplied with the network. Of course, if you enter a value for a variable into a Bayesian belief network, its probability distribution may, and in most cases actually will, change. So, P^* denotes this new, *updated*, probability distribution. Thus it holds, for example, that $P^*(\text{TEMP} > 37.5) = 1$ after you have entered the value '> 37.5' for the variable TEMP into the network; this is what could be expected, as we know that '$\text{TEMP} > 37.5$' must hold with absolute certainty (as we have observed it). However, it still holds that $P(\text{TEMP} > 37.5 \mid \text{FE} = y) = 0.9$, and the remaining probabilities $P(\text{TEMP} \mid \text{FE})$ are also unchanged. Also note that $P(X_i)$ (so again the original probability distribution P), which is the marginal probability distribution for variable X_i, has also not

changed, as it is by definition the probability distribution for a variable X_i without taking into account evidence $\mathcal{E}$. Of course, if you have entered evidence into the network, you will normally not see $P(X_i)$ displayed on the screen (but see below).

There is a straightforward relationship between the original probability distribution P and the *updated* probability distribution P^*:

$$P^*(X_i) = P(X_i \mid \mathcal{E})$$

or if we now focus on the variable TEMP, it holds that

$$P^*(\text{TEMP} > 37.5) = P(\text{TEMP} > 37.5 \mid \text{TEMP} > 37.5) = 1$$

So, we can compute this updated probability distribution P^* simply by using the probability distribution P. As the computation rules we use are the same, whether or not we take evidence $\mathcal{E}$ into account, we will simply denote any probability distribution resulting from the computations by $P^*(X_i)$ (the updated marginal probability distribution for variable X_i). And so we sometimes have that $P^*(X_i) = P(X_i)$, whereas sometimes it holds that $P^*(X_i) \neq P(X_i)$.

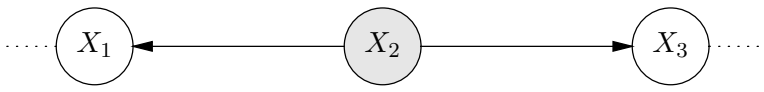
▶ *Remove the evidence 'TEMP > 37.5' from the network by left clicking on the red bar of the* TEMP *vertex. Now, enter the evidence 'FLU = no' into the network. Compare P^*(PNEUMONIA), which you obtain now, with the P^*(PNEUMONIA) as computed in the second exercise above, i.e. before entering any evidence into the network. Also compare this result with the results you obtain if you in addition enter 'TEMP > 37.5' into the network. Try to explain the behaviour you observe.*

3.2 A Bayesian network models (conditional) independence

A Bayesian network is really a graphical representation of stochastic (statistical) dependences and independences among variables. By adding arcs between two vertices we express that its two corresponding variables may influence each other, i.e. may be dependent. By removing arcs between vertices, we are saying that the corresponding variables are (conditionally) independent. The type of dependence and independence we are dealing with is determined by the direction of arcs and whether or not particular variables are instantiated. In the next few exercises, we have a more detailed look at this very important issue.

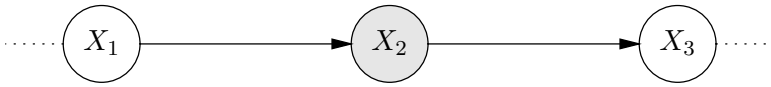
Firstly, consider the possible three patterns which can occur within any Bayesian network, dependent on the direction of the arcs with respect to a central variable X_2 [1]:

- **Diverging arcs:**



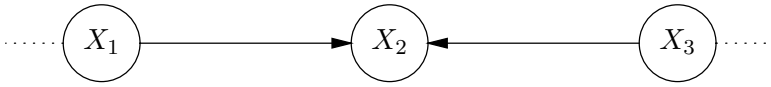
  Here we have that instantiating variable X_2 *blocks* flowing (probabilistic) information, i.e. the influence, from X_1 to X_3 (and vice versa; recall that it is possible to reason in the reverse direction). This means that if X_2 is *not* instantiated, instantiating X_1 may change the marginal probability distribution $P^*(X_3)$ (and vice versa), but if X_2 is instantiated, instantiating X_1 will not change $P^*(X_3)$ (and vice versa). This can be expressed formally using the independence relationship $\perp\!\!\!\perp$ as follows: $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

- **Serial arcs:**



  Here also we have that instantiating variable X_2 will *block* flowing (probabilistic) information from X_1 to X_3 (and vice versa): $X_1 \perp\!\!\!\perp X_3 \mid X_2$. So, the structure of this situation is different from the first one, but the independence relationship modelled is identical to the first one.

- **Converging arcs:**



  This situation is completely different (and actually the opposite) of the other two, as here it holds that instantiating X_2 will make X_1 and X_3 in fact *dependent*, i.e. $X_1 \not\perp\!\!\!\perp X_3 \mid X_2$ (the same holds for the successors of X_2 if present), whereas if X_2 is not instantiated, then X_1 and X_3 are *in*dependent, i.e. formally: $X_1 \perp\!\!\!\perp X_3 \mid \varnothing$.

Instead of the term *blocking*, one also speaks of *d-separation* (*d*irected separation). In the first two cases, it is said that X_2 *d-separates* X_1 and X_3. Instead of saying that two variables are made dependent, we also speak of *d-connection*. In the third case, it is said that X_1 and X_3 are d-connected by X_2 (or by a successor if present).

The notions of diverging, serial and converging arcs have practical significance (otherwise we would not have mentioned them). This can best be illustrated by looking again at the flu example network shown in Figure 1.

▶ *Have a look at the structure of the network in Figure 1, and write down on a sheet of paper which arcs are diverging, serial or converging. Each time you have to select a single (central) vertex, and to look at the arcs connecting this vertex to other vertices.*

▶ *The next step is to experimentally validate your solution to the previous problem. Start with the diverging and serial arcs, and examine the effect of instantiating the central vertex on the flow of probabilistic information. Next, do the same for the converging arcs. In particular, try to find out whether variables which were originally independent, have become dependent by instantiation of variables.*

3.3 Probabilistic inference

There is little doubt that for Bayesian networks of realistic size, having access to a Bayesian network package such as SamIam is really essential for probabilistic inference or reasoning, because otherwise you will run the risk of becoming mad.[1] However, it still makes sense to

[1] More than a decade ago, Peter Lucas worked at the Centre of Mathematics and Computer Science (CWI, `http://www.cwi.nl`) in Amsterdam. This centre is well known for its fundamental research in mathematics and computing science, even before computers became into wide-spread use. In the 1950s and early 1960s, a group of 20 ladies, all gathered in one big room, were employed by CWI to carry out complex numerical computations on paper. These numerical computations were of great importance, as the Dutch government had decided in 1956 that the sea dikes in the Netherlands had to be improved. CWI was involved in finding out how high the dikes had to be to ensure that the Netherlands would not be flooded by sea water in the next millennium, and this involved solving complicated numerical models.

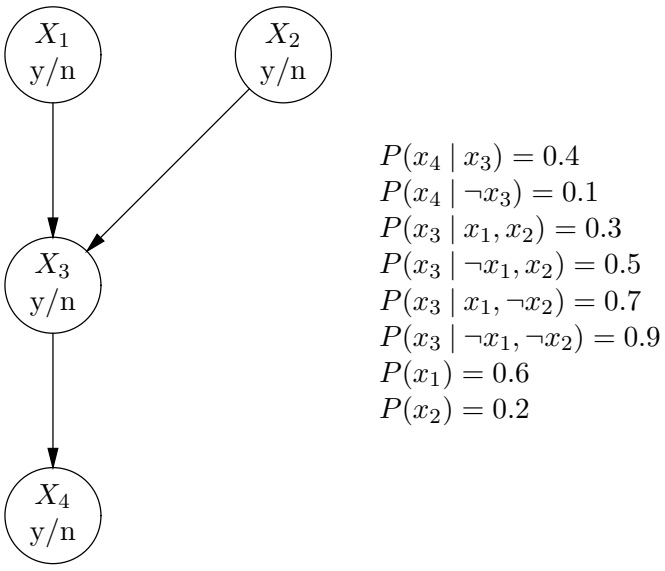


Figure 3: Example network from the lectures.

understand how one could compute probabilities by hand, certainly for a small network, as we as computer scientists are the people who in the end need implement software packages such as SamIam.

Consider the Bayesian network shown in Figure 3. Computation of $P(X_i)$, for $i = 1, \ldots, 4$, goes as follows, starting with X_4:

$$\begin{aligned} P(x_4) &= P(x_4, x_3) + P(x_4, \neg x_3) \\ &\quad \text{(marginalisation)} \\ &= P(x_4 \mid x_3)P(x_3) + P(x_4 \mid \neg x_3)P(\neg x_3) \\ &\quad \text{(conditioning)} \\ &= \sum_{X_3} P(x_4 \mid X_3)P(X_3) \end{aligned}$$

So, we need to know $P(X_3)$:

$$\begin{aligned} P(x_3) &= \sum_{X_1, X_2} P(x_3, X_1, X_2) \\ &= \sum_{X_1, X_2} P(x_3 \mid X_1, X_2)P(X_1, X_2) \end{aligned}$$

We know that

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

as X_1 and X_2 are independent. The conclusion is that:

$$\begin{aligned} P(x_4) &= \sum_{X_3} P(x_4 \mid X_3) \cdot \\ &\quad \sum_{X_1, X_2} P(X_3 \mid X_1, X_2)P(X_1, X_2) \\ &= 0.31 \end{aligned}$$

- *What are $P(X_1)$ and $P(X_2)$?*
- *The file corresponding to the Bayesian network in Figure 3 is* `slide-net.net` *in the directory* `http://cs.ru.nl/~peterl/teaching/CI/networks`. *Load this network into* SamIam *and check the computations above.*

As said above, normally we would use a Bayesian network tool in order to find out the effect of particular evidence on a variable X_i, i.e. $P^*(X_i) = P(X_i \mid \mathcal{E})$. As the computations carried out are similar to the one above, with the exception of the situation when reasoning in the reverse direction of an arc has to be accomplished, because then Bayes' rule comes into play. For example, assume that $X_2 = y$ is entered into the network. Then, the updated marginal probability distribution for X_4 can be computed as follows:

$$\begin{aligned}
P^*(x_4) &= P(x_4 \mid x_2) \\
&= \sum_{X_3} P(x_4 \mid x_2, X_3) P(X_3 \mid x_2) \\
&= \sum_{X_3} P(x_4 \mid X_3) P(X_3 \mid x_2) \\
&= \sum_{X_3} P(x_4|X_3) \sum_{X_1} P(X_3|X_1, x_2) P(X_1|x_2) \\
&= \sum_{X_3} P(x_4|X_3) \sum_{X_1} P(X_3|X_1, x_2) P(X_1) \\
&= 0.214
\end{aligned}$$

- *Check these computations using* SamIam.

Finally, assume that instead of X_2 the variable X_4 is now observed with value y, so it holds that $P^*(x_4) = 1$. We now get:

$$\begin{aligned}
P^*(x_2) &= P(x_2 \mid x_4) \\
&= \frac{P(x_4 \mid x_2) P(x_2)}{P(x_4)}
\end{aligned}$$

This is Bayes' rule. Note that:

$$\begin{aligned}
P(x_4 \mid x_2) &= 0.214 \\
P(x_4) &= 0.31
\end{aligned}$$

So, it holds that:

$$\begin{aligned}
P^*(x_2) &= 0.214 \cdot 0.2/0.31 \\
&\approx 0.1381
\end{aligned}$$

- *Check these computations using* SamIam.

4 Examples of real-world Bayesian networks

Since the beginning of the 1990s researchers have been developing Bayesian networks for many different problems, varying from hardware trouble shooting and user guidance in applying software, to medical diagnosis and treatment. However, most of the networks which have been developed are not in the public domain. This holds in particular for the networks which have been developed by commercial companies such as Hewlett Packard and Microsoft. Some of the networks which have been developed within the medical domain, though, have been placed in the public domain. This is understandable, as in biomedical research there is less emphasis on commercial interests. As a consequence, biomedical Bayesian networks have gained greater visibility than the industrial ones, and some people therefore maintain the opinion that biomedicine is the primary application field of Bayesian networks. Actually, the reverse is true. For example, the two most well-known research groups in Bayesian belief networks, the group at Microsoft Research in the USA and the group at the University of Aalborg, Denmark, are not doing any work in the biomedical field at all. There is currently also a growing interest in the automotive industry to use Bayesian networks for on-board diagnosis of car failure. Again, due to the stiff competition between car factories, the automotive industry is not willing to publish their models in the public domain.

The consequence of the above is that, due to the lack of public availability of non-medical Bayesian networks, we will focus on medical networks. At the end of this practical, however, there is a Bayesian network discussed that has been developed together with people from Hewlett-Packard by researchers from Aalborg University.

One should, however, keep the remarks made above in mind.

4.1 Treatment of non-Hodgkin lymphoma of the stomach

4.1.1 The problem

Non-Hodgkin lymphoma of the stomach, gastric NHL for short, is a relatively uncommon malignant disorder, accounting for about 5% of tumours of the stomach. Until recently, the cause of gastric NHL was unknown; it is now generally believed that the main factor in the development of this disease is a chronic infection with the bacterium *Helicobacter pylori*. This has had a major effect on treatment practice. Whereas originally, as in most cancers, treatment consisted of surgery (total or partial removal of the stomach), chemotherapy, radiotherapy or a combination of two or three of these, there is now also a place for antibiotics. Only 10 years ago, no medical doctor would have believed you when you had said that cancer can be treated by antibiotics. So, the impact of these recent findings has been dramatic.

Now, the selection of treatment for gastric NHL is a complicated process, because only part of the patient findings necessary for therapy selection may be known at a particular stage of the disease, and knowledge of adverse reactions to particular treatments in patient groups may influence treatment selection significantly. This explains why the Netherlands Cancer Institute in Amsterdam considered developing a Bayesian network. It was hoped that a Bayesian network of gastric NHL might help doctors in the prescription of *optimal* treatment of a patient. The network discussed here is still in prototype stage; further development needs to take place in order to introduce it in actual clinical practice.

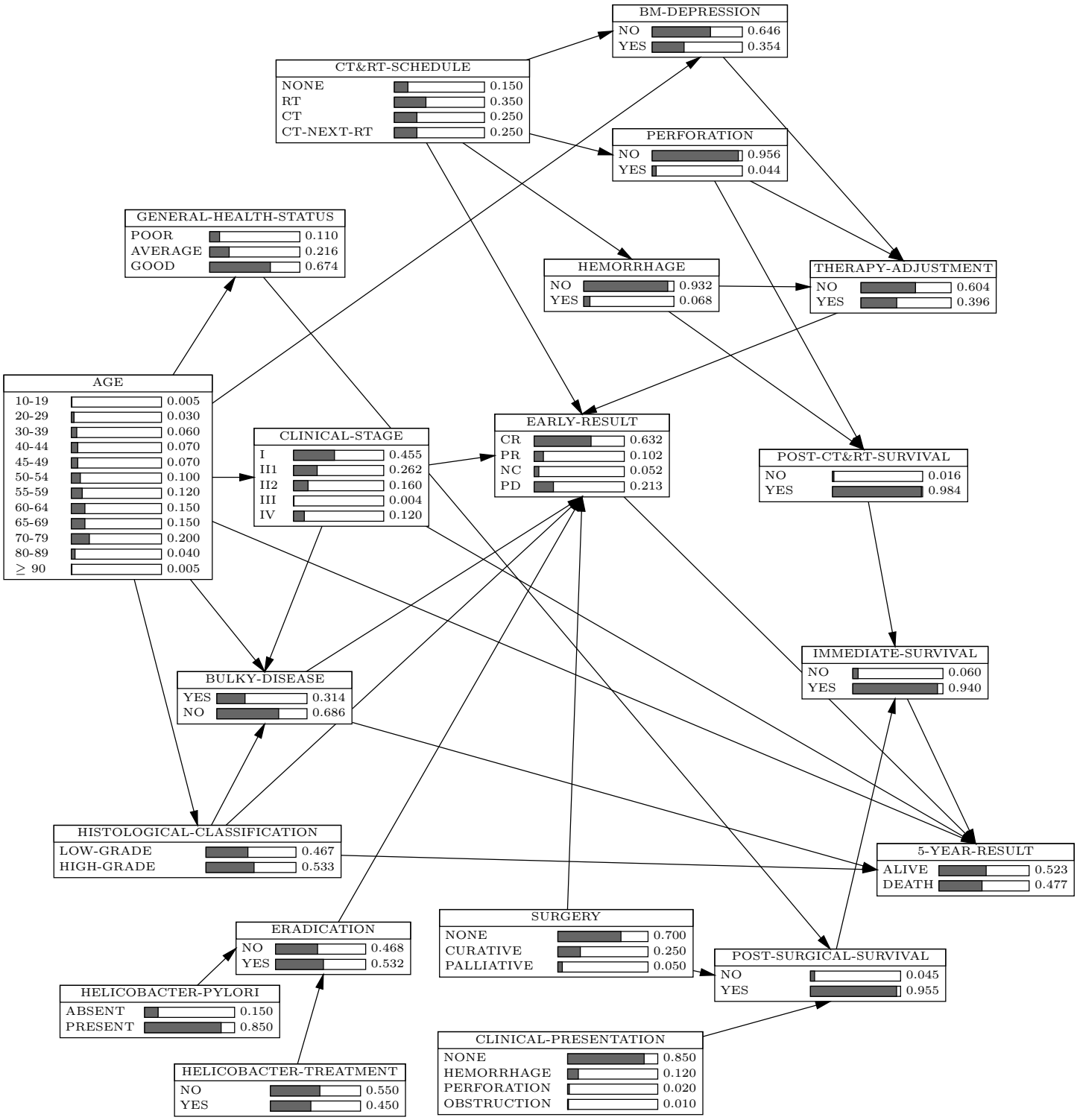


Figure 4: Bayesian network with prior probability distributions for gastric NHL.

4.1.2 Structure of the network

The gastric NHL Bayesian network only incorporates variables that are widely used by clinicians in choosing the appropriate therapy for patients [4]. The relevance of most of these variables is supported by literature on prognostic factors in gastric NHL.

First, the information used in the clinical management of primary gastric NHL was subdivided in *pretreatment information*, i.e. information that is required for treatment selection, *treatment information*, i.e. the various treatment alternatives, and *posttreatment information*, i.e. side effects, and early and long-term treatment results for the disease. The selected variables are shown in Figure 4. The most important pretreatment variables in the table are the variable 'clinical stage', which expresses severity of the disease according to a common clinical classification, and histological classification, which stands for the assessment by a pathologist of tumour tissue obtained from a biopsy.

Various treatments are in use for gastric NHL such as chemotherapy, radiotherapy, and a combination of these two, which has been represented as the single variable 'CT&RT-SCHEDULE' with possible values: chemotherapy (CT), radiotherapy (RT), chemotherapy followed by radiotherapy (CT-next-RT), and neither chemotherapy nor radiotherapy (none). Furthermore, surgery is a therapy with is modelled by the variable 'SURGERY' with possible values: 'curative', 'palliative' or 'none', where curative surgery means total or partial resection of the stomach with the complete removal of tumour mass. Finally, prescription of antibiotics is also possible.

The most important posttreatment variables are the variable 'EARLY RESULT', being the endoscopically verified result of the treatment, six to eight weeks after treatment (possible outcomes are: complete remission – i.e. tumour cells are no longer detectable –, partial remission – some tumour cells are detectable –, no change or progressive disease), and the variable '5-YEAR RESULT', which represents the patient either or not surviving five years following treatment.

4.1.3 Exercises

The following exercises are meant to give you some ideas of how a medical Bayesian network might be used in practice. Table 1 lists features of 11 actual Dutch patients.

▶ *Enter the findings of a number of patient as described in Table 1 into the network* `nhl.net`*, which is available in the directory*

`http://cs.ru.nl/~peterl/teaching/CI/networks`,

using SAMIAM. *For each of these patients, try to determine the treatment which yields the best results.*

▶ *A typical clinical research question might be whether there is a difference between those patients who live shorter and those that live longer than 5 years following treatment. Use the network to answer this question. Is age also a factor that affects the results of this question, i.e. can different patient groups be distinguished?*

N	Age	Gender	Stage	Grade	Mass	HP	Clinical Presentation
1	61	m	I	high	non-bulky	−	gastric perforation
2	39	m	I	high	bulky	−	—
3	64	f	I	low	non-bulky	+	—
4	63	m	II_1	high	borderline	+	reflux oesophagitis gastric obstruction
5	77	m	II_1	low	non-bulky	+	—
6	82	f	II_1	high	bulky	−	—
7	46	m	II_1	high	non-bulky	+	impaired right kidney
8	60	m	I	high	non-bulky	+	—
9	47	m	I	high	bulky	+	—
10	67	f	IV	high	non-bulky	−	—
11	73	m	IV	high	bulky	−	—

Stage: Clinical stage according to the Ann Arbor classification of NHL by Musshoff
Grade: histological MALT classification
HP: H. pylori present (+) or absent (−)

Table 1: Selected features of 11 actual patients.

4.2 Anaesthesia problems

4.2.1 The problem

During an operation, a patient's vital functions, such as heart rate and blood pressure, are measured in order to act as rapidly as possible if one of these functions deteriorate. However, many of these functions interact which each other, so that a change in one may change the other, and this can be very confusing for the anaesthetist who is responsible for keeping a patient stable during an operation.

The ALARM Bayesian network has been developed in order to assist anaesthetists in interpreting changes in vital signs in patients. Again, this network was a research prototype, and has never been used in practice. However, the network is very popular within the Bayesian network research community, where it has been used for the evaluation of all sorts of research ideas.

4.2.2 Structure of the network

We do not discuss the network in detail, but will focus on a small part of it. The stroke volume of the heart, which is the amount of blood pumped out of the heart at every beat, is determined by the amount of blood available and by the functional capabilities of the heart muscle. In hypovolaemia, the amount of blood is significantly decreased. The variable 'HYPOVOLEMIA' models this situation. If the heart muscle fails to meet its requirements, we say that the patient has a left-ventricular failure; this is modelled by the variable 'LVFAILURE'. In turn, the state of the stroke volume is modelled by the variable 'STROKEVOLUME'. The cardiac output of the amount of blood pumped out of the heart per minute is the *cardiac output* (CO). It is about 5 l/min. This is of course determined by the stroke volume and the heart rate. The heart rate is modelled by the variable 'HR'. Finally, the CO is one of the factors that determines the blood pressure, which is modelled in the network by the variable

'BP'. The resistance of the blood vessels to blood flow is another factor that determines blood pressure. This is modelled by the variable 'TPR' (Total Peripheral Resistance). This in turn may decrease significantly in the case of anaphylaxis. This is an allergic reaction characterised by a drop in TPR, e.g. after having been stung by a bee.

4.2.3 Exercises

▶ *Load the network* `alarm.net` *into* SamIam. *Find out what may happen to a patient with a history of left ventricular failure (LVF), and low blood pressure due to anaphylaxis, e.g. after particular drugs have been administered to the patient.*

4.3 Printer trouble shooting

4.3.1 The problem

Helping the customer tracking down problems with a printer is a typical issue investigated by hardware firms, such as Hewlett-Packard (HP). Their current printer assistant software does offer some limited support in this. Also Microsoft has been involved in the development of Bayesian-network models for printer trouble shooting.

4.3.2 Structure of the network

Modelling for trouble shooting is often done in terms of cause-effect relationships, i.e. an initial design of a Bayesian network is actually a causal graph, where arcs $C \rightarrow E$ have the meaning of a cause-effect relationship. This is then converted into a Bayesian network, by simply assuming that the given cause-effect structure yield the right structure of the Bayesian network.

4.3.3 Exercises

An example of such a network, which has been developed with engineers from HP is the `printer-ts.net` network, which can be downloaded at:

`http://cs.ru.nl/~peterl/teaching/CI/networks`

The structure of this network is clearly causal in nature, e.g.

$$(\text{Net Printer on and online} \wedge \text{Net Printer Paper Supply}) \rightarrow \text{Net Printer OK}$$

expresses that a net printer is operating correctly if it is on and online, and there is sufficient paper supply.

▶ *Load the network* `printer-ts.net` *into* SamIam. *Look at the structure of the network and decide whether you agree with the chosen structure. If you do not agree, modify the structure of the network.*

▶ *Now note that the probabilities are all uniform, and so the network is useless. Change the probabilities using your knowledge about printers and investigate whether the resulting Bayesian network is behaving according to your expectations.*

4.4 Other networks

One other interesting Bayesian network is the one described in [3], and concerns the diagnosis of a very rare disease: Wilson's disease. *Load the network* `wilson.net` *from:*

`http://cs.ru.nl/~peterl/teaching/CI/networks`

into SamIam*, and read the explanation in Ref. [3].*

5 Problem solving

As we have seen above, Bayesian networks can be used as a formalism to solve various tasks. A common application is *diagnosis*, for example in medicine or in engineering. Typically, a user enters case data into a Bayesian network, after which a diagnosis can be determined using a probabilistic inference algorithm. Other applications are: *population description*, *decision making*, i.e., selecting an action from a list of possible actions, *prediction*, and *profiling*.

5.1 Exercise

Consider the screenshot of the Bayesian network concerning the three different disorders tuberculosis, lung cancer and bronchitis shown in Figure 5. This network represents the following problem.

> Shortness-of-breath (dyspnoea) may be due to tuberculosis, lung cancer or bronchitis, or more than one of them. A recent visit to Asia increases the chances of tuberculosis, while smoking is known to be a risk factor for both lung cancer and bronchitis. A single chest X-ray does not discriminate between lung cancer and tuberculosis, as neither does the presence or absence of shortness-of-breath.

5.1.1 Population description: a priori probability distribution

- *Load the Bayesian network*

  `http://www.cs.ru.nl/~peterl/teaching/CI/chest_clinic.net`

  into SamIam *and try to establish the (in)dependences among the variables concerned.*

- *Try to determine which vertices can act as* observables *or* findings*, as* class vertices*, or as* conditioning vertices*.*

- *Compute the marginal probability distributions of the individual variables.*

  Note that class vertices are used for establishing diagnoses.

5.1.2 Establishing a diagnosis

Assume that we are dealing with a person with dyspnoea, who has visited south Asia.

- *What is the most likely diagnosis?*

- *Does the diagnosis change if we assume that the person is a heavy smoker?*

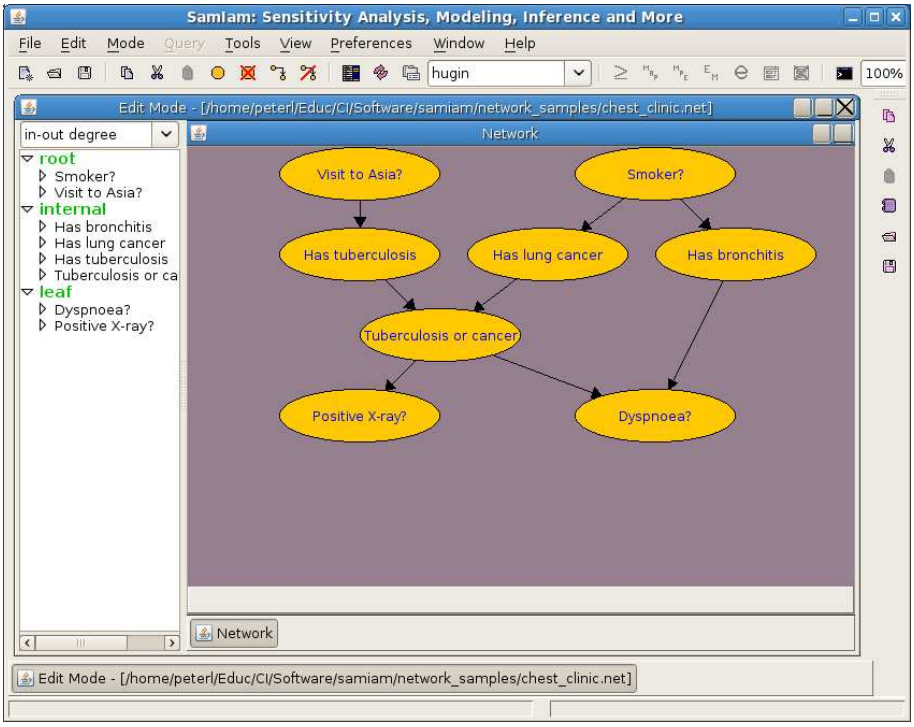


Figure 5: Screen-shot of SamIam with the chest-clinic network.

5.1.3 Prediction

Remove all the evidence $\mathcal{E}$ entered into the network.

▶ *Predict the likelihood that a smoker will develop dyspnoea.*

5.1.4 Decision making

In the first exercise in Section 4.1, we already investigated a method for decision making, here finding the optimal treatment (optimal combination of actions) for an NHL patient. This network is also suitable to illustrate description of a population and prediction, but it cannot be used for diagnostic purposes as the model assumes that a diagnosis is already made for a patient.

5.1.5 Profiling

Finally, in the second exercise in Section 4.1 we already studied profiling:

▶ *Determine the characteristics of the population of NHL patients who died within 5 years, and compare these characteristics with patients who survived more than 5 years.*

All examples together illustrate the big versatility of the Bayesian network formalism.

6 Practical assignment

Design a Bayesian network for a problem of your own choice together with your coworker. The resulting network model together with the associated report are part of practical assignment I.

- Select a domain of your interest in which cause-effect relationship play an important role, i.e. it must be possible to describe the problem domain in terms of cause-effect relationships.
- Design a causal graph in which all relevant causal relationships are linked to each other.
- Transform the resulting causal network into a Bayesian network by selecting the most relevant part of the network for the task for which the Bayesian network will be most suited, like diagnosis, prediction, simulation.
- Assess probabilities using information from the literature, possibly enhanced by information from people in your neighbourhood or your own judgments. Available datasets may also be used.
- Implement the resulting network using the SamIam package.

The resulting Bayesian network should consist of at least 10 vertices (variables). The following should be submitted via email to `plucas@liacs.nl`:

1. **Three- to five-page report describing:**
   - **the problem and the motivation for using Bayesian networks in modelling it**
   - **the causal model, including the variables, their values and relationships**
   - **the transformation of the causal graph to the Bayesian network**
   - **the assessment of probabilities,**
   - **the evaluation of the network in terms of the expected capabilities by:**
     - **Computing for at least 3 cases the posterior probabilities of variable(s) of interest (queries) given evidence (observations). Do the results match your intuition and/or domain knowledge? Explain your answer.**
     - **Explaining the advantages and limitations of the network.**

   Please make sure that the report has a clear structure and includes all the points mentioned above.
2. **The implementation of the Bayesian network in the SamIam package as a .net file.**

References

[1] F.V. Jensen. *Bayesian Networks and Decision Graphs*. Springer: New York, 2001.

[2] K.B. Korb and A.E. Nicholson, *Bayesian Artificial Intelligence*, Chapman & Hall: Boca Raton, 2004

[3] M. Korver and P.J.F. Lucas. Converting a rule-based expert system into a belief network. *Medical Informatics*, 1993; 18(3): 219–241. See: http://www.cs.ru.nl/∼peterl/hep2bel.pdf.

[4] P.J.F. Lucas, H. Boot, and B.G. Taal. Computer-based decision-support in the management of primary gastric non-Hodgkin lymphoma. *Methods of Information in Medicine*, 1998; 37: 206–219. See: http://www.cs.ru.nl/∼peterl/mim.pdf.

[5] P.J.F. Lucas and L.C. van der Gaag. *Principles of Expert Systems*. Addison-Wesley: Wokingham, 1991.