BayesBuilder: Introduction to Bayesian networks – Computer Practicum

Wim Wiegerinck
Foundation for Neural Networks, University of Nijmegen
`http://www.snn.ru.nl/~wimw`

About this practical

This computer practical aims to give you insight in basic modeling issues with Bayesian networks. The practical makes use of BayesBuilder, a software tool for building and inferencing Bayesian networks. BayesBuilder is free for non-commercial purposes. It can be downloaded from the web `http://www.mbfys.ru.nl/snn/nijmegen/index.php3?page=31`. A version for Linux will be available soon.
The practical is organized as follows. In section 1, the BayesBuilder software tool is explained and a short review of Bayesian networks is given. In section 2, we go a little bit deeper into the meaning of the graphical structure in Bayesian networks. Section 3 contains the main assignment for this practical. We will build a model for medical diagnosis, and we will use it to make some diagnoses.

1 Introduction to BayesBuilder and Bayesian networks

In this section we discuss the main functionalities of BayesBuilder. We do this on the basis of an example and a concise review of Bayesian networks.

1.1 BayesBuilder

To start the tool, open the BayesBuilder folder, and double-click the icon named **BayesBuilder.exe**. Wait for the license agreement, and click 'accept'.
The most important operations for BayesBuilder can be accessed via the icons on the menu bar. If the mouse is moved onto an icon, a short explanation is given. An icon is activated by a single left-mouse click.


Open network Click 'Open network', select 'icy.bbnet' and click on 'Open'. The network is now shown in the the right panel.

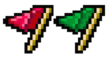
1.1.1 Inference in BayesBuilder

This network is related to a little story about reasoning with uncertainty (Jensen 2.1.1, [Jensen, 1996], see reader).

Icy roads

> Police Inspector Smith is impatiently awaiting the arrival of Mr. Holmes and Dr. Watson; they are late and Inspector Smith has another important appointment (lunch). Looking out of the window he wonders whether the roads are icy. Both are notoriously bad drivers, so if the roads are icy, they may crash.
>
> His secretary enters and tells him that Dr. Watson has had a car accident. "Watson? OK. It could be worse ... icy roads! Then Holmes has most probably crashed too. I'll go for lunch now."
>
> "Icy roads?" the secretary replies, " It is far from being that cold, and furthermore all the roads are salted." Inspector Smith is relieved. "Bad luck for Watson. Let us give Holmes ten minutes more."

A Bayesian network is a representation of a probabilistic model. In this case, the model has three variables: *IcyRoads*, *Watson*, and *Holmes*. These are the nodes in the network. The nodes have a discrete set of states, e.g., the node *Watson* has states *crash* and *save*.



Enable/disable calculating of network To compute, the flag should be toggled to green. If the flag is toggled to green you will see numbers behind the states, visually represented as the red bars in the nodes, refer to the probabilities ($\times 100$) of that node being in a certain state, i.e.,

$$\begin{aligned} P(\mathit{Watson} = \mathit{crash}) &= 0.59 \\ P(\mathit{Watson} = \mathit{save}) &= 0.41 \\ P(\mathit{IcyRoads} = \mathit{TRUE}) &= 0.70 \end{aligned}$$

etc. Note that the probabilities within a node add up to 1.

Clamp node If this icon is activated, nodes can be set to a value by left-clicks. In this way, evidence can be entered into the network. For instance, left-click on state *crashed* in the node *Watson*. The blue bar means that the node is clamped in that state. The remaining red bars (*IcyRoads*, *Holmes*), change value. (if not, check that the computation flag is set on green). They represent the conditional probability of that node given the clamped states, e.g.,

$$\begin{aligned} P(\mathit{IcyRoads} = \mathit{TRUE} | \mathit{Watson} = \mathit{crash}) &= 0.95 \\ P(\mathit{Holmes} = \mathit{crash} | \mathit{Watson} = \mathit{crash}) &= 0.76 \end{aligned}$$

By left-clicking on another state (e.g. *save*) in the node *Watson*, the node is clamped in another state. Unclamp by left-clicking on the node-name (*Watson*), or by left-clicking on the blue bar, or on its state-name.

A Bayesian network is a defined as a product of conditional probability tables. For each node, there is one table. These tables are the parameters of the network. The graphical structure of the network implies that the table of a node is conditioned on the states of the parents of that node, i.e., the nodes that point to that node. If there is no parent pointing to a node, the table consists of unconditional probabilities. So in this case the Bayesian network is defined as the product

$$P(Watson, IcyRoads, Holmes) = P(IcyRoads)P(Watson|IcyRoads)P(Holmes|IcyRoads)$$

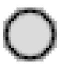
The entries of the tables can be viewed (and edited) by right-clicking on a node, and subsequently selecting **Probabilities...** in the pop-up menu. For instance, in Watson's table, the first row shows the conditional probabilities of Watson's state (*crash* or *save*) conditioned on $IcyRoads = True$. The second line shows the probabilities conditioned on $IcyRoads = False$. In other words, the table represents the following conditional probabilities

$$\begin{aligned} P(Watson = crash|IcyRoads = TRUE) &= 0.8 \\ P(Watson = save|IcyRoads = TRUE) &= 0.2 \\ P(Watson = crash|IcyRoads = FALSE) &= 0.1 \\ P(Watson = save|IcyRoads = FALSE) &= 0.9 \end{aligned}$$

Note that each row in the table add up to one. You may check these probabilities explicitely by clamping *IcyRoads* in BayesBuilder.

Excercise: Test the model, and check if the model agrees with the reasoning of inspector Smith and his secretary.

1.1.2 Changing models and creating new models



Create node A dialog pops up. The label of the node (default is 'label0', 'label1' etc.), the number of states of the node (default is 2), and the state names (comma separated - default is TRUE,FALSE) can be entered. Label and state names can be altered later if desired.

Delete the text 'label0' and enter 'Cold?'. Press OK. The node is created in the middle of the visible screen. It is drawn with a dashed line, because its probability parameters are still to be set.

Move node Left-click on a node, and drag it to another place. In this case, drag the newly created node right above the node *IcyRoads*. In a similar way, you can drag other nodes as well.

Create another node, label it 'Salted?' and drag it right to *Cold* and *IcyRoads*.

Link node There are several ways to link two nodes:

1. Left-click on the node *Cold*. Now *Cold* is purple, meaning that it is selected. Drag an arrow from *Cold* into *Salted*
2. Left-click on *IcyRoads*. Right-click on *Cold* and *Salted*.

Arrows can be removed by drawing arrows in the *same* direction. For instance, left-click *IcyRoads* and right click *Salted* removes the arrow again. Click *Salted* once again to redraw the arrow.

Note that drawing arrows in a loop (i.e. following the direction of the arrows) is not allowed.

Select node Left-click *Cold*, and right-click on the node. Select **Probabilities...**. A Node Probabilities Dialog Box appears in which you can inspect and edit the probability table of this node.

The Node Probabilities Dialog Box You may enter any positive value (you may prefer numbers between 0 and 100 instead of 0 and 1). The numbers will be normalized automatically after you close the dialog box. You can clear the table by the left button (small white squares), and you can normalize the table by the right button (calculator) in the tool-bar of the dialog box. Uniform and random values can be entered via the **Table** menu. Clear the table, enter the values 0.95 in the cell labeled 'TRUE' and 0.05 in the cell labeled 'FALSE'. Make sure that you press the `Enter` key, or activate another text-cell using the mouse after you entered the value. Press OK. Open the Node Probabilities Dialog Box for *Salted* Now you have to enter conditional probabilities for *Salted* conditioned on the states of *Cold*. Enter the table

Cold?	TRUE	FALSE
TRUE	0.5	0.5
FALSE	0.1	0.9

Finally, enter the probabilities for *IcyRoads*.

Salted?	Cold?	TRUE	FALSE
TRUE	TRUE	0.5	0.5
TRUE	FALSE	0.001	0.999
FALSE	TRUE	0.9	0.1
FALSE	FALSE	0.01	0.99

(If *Salted* and *Cold* are interchanged, the probabilities should be interchanged accordingly).

Warning: make sure that you press the `Enter` *key after the last entered probability, otherwise this probability will not be stored.*

After entering the probabilities, the network is again fully specified, as indicated by the red bars. Now you may test the model, again, e.g., check if the model agrees with the reasoning of inspector Smith and his secretary. Probably you will notice that the conclusions drawn in this network will not differ significantly from the original icy.bbnet network.

The Node Properties Dialog Box Selection of **Properties...** in the pop-up menu of a node allows you to change the label of a node and to change the names of the states. You also may enter some text in the 'descr.' and 'info' text fields.

Saving Use the **Save as ...** option under **File** in the menu bar to save the network under a new name. (Use **Save** if no different name is required.)

Other useful operations

File | New to create a new network.

Edit | Undo / Redo to undo or redo certain operations. Also available on the tool-bar

Network | Delete Selection select one or more nodes by activating **Select**, and clicking one or more nodes while pressing the `Ctrl` key. **Delete Selection** (or `Ctrl Delete`) deletes the nodes (you can undo this operation).

Network | Clear unclamps all nodes.

Module | Edit Name Edit the name in the upper left corner of the network panel.

2 Graphical Structure in Bayesian networks

2.1 Direction of the arrows

The graphical structure (i.e., connectivity of the nodes and directions of the arrows) of a Bayesian network fully specifies the conditional independency assumptions of the probabilistic model A structure without independency assumptions is fully connected.

2.1.1 Two nodes

Two-node networks that are connected are fully connected. Therefore the direction of the arrow does not imply any assumption. In other words, the two network

(a) Fire → smoke

(b) Smoke → fire

Figure 1: Two equivalent network structures

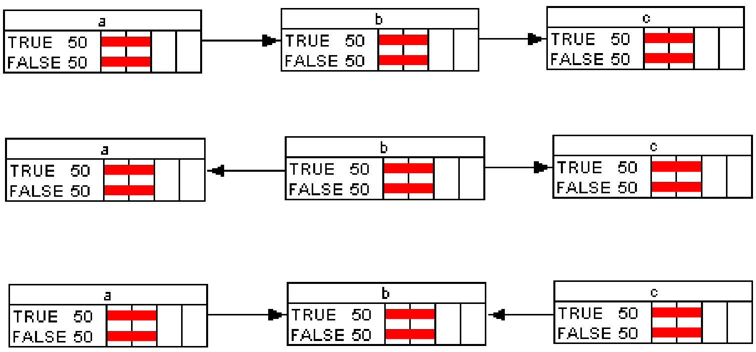
structures in figure 1 are equivalent. In algebraic form, the two network structures are clearly equivalent, since

$$P(smoke, fire) \quad = \quad P(fire)P(smoke|fire) = P(fire|smoke)P(smoke)$$

You may check this empirically by building the first model with some reasonable probability tables in BayesBuilder, and then try to make the second model such that all unconditional and conditional probabilities are the same. You can get the required conditional probabilities for the tables in the second model by appropriate clamping in the first model.

2.1.2 Three nodes

For three nodes, it is slightly more complicated.



These three networks represent

$$\begin{aligned} P(a,b,c) &= P(a)P(b|a)P(c|b) && \text{(serial connections)} \\ P(a,b,c) &= P(a|b)P(b)P(c|b) && \text{(diverging connections)} \\ P(a,b,c) &= P(a)P(b|a,c)P(b) && \text{(converging connections)} \end{aligned}$$

The serial and the one with diverging connections are equivalent: both are of the form $P(a,b,c) = P(c|b)P(a,b)$. The converging model, however, is not equivalent to these.

A typical example of a network with diverging arrows is the icy-roads network. A property of this network is that Watson's state and Holmes' state are dependent (via IcyRoad). If the state of IcyRoad is clamped, the states of Watson and Holmes become independent. You may explicitely check this in the network. In the serial and diverging network shown above a and c are independent given b.

A typical example of a network with converging arrows is the wet-grass network (wetgrass.bbnet). Open this network. The network is related to the following story (Jensen 2.1.2,[Jensen, 1996], see reader).

Wet grass

> Mr. Holmes now lives in Los Angeles. One morning when Holmes leaves his house, he realizes that his grass is wet. Is it due to rain, or has he forgotten to turn off the sprinkler? His belief in both events increases. Next he observes his neighbors grass, which is also wet. The only explanation for his neighbor's wet grass is that it has been raining. This also explains Holmes' wet grass, and there is no reason to believe that he forgot to turn off the sprinkler.

Check if the network agrees with Holmes' line of reasoning. The phenomenon that sprinkler becomes less likely due to other evidence explaining wetgrass is called 'explaining away'. What happens with the probability of sprinkler if it had not rained? Also compare the dependency of rain on sprinkler and vice versa for both the case that wet grass is unclamped with the case that it is clamped.
In this network, rain and sprinkler are marginally independent. Once the state of the grass is known, the rain and sprinkler become dependent.

2.1.3 Reversing the arrows?

Consider what would be the consequences in the model if wet grass had been modeled with diverging arrows. (One could say: (1) wet grass indicates rain, therefore there should be an arrow to rain (2) wet grass indicates sprinkler, therefore there should be an arrow to sprinkler). Consider how rain would depend a priori on sprinkler, and what would happen with the probability of sprinkler if you know that the grass is wet and that it has rained.

What are the consequences if you would model the icy roads networks with converging nodes. (One could say, car crashes are indications of icy roads). Suppose inspector Smith would follow the reversed model. If he would learn about Watson's crash, his belief in icy roads should increase, but this would not affect his belief in Holmes' car accident. On the other hand, if he only learned about an increase in the probability of icy roads, but not about Watson's car accident, his belief in Holmes' car accident would increase. If he then hears that Watson had a car accident, he would be relieved, since this would explain away the icy roads and his belief in Holmes car accident decreases again.

2.2 Bayesian networks and causal modeling

In general it is a good rule of the thumb is to construct a Bayesian network from cause to effect. Start with nodes that represent independent root causes, then model the nodes they influence, and so on until we end at the leaves, i.e. the nodes that have no direct influence on other nodes. For this procedure, it is often useful to have a 'story' in mind.

Sometimes this procedure fails, because it is to difficult to tell what is cause and what is effect. Is someone's behavior an effect of his environment, or is the environment a reaction on his behavior? In such a case, just avoid the philosophical dispute, and return to the basics of Bayesian networks: a Bayesian network is not a model for causal relations, but a joint probability model. The structure of the network represents the conditional independence assumptions in the model and nothing else.

In practice it is often difficult to decide whether two nodes are really (conditionally) independent. Usually, this is a matter of simplifying model assumptions. In the case of Icy roads, one could easily argue that Watson and Holmes are still dependent, even if we know that the roads are not icy, e.g., due to other conditions (traffic, other weather conditions, a mystery distracting their minds, etc) that increases the risks of an accident for both drivers simultaneously. In the true world, all nodes should be connected. In practice, reasonable (approximate) assumptions are needed to make the model simple enough to handle, but still powerful enough for practical usage...

3 Main assignment: a model for medical diagnosis

In this section we will build a model for medical diagnosis based on the information in the following subsection.

3.1 Common knowledge and statistical information

First, consider following piece of qualitative 'knowledge' (This is adapted Lauritzen's chest clinic [Lauritzen and Spiegelhalter, 1988]).

> The symptom *shortness of breath* may be due to the diseases *pneumonia*, *lung cancer*, and/or *bronchitis*. Patients with *pneumonia*, and/or *bronchitis* often have a very nasty *wet coughing*. *Pneumonia*, and/or *lung cancer* are often accompanied by a heavy *chest pain*. *Pneumonia* is often causing a severe *fever*, but this may also be caused by a *common cold*. However, a *common cold* is often recognized by a *runny nose*. Sometimes, *wet coughing*, *chest pain*, and/or *shortness of breath* occurs unexplained, or are due to another cause, without any

of these diseases being present. Sometimes diseases co-occur. A *weakened immune-system*[1] increases the probability of getting an *pneumonia*. Also, *lung cancer* increases this probability. *Smoking* is a serious risk factor for *bronchitis* and for *lung cancer*.

For each of the diseases (*pneumonia*, *lung cancer*, *bronchitis*, and *common cold*), gold standard tests exists. But those tests are not considered in the model.

Next, consider the quantitative figures of the population visiting the chest clinic:

- Statistics:

Finding	percentage
weakened immune-system	5
smoking	29
pneumonia	2
lung cancer	4
bronchitis	9
common cold	35
fever	9
wet coughing	19
chest pain	14
shortness of breath	18

- Correlations:
  - 80 % of the population that had been diagnosed with *bronchitis* suffered from *shortness of breath*.
  - 50 % of the population without bronchitis, but with *pneumonia*, and/or *lung cancer* suffered from *shortness of breath*.
  - 10 % of the population for which the mentioned diseases could not be established suffered from *shortness of breath*.
  - 90 % of the population with *pneumonia*, and/or *lung cancer* suffered from *chest pain*.
  - 10 % of the population for which these diseases could not be established suffered from *chest pain*.
  - 90 % of the population with *pneumonia*, and/or *bronchitis* suffered from *wet coughing*.
  - 10 % of the population for which these diseases could not be established suffered from *wet coughing*.

[1] For instance, homeless people, or HIV infected.

  - Of the population with a *common cold*, about 20 % has a high *fever*. In 95 % a *common cold* is accompanied by a *runny nose*. Only 1 %, of the people without *common cold* had a *runny nose*.
  - 90 % of the population with *pneumonia*, has a high *fever*

- Risk factors:
  - 10 % of the smokers has lung-cancer, (and only 1 % of the non-smokers).
  - 30 % of the smokers has bronchitis, (and only 1 % of the non smokers).
  - 30 % of the people with a *weakened immune-system*, and about 5% of the people with a *lung cancer* has been diagnosed with *pneumonia*. In only 0.1% of the population without these risk factors *pneumonia* has been diagnosed.

3.2 Assignment

3.2.1 Build a medical expert system for the chest clinic

Build a medical expert system based on the information provided in the previous subsection.

- *Hint (1):* First find out, which are the variables (these are printed in *italics* in the story with 'qualitative knowledge'). Then try to figure out a sensible graphical structure. Use causal relations! You can derive these from the 'qualitative knowledge' and some common sense. Finally, figure out the required conditional probabilities, using the quantitative information above.
- *Hint (2):* The quantitative information is not sufficient to fill in the table values. You have to use common sense and you should be a bit practical in your approach.
- *Hint (3):* It is recommended that you have your model checked by the practicum assistant before you proceed to the application of your model.

3.2.2 Apply your system

1) In a medical text book, the following diagnostic guidelines and comments are given. Check if your system agrees with these, and explain.

- In case of high *fever* in absence of a *runny nose*, one should consider *pneumonia*. Why?
- *Lung cancer* is often found in patients with *chest pain*, *shortness of breath*, no *fever*, and usually no *wet coughing*.

- *Bronchitis* and *lung cancer* are often accompanied, e.g patients with *bronchitis* often develop a *lung cancer* or vice versa. However, these diseases have no known causal relation, i.e., bronchitis is not a cause of lung cancer, and lung cancer is not a cause of bronchitis. Can you explain why these diseases are often accompanied?

2) Apply your system to the following cases:

- Mr. Appelflap calls. Mr. Appelflap is a book keeper in a middle sized entreprise. He lives with his wife and two children in a nice little house [2] in Almere. I.e, you may assume that he is not HIV infected. Mr. Appleflap complains about high fever and a nasty wet cough (although he is a non-smoker). In addition, he sounds rather nasal. What is your primary diagnosis? Why?

- The salvation army calls. An unknown person (looking not very well) has arrived in their shelter for homeless people. This person has high fever, and a nasty wet cough. What is your primary diagnosis? Why? If this person would have a runny nose, would that change your advice?

- A colleague calls. One of his patients suffers from a recurrent pneumonia. This patient is a heavy smoker but otherwise leads a 'normal', healthy live. What is your advice?

References

[Jensen, 1996] Jensen, F. (1996). *An introduction to Bayesian networks.* UCL Press.

[Lauritzen and Spiegelhalter, 1988] Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilties on graphical structures and their application to expert systems (with discussion). *J. Royal Statistical Society Series B*, 50:157–224.

[2] rijtjeshuis