A Bayesian Network Reasoner to compute risk factors for physical health problems

Knowledge representation 2021 - project 2

No author given

Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081HV Amsterdam, The Netherlands

1 Introduction

Every year, more risk factors arise for problems in physical health. Knowing what causes these health issues are important for reducing the chance on alarming consequences like getting a stroke or becoming paralysed. [5]. A model to predict certain outcomes of possible risk factors can hopefully gain more insight in what causes have bigger effects compared to other causes. In this paper a network is modeled that shows what conditions may appear when somebody is being old or having a poor lifestyle. Calculating the probability for a stroke to happen or for a person to become paralysed represents a complex problem with many hidden variables. A Bayesian approach is used in this case, together with graphical models.

This model that is created is a Bayesian Network. This is a representation of a joint probability distribution of a set of random variables with a possible mutual causal relationship. The network has the form of a graph with nodes and edges. The nodes represent the variables, while the edges represent the causal relationship of the nodes[4]. The quantitative component of the networks are the conditional probability tables. They help to quantify the dependencies between the variables and their parents in the directed acyclic graph.

Bayesian networks are considered representational tools with the ability to represent large probability distributions compactly, and also the availability of inference algorithms that can answer queries about these distributions without necessarily constructing them. Bayesian networks can be constructed in many ways. It all depends on what the application is about and what the creator wants to obtain from it. They can be constructed using traditional knowledge engineering sessions with domain experts, for example. Bayesian networks have two interpretations, a standard one represented by probabilistic independence and a stronger one in terms of causality [10]. In some situations the probability of an event to happen, is changed when there exists an evidence. In this paper we present a Bayesian reasoner, a tool that uses different inference mechanisms to compute such probabilities. The calculation time for calculating the probability given something else increases very rapidly when the amount of nodes in a Network increases. By doing some smart steps in between, this reasoner is able to solve bigger Bayesian Networks in less time. Diverse methods are created that lead to the reasoner, those methods are tested on different given test cases.

First some explanation about the implementation of the reasoner is given. Thereafter, a performance evaluation of the Bayesian reasoner with different order methods. In the last part of the paper the network about risk factors for different diseases is created and described, and is tested by the reasoner for some interesting queries.

2 Implementation

2.1 d-Separation

D-Separation represents a graphical test for probabilistic independence in a Bayesian Network. The pruning method was applied in the implementation in order to determine if two sets of events (X and Y) are independent given the evidence set (Z). This method provides the decision in linear time. Given the fact that the Bayesian Network is a direct acyclic graph(G), the pruning method relies on: [1]

Theorem 1. *X and Y are d-separated by Z in G' if X and Y are disconnected in the pruned G' w.r.t. Z.*

G' is obtained by repetitively applying two steps. The first one deletes all leaf nodes, which do not belong to any of the X, Y and Z sets, whilst the second one deletes all edges ongoing from leaf nodes.

2.2 Ordering

The order of the variables during the variable elimination is important to reduce the complexity of a Bayesian reasoner. Both space and time can be reduced if the elimination of specific variables is done in such a way, that each computation of the reasoner takes as least variables as possible. Finding an optimal ordering is an NP-complete problem[3]. Therefore various heuristics were developed to find such optimal orderings.

In this project, three heuristics are used. The most simple and naive approach is using a random ordering of the variables.

The second heuristic which was implemented is called minimum degree ordering. This heuristic will give a score to each node in the interaction graph according to the number of neighbours that node has. The node with the lowest number of neighbours is picked and added to the list. Afterwards, the node is removed from the interaction graph and the neighbours are linked to each other. The whole process repeats until all the nodes are eliminated from the interaction graph.

The third used heuristic is called minimum fill ordering. By using this heuristic, the node which adds the least number of edges after its elimination is chosen first. As in the previous heuristic, every node has a score. The score is the difference between the degree of the node (the number of its neighbours) and the number of edges created after the node is eliminated. The node with the lowest score is chosen, added to the ordering list, eliminated from the interaction graph and then the process repeats until the interaction graph is empty.

2.3 Network Pruning

Network pruning is a process required in computing MAP and MPE queries, that simplifies the network without affecting its ability to answer the query correctly. This process involves two steps, Node Pruning and Edge Pruning that rely on: [2]

Theorem 2. *Let N be a Bayesian network and let (Q,e) be a corresponding query. If $N' = pruneNodes(N, Q \cup E)$, then $Pr(Q, e) = Pr(Q, e)$ where Pr and Pr' are the probability distributions induced by networks N and N', respectively.*

Theorem 3. *Let N be a Bayesian network and let e be an instantiation. If $N' = pruneEdges(N, e)$, then $Pr(Q, e) = Pr'(Q, e)$, where Pr and Pr' are the probability distributions induced by networks N and N, respectively.*

The Node Pruning Process consists of eliminating all leaf nodes that are not in $Q \cup E$, where Q is the set of query variables and E is the set of evidence events. The Edge Pruning Process involves deleting all edges outgoing from the evidence set of nodes and reducing the CPTs. The CPTs containing nodes from the evidence set are filtered and only the lines according to the evidence are kept in the new CPT. The columns of each evidence node are removed in the new CPT of each one of its children, as a result of the previous elimination of the edge between the two.

2.4 Marginal Distributions

Marginal distributions describe the probability of each variable in the probabilistic graphical model, given some other variables in the model. So, given a Bayesian (G, Θ) over X, inducing a probability distribution, each probability of x, given some evidence e can be given. [10] These can be used to answer probability queries on the graph [7]. Therefore, with a given evidence, the probability of evidence and posterior marginals on all network variables will follow from the marginal distribution.

Theorem 4. *Marginal distributions of a set of variables X is the set of all posterior marginals x, given their $P(x_s \mid e)$ with evidence $e \subset E$, the set of all evidence.*

To model $P(x_s \mid e)$, all conditional probabilities are taken for all variables in X: the joint marginal of the total set. Each variable in the network has a conditional probability table (CPT) assigned to it, containing the probabilities of the variable's status being the case (*True*) or not the case (*False*). Using the

Chain Rule, the joint probability $P(X)$ can be computed by multiplying the probabilities of each variable, with all intermediate dependencies included:

$$P(X) = \sum_{x,y \subseteq X} \theta x \mid y$$

A joint marginal $p(x_s, e)$ is the subset of such a joint probability, under evidence e [10]. Then, in order to compute the adherent posterior marginal $p(x_s \mid e)$, we take the joint marginal and normalize it with the evidence.

In order to marginalise, so to get to the $p(x_s, e)$, we need to compute a conditional probability of x_s and all its dependencies. This means that we first have to create a CPT that contains the variable, combined with all other conditional probabilities the variable appears in. We do so by multiplying the factors of these CPTs, which are the mappings of each instantiation in the CPT. After, the unwanted variables can be summed out / marginalized, by taking its probabilities of True and False into account for all the other variables in the CPT, so that in a set of f variables in X:

$$(\sum_X f)(y) = \sum_X f(x, y)$$

As earlier mentioned, to reduce the amount of computational effort, the multiplication order and elimination of each factor in the joint marginal matters. Instead of computing the complete joint probability and eliminating all variables that are not in the query afterwards, for each variable the joint marginal is taken and used in the joint marginal of the next variable, after which the variable is summed out of the CPT. The CPT that is left will be used for the multiplication of with the next variable in the total sum. In our implementation we used the random, minimum degree and minimum fill ordering.

2.5 MAP and MPE

An MAP query to a Bayesian network will answer the question of the most a posteriori estimate, thus the most likely instantiation of some network variables M, given some evidence e [10].

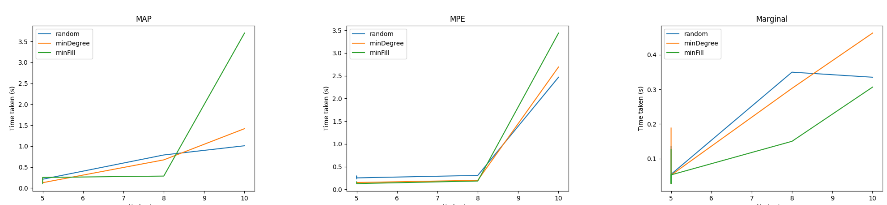
An MPE query is a specific type of MAP query, where no specific variables such as (M) are queried. The interest for an MPE query is just in the most likely instantiation of the total set of variables X, under evidence e [10].

In our implementation, the same way as in the marginal distribution function a CPT is created of either only the queried variables (MAP) or of all variables of the network (MPE), with their most likely outcomes. To do this, in MPE the sum out function of the marginal distribution is changed into the maxing out function. This function takes instead of the sum of two same rows, the maximum value and saves the instantiation that belongs to that maximum value. In the MAP function, this is somewhat the same, the only difference is that in this function first the variables that are not in M, are summed out and thereafter the M variables are removed by using the maxing out function. By using this maxing out function, the most likely instantiations of M are saved and there corresponding p-value. The maximum probability and the most likely instantiations belonging to that are returned.

3 Performance Evaluation

To test the performance of our Bayesian Reasoner we computed MAP and MPE with different elimination order heuristics. We additionally computed the marginal distribution. The networks that were used to show the performance were found in a repository on BNLearn[**?**]. The Bayesian networks have different dimensions being ranked from small to large. Our algorithm was not able to run networks as large as 30 nodes. Therefore we used 3 networks consisting of 5 nodes, one of 8 nodes and one of 10 nodes.

The performance is measured using the run-time of the program and it is shown in the following charts (Figure 1). The X-axis values represent the number of variables of the network, whilst the Y-values show the running time. The comparison was made between the heuristics implemented in the task1: random elimination order, minimum fill ordering and minimum degree ordering. To ensure the results are relevant, we used the same test and only changed the network size and ordering.



(a) First sub-figure (b) First sub-figure (c) First sub-figure

Fig. 1: Run times for MAP, MPE and marginal distribution

Because the timescale was small, we more thoroughly wanted to see the effect of ordering on running time. To do this, we calculated the MAP for the network with 10 nodes a total of 10 times for each ordering method. The median and spread of all three methods are shown in (Figure 2).

4 Use-case Model

A use-case model is created to show how the Bayesian Network Reasoner works and to get an interesting inside in how diverse conditions each other influences inside the Bayesian Network.

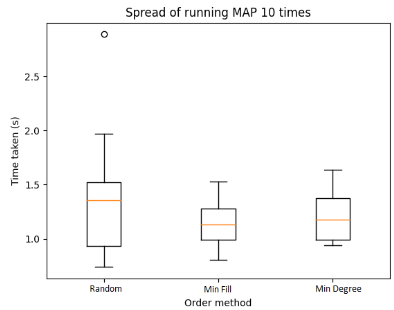


Fig. 2: Comparing difference in running time over 10 iterations

The diagram of the Bayesian Network shows possible conditions that can develop from age, from bad eating habits (being overweight or obese) or from using drugs regularly. The diagram can been seen in appendix A. Based on research, elderly people have a higher chance to have a stroke. A stroke occurs when the blood supply to a part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients[6]. After having a stroke, a patient can lose muscle movements (become paralysed) or develop sight problems[9]. Many people that become paralysed will have to use a wheelchair for the rest of their life. Furthermore, not only that being overweight can increase your chances of having a stroke, but also can increase the chance to develop type 2 diabetes, which has no cure. Diabetes is an impairment in the way the body regulates and uses sugar as a fuel. Diabetes can also appear as a cause of excessive use of a substance such as alcohol or drugs[5]. Diabetes without being treated correctly can help the development of other conditions, such as losing the ability to see well or fatigue[8]. As a consequence of fatigue or sight problems, a person can start having problems to walk and need a walking stick or the assistance of a caretaker.

We created this model with the idea in mind that a few factors can affect a person's quality of life. Aging is another factor that can help developing some of the conditions in the Bayesian Network. This factor is impossible to be avoided by anyone, but together with other factors such as being overweight, can accelerate the evolution of these conditions. Having poor eating habits will result in becoming overweight or obese, which afterwards increases the change of developing other conditions in the Bayesian Network. These conditions may even end with the death of a person when they are very severe.

The conditional probability tables were created with regard to how likely one condition might appear. The CPT's are visualised in the appendix of the document together with the Bayesian Network. For example, overweight is true in 30 percent of the cases. Drug abuse is only in 0.05 percent of the time the case. For the nodes in the Bayesian Network with parents, CPT's are made given the outcome of the parent node. So for example the CPT for Diabetes is given Drug Abuse and Overweight. Important to node is that those values are not totally based on real research. If this was the case, the probabilities would have probably been lower for a lot of values because for example having a stroke and paralysis together are very unlikely. However, for this research we want to investigate some queries and see what happens, and therefore it is better to give it a little bit higher chances, so actual differences are visible.

To test some of the inferences that can be made from our model, we queried it with "What is the probability of an individual to become paralysed if a person has overweight and is old, using a Bayesian Reasoner?" This probability is also calculated for when somebody without overweight and without being old and everything in between. The expectation is that being old and having overweight will increase the probability of somebody becoming paralysed, and when both are not the case, it will decrease the probability. Because both factors increase the chance of getting a stroke. The interesting part is if being old and no overweight gives a higher probability on becoming paralysed or the other way around. The expectation is that being old increases the chance of paralysing more rapidly compared to overweight because research found out that age is one of the important factors in having a stroke [6]. Prior marginal distribution is calculated to investigate the probability of drug abuse, being and having overweight together. The prior marginal is interesting to see what probability will arise when the conditions are arising together. The MPE is computed to find the most likely instantiations and probability of an individual that is drug addicted (or not). The MAP is computed to find the most likely instantiations and probability of use a wheelchair of an individual that is drug addicted and fatigue. Another MAP that is computed to find the most likely instantiations and probability of being paralysed given being drug addicted and having a stroke.

5 Use-case Model Results and discussion

All the tables of the results of the query tests can be found in appendix C. The probabilities for the prior marginal distribution is calculated for being old, abusing drugs and having overweight. The probability of all three the variables being False is the highest. This is what is expected since the chance of all three separate is as well, being higher for being False. Another interesting insight is that the chance is very low when the Drug Abuse is True. This is also in line with the expectations since drug abuse on it's own has a very low probability.

The results show that the posterior marginal query for paralysis, is changed when the given of being old and having overweight is changed. Indeed, as we inspected the change of being paralysed when not being old and not having overweight is low (5 percent). The change is the other way around when somebody is old and having overweight (84 percent). This can been seen in the figure 6. The other results are shown in the appendix. Somebody that is not old but with overweight has a 44 percent probability of being paralysed and when somebody is old but does not have overweight this is 78 percent. This shows that being old increases the chance of being paralysed more rapidly compared to having overweight.

	Paralysis	p
0	0.0	0.16
1	1.0	0.84

Fig. 3: Posterior marginal query for paralysis given Old=True and Overweight=True. 0 means False and 1 means True.

The MPE's that are calculated, are as well shown in appendix C. As can been seen is the most likely probability for all the different conditions together, higher when the evidence of drug abuse is false. Besides that, all the other conditions are false. This is expected since all the conditions are less likely to be true. This is because in this Bayesian Network all the nodes are exceptions of somebody that is 'healthy' which is the majority of the humans. When there is given that somebody drug abuse, the overall probability decreases. This could be due to the fact that the probability of somebody with drug abuse is already very low. Besides that, the most likely instantiation change to True for some of the variables. In real life this is probably not the case for all of them, since the probability of for example having diabetes is very low, also when somebody abuse drugs. However, as said before, this model is a simplified version of the real life and therefore higher values where chosen for the different probabilities than they would have been in real life.

The most likely instantiation and the corresponding probability of being paralysed given being drug addicted and having a stroke is calculated. Shown is that the most likely instantiation is true for being paralysed given this evidence. This is expected since the probability of being paralysed increases when somebody has a stroke. However, again should be taken into account that now higher probabilities are used for the nodes being true, and therefore this would have been different in a case based on real data. The chance of being paralysed is higher given that somebody got a stroke, however this could be still a very low probability. But in this bayesian network, we can conclude that those both evidences lead to a most likely instatiation of being paralysed. The other MAP that is computed to find the most likely instantiations and probability of using a wheelchair of an individual given that this individual is not drug addicted and not fatigue. It is most likely that this individual is not sitting in a wheelchair. This is with a probability of 37 percent which is high, as expected. However, it should be taken into account that a lot of other things have an influence on the probability of sitting in an wheelchair. Following our model wheelchair and being fatigue have both an indirect influence on the probability of sitting in an wheelchair.

6 Conclusion

In conclusion, it can be stated that the Bayesian Network reasoner that is created has helped to solve bigger Bayesian Networks. For the effects of network size on performance, we unfortunately could not test with larger network sizes. Therefore, conclusions taken from our results should be interpreted with caution. It seems evident that running time increases as networks get larger, which is expected. However, comparison between the performances of the different types of ordering for different network sizes is less obvious from our results. We additionally included a comparison between the ordering methods using only one size network. Differences in median are small, however random ordering has a trend towards needing a longer running time. Also the spread of running time for the random method is bigger than for the other two. These are both results that can be expected to even become more significant as network size increases. The reasoner has helped to answer interesting queries, about the Bayesian Network that was created. It is found that the probability of an individual to become paralysed if a person is old and having overweight increases. And decreases when somebody is not old and not having overweight. This is a very simplified network, and future research could take a look into more representative Bayesian Networks, with more nodes about possible risk factors. However, this Bayesian Network reasoner can be used to solve Bayesian Network questions.

References

1. Acar, E.: Orobabilistic graphical models (lecture 2), `https://canvas.vu.nl/courses/55684/files/4391980?wrap=1`
2. Acar, E.: Orobabilistic graphical models (lecture 4), `https://canvas.vu.nl/courses/55684/files/4433217?wrap=1`
3. Arnborg, S.: Efficient algorithms for combinatorial problems on graphs with bounded decomposability—a survey. BIT Numerical Mathematics **25**(1), 1–23 (1985)
4. Ben-Gal, I.: Bayesian networks. Encyclopedia of statistics in quality and reliability **1** (2008)
5. Clayton, R.R.: Transitions in drug use: Risk and protective factors. (1992)
6. Coco, D.L., Lopez, G., Corrao, S.: Cognitive impairment and stroke in elderly patients. Vascular health and risk management **12**, 105 (2016)
7. Darwiche, A.: A differential approach to inference in bayesian networks. Journal of the ACM (JACM) **50**(3), 280–305 (2003)
8. Meneilly, G.S., Tessier, D.: Diabetes in elderly adults. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences **56**(1), M5–M13 (2001)
9. Ring, H., Weingarden, H.: Neuromodulation by functional electrical stimulation (fes) of limb paralysis after stroke. Operative Neuromodulation pp. 375–380 (2007)
10. Van Harmelen, F., Lifschitz, V., Porter, B.: Handbook of knowledge representation. Elsevier (2008)

A Appendix - Use-case model

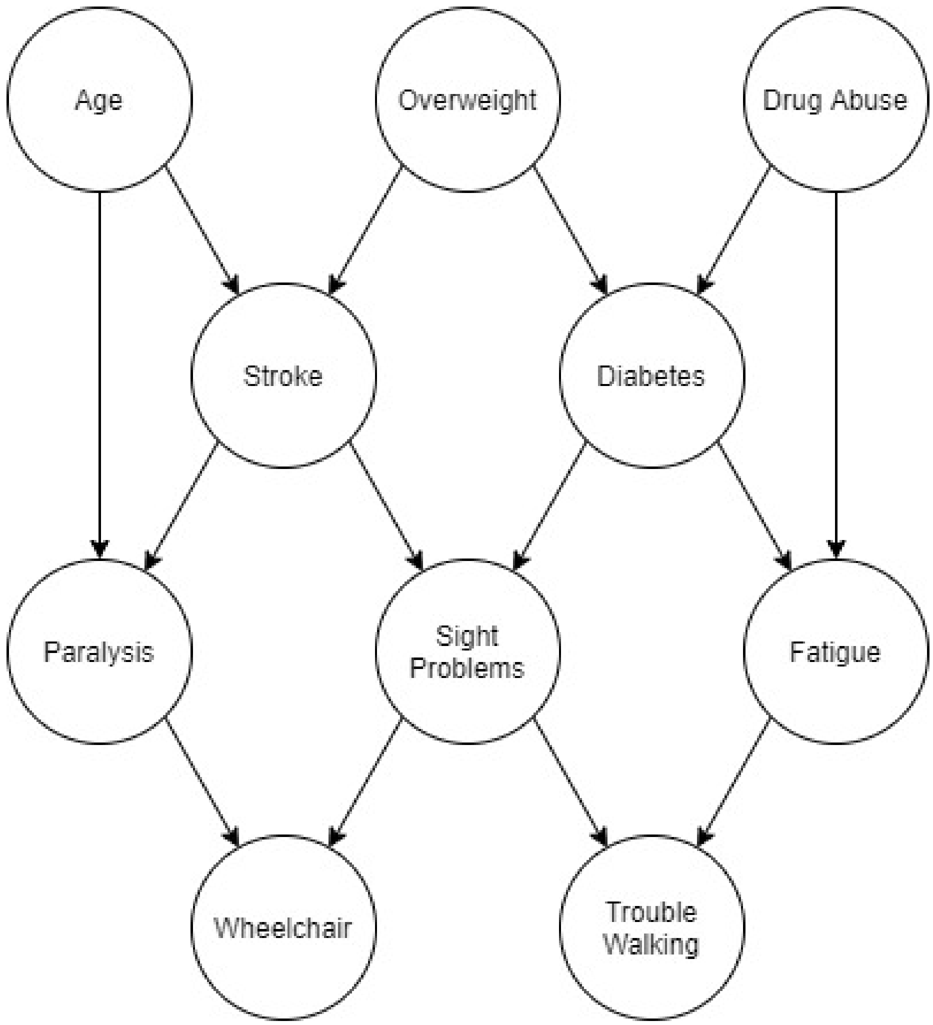


Fig. 4: A Bayesian Network with possible condition that can appear as you grow older, by being overweight or if you use drugs

B Appendix - CPTs for the use-case model

	Old	p
0	False	0.4
1	True	0.6

Fig. 5: CPT for Age

	Overweight	p
0	False	0.7
1	True	0.3

Fig. 6: CPT for Overweight

	Drug Abuse	p
0	False	0.95
1	True	0.05

Fig. 7: CPT for Drug Abuse

	Age	Overweight	Stroke	p
0	False	False	False	1.0
1	False	False	True	0.0
2	False	True	False	0.4
3	False	True	True	0.6
4	True	False	False	0.3
5	True	False	True	0.7
6	True	True	False	0.2
7	True	True	True	0.8

Fig. 8: CPT for Stroke given Age and Overweight

	Overweight	Drug Abuse	Diabetes	p
0	False	False	False	0.9
1	False	False	True	0.1
2	False	True	False	0.5
3	False	True	True	0.5
4	True	False	False	0.3
5	True	False	True	0.7
6	True	True	False	0.1
7	True	True	True	0.9

Fig. 9: CPT for Diabetes given Overweight and Drug Abuse

	Age	Stroke	Paralysis	p
0	False	False	False	0.95
1	False	False	True	0.05
2	False	True	False	0.30
3	False	True	True	0.70
4	True	False	False	0.60
5	True	False	True	0.40
6	True	True	False	0.05
7	True	True	True	0.95

Fig. 10: CPT for Paralysis given Age and Stroke

	Stroke	Diabetes	Sight Problems	p
0	False	False	False	0.8
1	False	False	True	0.2
2	False	True	False	0.6
3	False	True	True	0.4
4	True	False	False	0.2
5	True	False	True	0.8
6	True	True	False	0.1
7	True	True	True	0.9

Fig. 11: CPT for Sight Problems given Stroke and Diabetes

	Drug Abuse	Diabetes	Fatigue	p
0	False	False	False	0.80
1	False	False	True	0.20
2	False	True	False	0.30
3	False	True	True	0.70
4	True	False	False	0.30
5	True	False	True	0.70
6	True	True	False	0.05
7	True	True	True	0.95

Fig. 12: CPT for Fatigue given Drug Abuse and Diabetes

	Paralysis	Sight Problems	Wheelchair	p
0	False	False	False	0.99
1	False	False	True	0.01
2	False	True	False	0.60
3	False	True	True	0.40
4	True	False	False	0.05
5	True	False	True	0.95
6	True	True	False	0.00
7	True	True	True	1.00

Fig. 13: CPT for Wheelchair given Paralysis and Sight Problems

	Sight Problems	Fatigue	Trouble Walking	p
0	False	False	False	0.8
1	False	False	True	0.2
2	False	True	False	0.2
3	False	True	True	0.8
4	True	False	False	0.7
5	True	False	True	0.3
6	True	True	False	0.0
7	True	True	True	1.0

Fig. 14: CPT for Trouble Walking given Sight Problems and Fatigue

C Appendix - Use-case model results

The 0 in the tables means False and 1 means True.

	Drug Abuse	Old	Overweight	p
0	0.0	0.0	0.0	0.4655
1	1.0	0.0	0.0	0.0245
2	0.0	0.0	1.0	0.1995
3	1.0	0.0	1.0	0.0105
4	0.0	1.0	0.0	0.1995
5	1.0	1.0	0.0	0.0105
6	0.0	1.0	1.0	0.0855
7	1.0	1.0	1.0	0.0045

Fig. 15: Prior marginal query for being old, having overweight and drug abuse

	Paralysis	p
0	0.0	0.95
1	1.0	0.05

Fig. 16: Posterior marginal query for paralysis given Old=False and Overweight=False

	Paralysis	p
0	0.0	0.56
1	1.0	0.44

Fig. 17: Posterior marginal query for paralysis given Old=False and Overweight=True

	Paralysis	p
0	0.0	0.215
1	1.0	0.785

Fig. 18: Posterior marginal query for paralysis given Old=True and Overweight=False

Diabetes	Drug Abuse	Fatigue	Old	Overweight	Paralysis	Sight Problems	Stroke	Trouble Walking	Wheelchair	p
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.20174

Fig. 19: MPE query given drug abuse = False

Diabetes	Drug Abuse	Fatigue	Old	Overweight	Paralysis	Sight Problems	Stroke	Trouble Walking	Wheelchair	p
1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.005254

Fig. 20: MPE query given drug abuse = False

	Paralysis	p
0	1.0	0.127944

Fig. 21: MAP query for Paralysis given drug abuse = True and Stroke = True

Wheelchair	p
0.0	0.375059

Fig. 22: MAP query for sitting in a wheelchair given fatigue = False and Drug Abuse = False