Study Manual

Probabilistic Reasoning with BNs

2019 – 2020

Silja Renooij

August 2019

General information

This study manual was designed to help guide your self studies. As such, it does not include material that is not already present in the syllabus or discussed in class. Hopefully, it will help you take a deeper dive into the subjects and enhance your understanding.

This manual enumerates some details on the knowledge and abilities students enrolling in this course are expected to learn. In addition, this manual lists for each chapter from the syllabus a number of questions that can help to guide your self-study. In the syllabus itself, a number of exercises and answers to some of them (indicated by a *) are given. The exercises enable you to practice what you have learned and give an idea of what you should be able to do if you fully understand the material. Most of the exercises are former exam questions. Examples of previous exams are available through the course web-site. The questions and exercises should indicate how you are doing thus far.

Examination

The INFOPROB course is a 7.5 ECTS course, which means that an average student is expected to require 210 hours of work to complete and pass the course.

The course is graded based on a number of practical assignments and a written exam. For the exam, you are expected to have knowledge of, insight in and, most importantly, understanding of the subjects presented in the syllabus and (sometimes more detailed) in the course slides. This also means that you have practised with applying different procedures and algorithms and can correctly do so in little time. In general, I expect at least the following:

- you know all preliminaries by heart, as well as any formula that can be easily reconstructed from understanding the idea behind it;
- you understand the origin of the components of all parameters/messages from Pearl's algorithm (Chapter 4), but you need not know the exact way they are computed (the lemmas);
- you understand the entropy formula (Chapter 5) and its effects, but you are not required to reproduce it;
- you know what different symbols and notations mean, how different concepts are defined, and you are able to apply the different methods and algorithms discussed; (NB make sure you practice with Pearl's algorithm; this will save you at lot of time during the exam!!)
- you have enough understanding of all subjects treated to be able to identify advantages and drawbacks of different concepts and algorithms, and to use these insights in generating or understanding variations on what was treated during the course;
- ...

More specifically:

- Chapter 2 + course slides: you know all definitions, propositions and theorems by heart;
- Chapter 3 + course slides: you know all definitions, theorem, corollaries and lemmas by heart;
- Chapter 4 + course slides:
  - you know by heart: all definitions, corollaries, proposition 4.1.3, lemmas 4.2.1, 4.2.8, all algorithms (test your inference answers with one of the Bayesian network software packages (see links on the course website))
  - you will be given during the exam: lemmas 4.2.4, 4.2.5, 4.2.6, 4.2.7, 4.2.13 (recall that the formulas for trees are just a special case of those for singly connected graphs)
- Chapter 5 + course slides
  - you know by heart: all definitions, except 5.2.2, 5.2.3, all lemmas, corollaries and algorithms, all formulas for the (leaky) noisy-or gate
  - if necessary, you will be given during the exam: definition 5.2.2, 5.2.3
- Chapter 6 + course slides: you know all concepts and formulas in 6.1 and 6.2 by heart; you understand the global ideas in 6.3
- Material on selected topics (if any): you should understand the global ideas presented, and their relation to the standard course material. More details will be supplied on the course website (literature tab).

An important note on applying procedures and algorithms: there is a difference between solving a problem and applying an algorithm that solves the problem. If an exercise or exam asks you to solve a problem, you can use whatever method you think is most suitable to do the job. If, however, you are required to solve a problem using a certain algorithm, then you are basically asked to illustrate how the algorithm can be used to solve the problem and you should execute the steps prescribed by the algorithm yourself. For example, if you are asked to compute a certain probability from a set of given probability, you may use whatever rules from probability theory you can apply. If you are asked to compute a probability using Pearl's algorithm (see Chapter 4), then you should perform exactly those computations prescribed by the algorithm, even if the example is such that there are more efficient ways of solving the problem by hand! In illustrating Pearl's algorithm, it is sufficient, though, to do only those computations that are required to answer the question at hand, instead of computing each and every parameter for all variables. Basically, if you are asked to apply an algorithm, I am testing whether or not you understand how the algorithm works, not if you are able to solve the problem for which the algorithm is designed.

Chapter 1

The first chapter gives some historical background about the use of probability theory and other uncertainty formalisms for reasons of decision support. It briefly motivates the emergence of Bayesian networks and the reason why Bayesian network applications historically have often concerned the medical domain.

Questions

- What is mutually exclusive?
- What is collectively exhaustive?
- Which assumption is made in a naive Bayes approach for reasons of *time* complexity, what is the reduction achieved and why?
- Which assumption is made in a naive Bayes approach for reasons of *space* complexity, what is the reduction achieved and why?
- Why is a naive (or idiot) Bayes approach naive? Is it still used?
- What is the relation between a naive Bayes approach and GLADYS?

Chapter 2

The second chapter refreshes the necessary concepts from graph- and probability theory that play a central role in the Bayesian network framework. These concepts have already been discussed in other courses and can be found in any textbook on graph theory and probability theory. The chapter also introduces the notation that is used throughout the syllabus and which may differ from what you encountered previously.

Two important things to note here are the following:

- In this course we often address paths in graphs; unless stated otherwise, these paths are assumed to be *simple* paths.
- A lot of formulas in the course material contain a summation of the following form:

$$\sum_{c_V} f(c_V)$$

  for some expression $f(c_V)$ depending on c_V. Note that this summation is a summation over the set of *configurations* c_V of a set V, not over the elements of set V itself. If this set V is empty, this therefore does not mean the summation is undefined! Rather, it means the summation is over the single element $c_V = c_\emptyset = \mathsf{T}$ (true). In that case the summation reduces to the single term $f(\mathsf{T})$.

Questions

- Is there a limit to the number of times a certain vertex may be included in a path?
- What is the difference between a walk, a path, and a simple path?
- What is the difference between a path and a chain in a directed graph?
- What is the difference between a loop and a cycle in a directed graph?
- What is the difference between a tree, a singly connected digraph and a multiply connected digraph?
- When can we use the operators $\cap$ and $\cup$?
- When can we use the operators $\wedge$ and $\vee$?
- What is the difference between the proposition True and a value *true* for a variable?
- What is the difference between a *joint* probability distribution, a *conditional* probability distribution, and a *marginal* probability distribution?
- What is the difference between a *value assignment*, a *configuration* and a *configuration template*?
- Why is it the case that for two sets of variables X and Y $C_{X\cup Y} = C_X \wedge C_Y$ is the only correct way of expressing the configuration template over both sets X and Y in terms of the separate configuration templates?
- What is the relation between the *marginalisation* and the *conditioning* property?
- What is the difference between independence of *propositions* and independence of *variables*?
- Suppose a (not necessarily binary) random variable V can take on the values $v_1, \ldots, v_n$. Why does the definition of probability distribution imply that $\sum_{i=1}^{n} \Pr(v_i) = 1$?
- What is the difference between (in)dependence and conditional (in)dependence?

"Refreshing" exercises

The following exercises were designed for a bachelor course on statistics and can be used to refresh your memory and intuition.

1 Joost has been suffering from fits of sneezing. He decides to see his general practitioner (GP) and have himself tested for hay fever. The GP tells Joost that about 1 in 10 people in the Netherlands get hay fever.

   a. The **prior** probability $\Pr(h)$ that Joost has hay fever is the probability of hay fever in case we know nothing about Joost. What is this probability?

   The GP performs a skin-test; this test turns up positive, i.e. the test indicates that Joost indeed has a hay fever.

   b. Can you now give an estimation of the probability that Joost does *not* suffer from hay fever?

Joost asks his GP how reliable the skin-test is. The GP indicates that the test gives a correct prediction for 70% of all people with hayfever, and for 90% of all people without hayfever.

c. Which conditional probabilities are described here?

d. Without calculating, what do you think the probability is of Joost *not* having a hay fever, despite the positive test result?:
between 0 and 0.2,
between 0.2 and 0.4,
between 0.4 and 0.6,
between 0.6 and 0.8, or
between 0.8 and 1

e. Use the *conditioning rule* to compute the prior probability of a positive test result. NB This rule is also referred to as the rule of **total probability**.

The reliability of a test is given in terms of the probability of obtaining a certain test result given that the patient tested has, or does not have, respectively, the disease tested for. In designing a treatment for a patient, however, it is important to know the opposite: what is the probability of the disease given the test result.

f. Which rule can be used for 'reversing' conditional probabilities?

g. Compute the probability of Joost not having a hay fever, despite the positive test result. Does the answer coincide with your estimate for part d.?

Intermezzo Two equivalent definitions of independence of two propositions x and y exist:

(a) x and y are *independent* if $\Pr(x \wedge y) = \Pr(x) \cdot \Pr(y)$;

(b) x and y are *independent* if $\Pr(x \mid y) = \Pr(x)$.

In addition, two equivalent definitions exist of *conditional* independence of two propositions $x$and y given a third proposition z:

(a) x and y are *independent given* z if

$$\Pr(x \wedge y \mid z) = \Pr(x \mid z) \cdot \Pr(y \mid z);$$

(b) x and y are *independent given* z if $\Pr(x \mid y \wedge z) = \Pr(x \mid z)$.

2 Suppose that the GP from the previous exercise also performs a blood test on Joost.

a. Are the results from the skin-test and the blood test dependent or independent?

b. Suppose you are clairvoyant (all-seeing) and know for sure that Joost has a hay fever. Are the results from the skin-test and the blood test dependent or independent, given this knowledge?

Suppose that you have available for both tests the probabilities of a positive result given the presence or absence, respectively, of hay fever. You want to use Bayes' rule to compute the probability of Joost having a hay fever given results of both tests.

c. Write down Bayes' rule for this probability.

d. Which assumptions with respect to (conditional) (in)dependence do you have to make in order to compute this probability from the information about the reliability of the tests?

3 Consider a population of a 1000 males and females. Consider the following propositions: $m =$ {is male}, $f =$ {is female}, $a =$ {wears make-up} en $b =$ {doesn't wear make-up}.

a. How many random variables do we consider here, and what are their values?

b. Use your intuition to determine whether the propositions m and b are dependent or independent.

Suppose that the following frequency table is given for the population under consideration:

	m	f
a	50	300
b	350	300

c. Use the definition of independence to determine whether the propositions m and b are independent.

d. Suppose that we only consider the subset of grown-ups from the above population. Use your intuition to determine whether, given the proposition {is grown-up}, the propositions m and b are dependent or independent.

e. Suppose that we only consider the subset of children from the above population. Use your intuition to determine whether, given the proposition {is child}, the propositions m and b are dependent or independent.

Suppose we divide the population into children c and grown-ups g:

	m		f	
	g	c	g	k
a	50	0	300	0
b	150	200	50	250

f. Use the definition of independence to determine whether, given the proposition k, the propositions m and b are independent.

4 Consider a population of 1000 individuals. Consider the following propositions: h = {dog owner}, k = {no dog owner}, s = {wears glasses} en g = {does not wear glasses}.

a. How many random variables do we consider here, and what are their values?

b. Use your intuition to determine whether the propositions h and s are dependent or independent.

Suppose that the following frequency table is given for the population under consideration:

	s	g
h	50	200
k	150	600

c. Use the definition of independence to determine whether the propositions h and s are independent.

d. Suppose we consider only the women in the above population. Use your intuition to determine whether, given the proposition {is female}, the propositions h and s are dependent or independent.

e. Suppose we consider only the severely visual disabled from the population. Use your intuition to determine whether, given the proposition {is 'blind'}, the propositions h and s are dependent or independent.

Suppose we divide the population into 'blind' b and not 'blind' z:

	s		g	
	z	b	z	b
h	30	20	195	5
k	140	10	595	5

e. Use the definition of independence to determine whether, given the proposition b, the propositions s and h are independent.

5 Consider the propositions a and $b_1, \ldots, b_n$, and the *marginalisation rule*:

$$\Pr(a) = \sum_{i=1}^{n} \Pr(a \wedge b_i)$$

Invent a proposition a and at least three proposition b_i to motivate the existence of the marginalisation rule.

6 Suppose a course is graded based on n assignments; each assignment T_i concerns the material discussed in week i, $i = 1, \ldots, n$. Let t_i be the proposition {student passes assignment T_i}.

a. Suppose the course requirements are such that you can only pass the assignment in week i, if you have passed assignments T_1 through T_{i-1}. Give a formula for the probability $P(t_1 \wedge t_2 \wedge t_3 \wedge t_4 \wedge t_5 \wedge t_6)$ that someone passes the first 6 assignments.

The formula that you constructed in part a. (if correct) is known as the *chain rule*.

b. Now suppose that the course consists of various subjects, each covering two weeks. The material for the assignments in uneven weeks is new, the material for the assignments in the even weeks is related to the material of the previous week. Rewrite the formula in part a. to reflect these properties.

Chapter 3

This chapter formalises two types of independence relation. The first, $I_{\Pr}$, is the type of relation that can be captured by a probability distribution. These independence relations form a proper subset of a more general type of independence relation I that abstracts away from probability distributions. The chapter also discusses different representations of independence relations, most notably (in)directed graphs. An important notion introduced in this chapter is the concept of d-separation.

Independence relations and their representation are still an area of ongoing research, see for example the work by Milan Studený from Prague. This chapter, however, describes the full body of research as far as required for our discussion of Bayesian networks in the subsequent chapters.

Questions

- Intuitively, what is the difference between independence relations $I_{\Pr}$ and I?
- What is the difference between the properties satisfied by $I_{\Pr}$ compared to those satisfied by I?
- What are possible encodings of an independence relation?
- Why does an I-map represent independencies and a D-map dependencies?
- Why is it unfortunate that not every independence relation has an (un)directed P-map?
- Why is an independence relation for which a P-map exists termed *graph-isomorphic*?
- Does a graph-isomorphic independence relation have a *unique* P-map?
- Does the (d-)separation criterion require a graph to be acyclic? Why/why not?

- Why does the following statement hold for both directed and undirected graphs: If a set Z blocks a path, then $Z \cup Y$ blocks the path for any Y?
- Why does the following statement for directed graphs only hold if $Z \neq \emptyset$ and at least one $Z_i \in Z$ is on the chain under consideration: If a set Z blocks a chain, then $Z \cup Y$ blocks the chain for any Y?
- Why does the following statement hold for undirected graphs: If a set Z separates two other sets, then $Z \cup Y$ separates the sets for any Y?
- Why does the following statement *not* hold for directed graphs: If a set Z d-separates two other sets, then $Z \cup Y$ d-separates the sets for any Y?
- How do we transform a directed I-map into an undirected I-map? Is minimality inherited?

Chapter 4

This chapter defines the concept of Bayesian network and reconstructs the algorithm for exact probabilistic inference as designed by Judea Pearl; for ease of exposition only binary variables are considered. Other, and often more efficient, algorithms for exact inference exist such as *jointree propagation* (S.L. Lauritzen, D.J. Spiegelhalter, F.V. Jensen, P.P Shenoy, G. Shafer) and *variable elimination* (R. Dechter, G.F. Cooper) methods. The reason for discussing Pearl's algorithm is that it explicitly exploits the graphical structure of the network without transforming it, and therefore seems the one easiest to explain and comprehend.

Current research focuses on finding still more efficient algorithms, both for exact and approximate inference, and also for dealing with networks that include continuous variables. Also, now and again, researchers look into better algorithms for loop cutset conditioning (A. Becker, D. Geiger), the extension to Pearl's algorithm that is required to do inference in multiply connected graphs.

To gain more insight in inference and the effects of loop cutset conditioning try one of the free software packages available through the course website.

Questions

The network as graphical model:

- Why is a Bayesian network often called a *belief network*, or *causal network*?
- What is the independence relation I modelled by a Bayesian network?
- Why have people chosen the graph of a Bayesian network to represent a (minimal) I-map of the independence relation as opposed to a D-map?
- Why have people chosen a directed graph to capture the independence relation for a Bayesian network?

The probability distribution:

- Why is the digraph of a Bayesian network assumed to be acyclic?
- What is the difference between an assessment function for a vertex and a probability distribution for that vertex?
- Which piece(s) of information is/are necessary to determine the number of assessments functions required for specifying a Bayesian network?
- Which pieces of information are necessary to determine the number of (conditional) probabilities required for specifying all assessments functions?
- How does the joint probability distribution Pr *defined* by a Bayesian network exploit the fact that the digraph of the network is an I-map of the independence relation of Pr?
- How do constraints for space requirements for Bayesian networks compare to the naive Bayes approach?

Probabilistic Inference – general:

- How do constraints for time requirements for Bayesian networks compare to the naive Bayes approach?
- What is the difference between a prior and a posterior probability?
- Is there actually a difference between $\Pr(v_i \mid v_j)$ and $\Pr^{v_j}(v_i)$?
- What is the difference between $\Pr(v_i)$ and $\Pr^{v_i}(v_i)$?
- How can a set of instantiated or observed vertices be related to the *blocking set* introduced in Chapter 3?
- What is the relation between *configuration*, *instantiation*, and *partial configuration*?
- What is the computational time complexity of probabilistic inference?

Pearl's algorithm – definitions:

- Why does the definition of V_i^- differ across different types of graph?
- How do the messages of Pearl's algorithm exploit the fact that the digraph is an I-map of the independence relation of Pr?
- Why do the formulas for computing messages differ across different types of graph?
- Which lemma represents the first step of Pearl's algorithm?
- What is the difference between a causal and a diagnostic parameter?
- What are the differences and similarities between causal/diagnostic parameters and probabilities?
- Why do the *compound* causal parameters for a vertex sum to 1? Why do the causal parameters *sent* by a vertex sum to 1?

- Why do the compound diagnostic parameters for a vertex in general *not* sum to 1? Why do the diagnostic parameters *sent* by a vertex in general *not* sum to 1?
- Why is it the case that if the compound diagnostic parameter of a vertex equals 1, that all diagnostic parameters sent by that vertex are equivalent (in a directed tree: equal to 1)?
- When is a vertex ready to send a causal parameter?
- When is a vertex ready to send a diagnostic parameter?
- When is a vertex ready to calculate its (posterior) probability?
- To calculate prior probabilities, do you require causal parameters, diagnostic parameters, or both?
- To calculate posterior probabilities, do you require causal parameters, diagnostic parameters, or both?
- Why do the formulas in Lemmas 4.2.5 and 4.2.7 and similar formulas for singly connected digraphs assume V_i to be an uninstantiated vertex?
- Why do observations have to be entered by means of *dummy nodes*?
- Why does the *dummy node* approach enable us to use lemmas 4.2.5, 4.2.7 etc. for instantiated vertices as well?
- Why can Pearl's algorithm lead to incorrect results when applied to multiply connected graphs?

Pearl's algorithm – normalisation:

- What is the use of a normalisation constant α?
- Why does the normalisation constant α differ across messages?
- Is normalisation always used for messages that sum to 1, or could they sum to a different constant? (Compare α in the formula for separate diagnostic parameters in singly connected graphs to the other α's).
- Why can normalisation be postponed to the final step of Pearl's algorithm?
- When does it come in handy to not postpone normalisation?
- Why is normalisation not actually required when computing prior probabilities? (Can be used as correctness check)

Loop Cutsets:

- Why does Pearl's algorithm as presented in 4.2.1 and 4.2.2 not work in multiply connected graphs?
- What is the idea behind the method of loop cutset conditioning?

- Which of the following statements is correct (the difference is rather subtle): "Loop cutset conditioning is a method for probabilistic inference in multiply connected networks that uses Pearl's algorithm for computing some of the probabilities is requires." or, "Loop cutset conditioning is a method that is used within Pearl's algorithm in order to enable its correct application to multiply connected networks."?
- Consider a Bayesian network with loops in its digraph G; can the loop cutset for G be the empty set $\emptyset$?
- Given the definition of a loop cutset, can a vertex with two incoming arcs in G be included in a loop cutset for G?
- For a graph G, will the Suermondt & Cooper algorithm return a loop cutset that includes a vertex with two or more incoming arcs in G?
- Why does every loop in a Bayesian network's digraph have at least one vertex with at most one incoming arc?
- Is a loop cutset a property of a graph or a property of a Bayesian network?
- Does a loop cutset change after entering evidence in a Bayesian network?
- Does the probability distribution over configurations of the loop cutset change after entering evidence into a Bayesian network?
- Why can the prior probability distribution over configurations of the loop cutset *not* be computed using Pearl's algorithm, but has to be computed by marginalisation from the joint distribution instead?
- In loop cutset conditioning, in general, which probabilities can and which probabilities cannot be computed using Pearl's algorithm (as a "black box")?
- What is the use of a *global supervisor*?
- What is the first step in performing loop cutset conditioning (not: in finding a loop cutset!)?
- Why is loop cutset conditioning computationally expensive?
- Is finding a loop cutset computationally expensive?
- Is a minimal loop cutset always optimal? Is an optimal loop cutset always minimal?
- Why should vertices with a degree of one or less not be included in a loop cutset?
- Why are vertices with an *in*degree of at most one selected as candidates for the loop cutset by the heuristic Suermondt & Cooper algorithm? What are possible effects of this choice?
- Why does the heuristic Suermondt & Cooper algorithm for finding a loop cutset choose to add a candidate with highest degree to the loop cutset? What if there are several candidates with the highest degree?
- What are properties of the graph and the loop cutset before and after performing the algorithm for finding a loop cutset?

Chapter 5

This chapter and the next one differ from previous chapters in the sense that the latter describe more or less finished research, whereas the following chapters discuss topics that are subject of ongoing research[1]. As a result, the chapters tell a number of short stories with lots of open endings. We provide the basics of various algorithms and methodologies that still underlie, and are necessary for understanding, state of the art results. The assumption that we consider binary variables only is now lifted, although employed in some specific situations.

This chapter discusses the construction of a Bayesian network: how do we get the graphical structure (by hand, or automatically?) and where do the probabilities come from (experts, data, ...?). Try and construct a (small) Bayesian network yourself using one of the numerous software packages or online tools developed for constructing and reasoning with Bayesian networks (see the course website)

Remark: in this chapter you will find formulas including a term of the form $a \cdot \log b$, where a and/or b can be zero. A standard convention is that $0 \log 0 = 0$.

Questions

Construction in general

- What are typical trade-offs made when constructing a Bayesian network for a real application domain?
- What is the difference between domain-variables and Bayesian network variables?
- What is the difference between single-valued variable, multi-valued variables and random variables?
- Given the definition of a Bayesian network in Chapter 4, why is it, in general, not possible to allow continuous random variables?
- What are advantages and disadvantages of the different ways of modelling a multi-valued domain variable in a Bayesian network?

Construction by hand

- What is the problem with using the notion of causality for constructing the digraph?
- Why is it important to distinguish between direct and indirect relations?
- Why do digraphs which are constructed by hand often contain cycles at some stage of the construction?
- Why does the assumption of a disjunctive interaction simplify probability assessment? Is the assumption realistic?

[1] As a matter of fact, the Decision Support Systems group of our Department does research into quite a number of subjects discussed in this chapter (see the group website).

- Why is a noisy-or gate called noisy? Why is its leaky version called leaky?
- This chapter formulates the noisy-or gate for binary variables. Is it easy to generalise to non-binary variables?
- Why does the inhibitor probability in the leaky noisy-or gate differ semantically from the inhibitor probability in the noisy-or gate?
- Why is it difficult to assess probabilities from literature?
- Why is it difficult to use domain experts for assessing probabilities?
- What is a typical trade-off when eliciting probabilities from experts?
- What is the use of sensitivity analysis?

Automated construction

- Are the assumptions required of a database in order to use it for learning a network realistic?
- What are possible solutions to and the effects of handling *missing values* in a database?
- $N(c_X)$ gives the number of cases in a database for which variable X has the given configuration. Why can we correctly assume that $N(c_\varnothing) = N$?
- Is frequency counting an efficient way of probability estimation?
- What is the difference between a conditional independence learning algorithm and a metric learning algorithm?
- Does the order of variables used for conditional independence learning affect the resulting graph?
- Why are a quality measure and a search heuristic necessary ingredients for a metric learning algorithm?
- What is the function of the different ingredients of the MDL quality measure?
- Why does the MDL measure as stated assume variables to be binary? Can this restriction be lifted?
- What are the ranges of the different terms in the MDL measure?
- Why is the $\log P$ term a constant if you assume a uniform prior distribution over possible digraphs?
- You could say "Entropy $\equiv$ Chaos $\equiv$ Uncertainty". Why does lower entropy go hand in hand with denser digraphs?
- Why do denser digraphs go hand in hand with less reliable probability estimates? What solution to this problem is present in the MDL measure?
- What are the benefits and drawbacks of using a search *heuristic*?
- When employing the search heuristic as described in Chapter 5 it is sufficient to consider only *gain* in quality. Why?

- For the search heuristic described: why does a difference in *vertex quality* for a single vertex correspond to a difference in *graph quality*?
- When the search heuristic adds an arc (V_i, V_j), why does the quality of V_j change and not the quality of V_i?
- If adding a certain arc *decreases* the quality of the graph, will that arc *ever* be added?
- If adding a certain arc *increases* the quality of the graph, will that arc *always* be added?
- Does the choice of an arc to add affect the set of arcs that may be subsequently added?
- Does the procedure of adding arcs that give the largest increase in quality result in a graph with maximum quality?
- Is the MDL measure a suitable measure?
- When should you learn networks of restricted topology?
- What are possible effects of learning a network from a database that does not obey the required assumptions?

Chapter 6

This chapter tries to bridge the gap between construction and actual use of a Bayesian network application. It first discusses possible methods for evaluating the behaviour of your network; then some example applications where a Bayesian network is used as a component in a decision support process are discussed[2].

Questions

Sensitivity analysis:

- What is the use of sensitivity analysis?
- Compare the following terms: 'sensitivity', 'robustness', and 'correctness'.
- Is it computationally feasible to perform a sensitivity analysis?
- Why can no parameter probability have a non-linear influence on a prior probability of interest?
- Is the sensitivity set a property of a Bayesian network's digraph?
- Why can variables in the sensitivity set be excluded from a sensitivity analysis?
- What do sensitivity functions look like in general?
- Why are sensitivity functions for a prior joint probability linear?

[2]The subjects of evaluation and of test-selection, that are briefly touched upon here, are also directions of research pursued in the Decision Support Systems group of our Department.

- What is the amount of data you get from a sensitivity analysis?
- How can you select parameters of interest?
- What are advantages and disadvantages of performing a two-way sensitivity analysis as opposed to a one-way analysis?

Evaluation:

- In evaluating a Bayesian network against a standard of validity, why would you ideally use a *gold standard*? Why is this not always possible?
- What information does and doesn't a percentage correct convey?
- What is an acceptable percentage correct?
- What information does and doesn't the Brier score convey?
- What is an acceptable Brier score?

Application:

- What is the use of a two-layer architecture?
- In the threshold model for patient management, are the utilities patient-specific? What about the threshold probabilities?
- Can the threshold model for patient management be applied to a series of tests?
- Why are utilities in the threshold model said to subjective and in diagnostic problem solving objective?
- In diagnostic problem solving utilities are defined for values of binary variables only. Can this be generalised to non-binary variables?
- In diagnostic problem solving utilities represent the shift in probability of an hypothesis as a result of evidence. Why are large shifts in *either* direction awarded large utilities?
- In diagnostic problem solving, is the utility of a variable's value a fixed number? What about the expected utility of a node?
- In selective evidence gathering, what is the use of a *stopping criterion*?
- In selective evidence gathering, what are the effects of the *single-disorder* and *myopic* assumptions?