COMP3308 Assignment 2

Mohammad Heydari
311223109

University of Sydney

Contents

1. Introduction---------- 3
2. Methods---------- 3
2.1. Bayesian Networks---------- 3
2.2. Variable Elimination ---------- 3
2.3. Likelihood Weighting---------- 3
3. Questions and Results ---------- 3
3.1. Task 1 ---------- 3
3.2. Task 2---------- 3
3.3. Task 3 ----------6
3.4. Task 4 ---------- 7
4. Conclusion ---------- 11
5. Reflection ---------- 11
6. Code ----------11
7. Reference ----------11
8. List of Figures ---------- 11
9. List of Tables ----------12

1. Introduction

The aim of this study is to build a Bayesian network for diagnostic problems and verify independence statements by using JavaBayes [1]. We also Implement an inference algorithm on a Bayesian network and compute posterior probabilities.

2. Methods

2.1. Bayesian Networks

Bayesian networks are probabilistic graphical models that represent sets of random variables and their conditional dependencies via a directed acyclic graph (DAG) [2]. They are used to define a scenario and calculate the event probabilities based on evidence and dependencies.

2.2. Variable Elimination

It is a simple and general exact inference algorithm in probabilistic graphical models, such as Bayesian networks and Markov random fields[3]. Variable elimination simplifies the network by removing variables. It eliminates or marginalize the non-evidence non-query variables individually by distributing the sum over the product. This method tries to avoid duplicate calculations and it also simplify large queries to a normalized message which can be calculated more easily.

2.3 Likelihood Weighting

It simply means that we use evidence to weight samples. Likelihood weighting is implemented to avoid the large number of sample rejection sampling wasting [4]. It is built similar to rejection sampling, however, we don't reject an observed node if it doesn't match the evidence. Instead, we keep a running value of the probabilities of the observed nodes encountered given the evidence encountered. The product of all these probabilities is then the weight of the sample over the entire network.

3. Questions and Results

3.1. Task 1

N/A

3.2. Task 2

Metastatic cancer is a possible cause of a brain tumor and is also an explanation for increased total serum calcium. In turn, either of these could explain a patient falling into a coma. Severe headache is also possibly associated with a brain tumor.
The prior probability of metastatic cancer $P(m)$ is 0.20.
The conditional probability of increased total serum calcium $P(I|M)$ is:

- $P(i|m) = 0.80$
- $P(i|\neg m) = 0.20$

The conditional probability of brain tumor $P(B|M)$ is:

- $P(b|m) = 0.20$
- $P(b|\neg m) = 0.05$

The conditional probability of coma $P(C|I, B)$ is:

- $P(c|i, b) = 0.80$
- $P(c|\neg i, b) = 0.80$
- $P(c|i, \neg b) = 0.80$
- $P(c|\neg i, \neg b) = 0.05$

The conditional probability of severe headache $P(S|B)$:

- $P(s|b) = 0.80$
- $P(s|\neg b) = 0.60$

3.2.1. Figure 1 below shows the equivalent graphical model

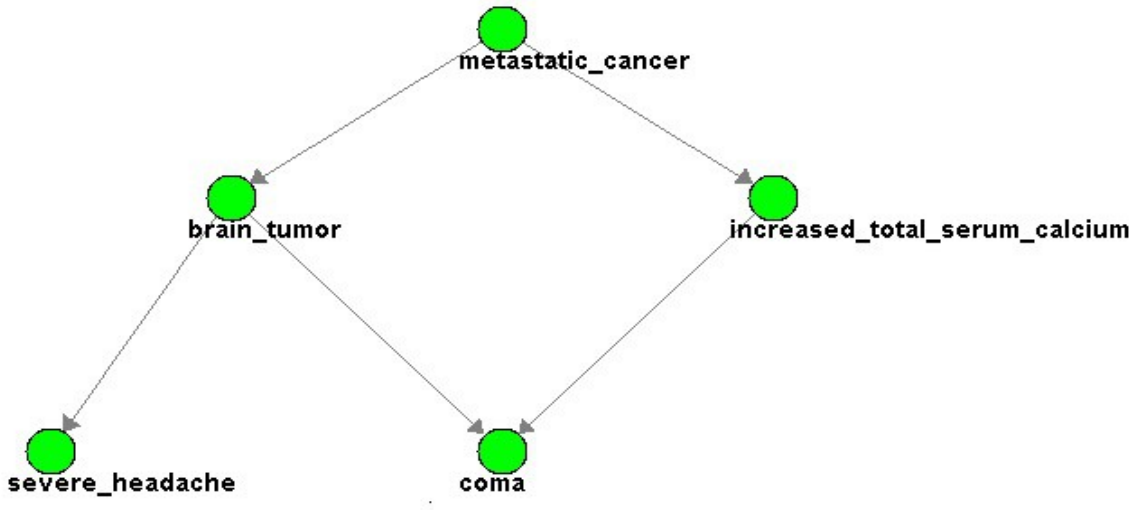

Figure 1 – Brain problems graphical network

3.2.2. P(C)?

Output of JavaBayes query:

```
Posterior distribution:
probability (  "coma" ) { //1 variable(s) and 2 values
        table
                0.32    // p(true | evidence )
                0.68;   // p(false | evidence );
}
```

So we have $P(C) = 0.32$

3.2.3. $P(M|S, \neg C)$?

We need to set the observed elements in JavaBayes. Output of JavaBayes query:

```
Posterior distribution:
probability (  "metastatic_cancer" ) { //1 variable(s) and 2 values
        table
                0.09727626459143972          // p(true | evidence )
                0.9027237354085603;          // p(false | evidence );
}
```

Based on query output from JavaBayes, $P(M|S, \neg C)$ = 0.09727626459143972

3.2.4. Markov blanket of coma?

A Markov blanket for a node in a Bayesian network is the set of nodes composed of node's parents, its children, and its children's other parents[5]. In our network, "coma" doesn't have any children and it has 2 parents, i.e. "brain_tumor" and "increased_total_serum_calcium". So the Markov blanket of coma is "brain_tumor" and "increased_total_serum_calcium".

3.2.5. No, "brain_tumor" and "increased_total_serum_calcium" are marginally independent, but given "coma" they become conditionally dependent. This is known as explaining away or Berkson's paradox.

3.2.6. $P(C \mid M)$?

We use JavaBayes tool. We use the "Observe" and "Query" tool to get what we want
Output from JavaBayes is:

```
Posterior distribution:
probability (  "coma" ) { //1 variable(s) and 2 values
        table
                0.68     // p(true | evidence )
                0.32;    // p(false | evidence );
}
```

so P(C|M) = 0.68

3.3. Task 3
3.3.1

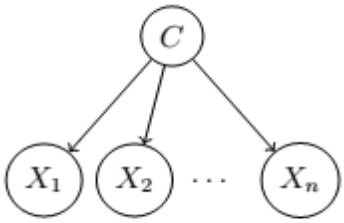

Figure 2 – Naive Bayes classifier network

$$P(C|X_1,,,,,X_n)=\frac{P(C)P(X_1,,,,,X_n|C)}{P(X_1,,,,,X_n)} \qquad \text{Bayes Theorem}$$

$$P(C|X_1,,,,,X_n)=\frac{P(C,X_1,,,,,X_n)}{P(X_1,,,,,X_n)} \qquad \text{Conditional Probability}$$

$$P(C|X_1,,,,,X_n)P(X_1,,,,,X_n)=P(C)P(X_1,,,,,X_n|C)$$

$$P(C|X_1,,,,,X_n)=P(X_1,,,,,X_n)\frac{P(C)P(X_1,,,,,X_n|C)}{P(X_1,,,,,X_n)}$$

$$P(C,X_1,,,,,X_n)=P(C)P(X_1,,,,,X_n|C)$$

$$P(X_1,,,,,X_n|C)=\prod_{i=1}^{n}P(X_i|C)$$

the last equation is the distribution of $X_1, , , , , X_n$. This can be represented as the product of each node's probability given *C*. hence we have following equation as we wanted:

$$P(C,X_1,,,,,X_n)=P(C)\prod_{i=1}^{n}P(X_i|C)$$

3.3.2.

$$\log\left(\frac{P(C=c|X_1,,,,,X_n)}{P(C=\neg c|X_1,,,,,X_n)}\right)=\log\left(\frac{P(C=c)}{P(C=\neg c)}\prod_{i=1}^{n}(\frac{P(X_i|C=c)}{P(X_i|C=\neg c)})\right)$$

$$= \log\left(\frac{P(C=c)}{P(C=\neg c)}\right)+\log\left(\prod_{i=1}^{n}(\frac{P(X_i|C=c)}{P(X_i|C=\neg c)})\right)$$

$$= \alpha_0+\log\left(\prod_{i=1}^{n}(\frac{P(X_i|C=c)}{P(X_i|C=\neg c)})\right) \quad \text{where} \quad \alpha_0=\log\left(\frac{P(C=c)}{P(C=\neg c)}\right)$$

$$= \alpha_0+\sum_{i=1}^{n}\log\left(\frac{P(X_i|C=c)}{P(X_i|C=\neg c)}\right)$$

$$= \alpha_0+\sum_{i=1}^{n}\alpha_i X_i \qquad \text{where} \quad \alpha_i=\log\left(\frac{P(X_i|C=c)}{P(X_i|C=\neg c)}\right)$$

Where (X_i =0 if X=¬x and 1 otherwise)

3.4. Task 4

In this study, we dealt with health related problems like a brain tumor. As a result we want to pay attention to another common problem, which is lung diseases. We build a Bayesian network that helps us to predict if a person has lung problem or not.

3.4.1 Random Variables

- **Age (A):** patient's age in 4 groups: <19, 20-30, 31-50, 50<.
- **Gender (G):** Male or female.
- **Physically active (PA):** True or False.
- **Smokes (S):** True or False.
- **Overweight (O):** True or False.
- **High cholesterol (HC):** True of False.
- **High blood pressure (HBP):** True or False.
- **Diabetes (DI):** Type1, Type2 or None.
- **Heart disease (HD):** True or False.

3.4.2 Probability Distributions

Prior probabilities for Age, Gender, Smokes, and Physically Active are shown in table 2, 3, 4, and 5 respectively.

Age			
<=19	(19 – 30]	(30 – 50]	50<
0.16	0.21	0.24	0.39

Table 2- prior probability table for Age

Gender	
Male	Female
0.62	0.38

Table 3- prior probability table for Gender

Smokes	
True	False
0.367	0.633

Table 4 – prior probability table for Smokes

Physically Active	
True	False
0.54	0.46

Table 5- prior probability table for Physically Active

P(O|PA) = 0.66
P(O|¬PA) = 0.01
P(HC|O) = 0.374
P(HC|¬O) = 0.15
P(HBP |S, O) = 0.215
P(HBP |S, ¬O) = 0.11
P(HBP |¬S, O) = 0.08
P(HBP |¬S, ¬O) = 0.05
P(D = type1|O) = 0.5, P(D = type1|¬O) = 0.4
P(D = type2|O) = 0.5, P(D = type2|¬O) = 0.4
P(D = none|O) = 0.2, P(D = none|¬O) = 0.7
P(HD|A ≤ 19, G = male, D = type1, HBP, HC) = 0.3
P(HD|A ≤ 19, G = male, D = type1, HBP, ¬HC) = 0.2
P(HD|A ≤ 19, G = male, D = type1, ¬HBP, HC) = 0.2
P(HD|A ≤ 19, G = male, D = type1, ¬HBP, ¬HC) = 0.1
P(HD|A ≤ 19, G = male, D = type2, HBP, HC) = 0.3
P(HD|A ≤ 19, G = male, D = type2, HBP, ¬HC) = 0.2
P(HD|A ≤ 19, G = male, D = type2, ¬HBP, HC) = 0.2
P(HD|A ≤ 19, G = male, D = type2, ¬HBP, ¬HC) = 0.1
P(HD|A ≤ 19, G = male, D = none, HBP, HC) = 0.2
P(HD|A ≤ 19, G = male, D = none, HBP, ¬HC) = 0.1
P(HD|A ≤ 19, G = male, D = none, ¬HBP, HC) = 0.1
P(HD|A ≤ 19, G = male, D = none, ¬HBP, ¬HC) = 0.05
P(HD|A ≤ 19, G = female, D = type1, HBP, HC) = 0.2
P(HD|A ≤ 19, G = female, D = type1, HBP, ¬HC) = 0.15
P(HD|A ≤ 19, G = female, D = type1, ¬HBP, HC) = 0.1
P(HD|A ≤ 19, G = female, D = type1, ¬HBP, ¬HC) = 0.05
P(HD|A ≤ 19, G = female, D = type2, HBP, HC) = 0.2
P(HD|A ≤ 19, G = female, D = type2, HBP, ¬HC) = 0.15
P(HD|A ≤ 19, G = female, D = type2, ¬HBP, HC) = 0.1
P(HD|A ≤ 19, G = female, D = type2, ¬HBP, ¬HC) = 0.05
P(HD|A ≤ 19, G = female, D = none, HBP, HC) = 0.1
P(HD|A ≤ 19, G = female, D = none, HBP, ¬HC) = 0.05
P(HD|A ≤ 19, G = female, D = none, ¬HBP, HC) = 0.1

P(HD|A ≤ 19, G = female, D = none, ¬HBP, ¬HC) = 0.05
P(HD|19 < A ≤ 30, G = male, D = type1, HBP, HC) = 0.4
P(HD|19 < A ≤ 30, G = male, D = type1, HBP, ¬HC) = 0.3
P(HD|19 < A ≤ 30, G = male, D = type1, ¬HBP, HC) = 0.3
P(HD|19 < A ≤ 30, G = male, D = type1, ¬HBP, ¬HC) = 0.2
P(HD|19 < A ≤ 30, G = male, D = type2, HBP, HC) = 0.4
P(HD|19 < A ≤ 30, G = male, D = type2, HBP, ¬HC) = 0.3
P(HD|19 < A ≤ 30, G = male, D = type2, ¬HBP, HC) = 0.3
P(HD|19 < A ≤ 30, G = male, D = type2, ¬HBP, ¬HC) = 0.2
P(HD|19 < A ≤ 30, G = male, D = none, HBP, HC) = 0.3
P(HD|19 < A ≤ 30, G = male, D = none, HBP, ¬HC) = 0.2
P(HD|19 < A ≤ 30, G = male, D = none, ¬HBP, HC) = 0.2
P(HD|19 < A ≤ 30, G = male, D = none, ¬HBP, ¬HC) = 0.1
P(HD|19 < A ≤ 30, G = female, D = type1, HBP, HC) = 0.3
P(HD|19 < A ≤ 30, G = female, D = type1, HBP, ¬HC) = 0.2
P(HD|19 < A ≤ 30, G = female, D = type1, ¬HBP, HC) = 0.2
P(HD|19 < A ≤ 30, G = female, D = type1, ¬HBP, ¬HC) = 0.1
P(HD|19 < A ≤ 30, G = female, D = type2, HBP, HC) = 0.3
P(HD|19 < A ≤ 30, G = female, D = type2, HBP, ¬HC) = 0.2
P(HD|19 < A ≤ 30, G = female, D = type2, ¬HBP, HC) = 0.2
P(HD|19 < A ≤ 30, G = female, D = type2, ¬HBP, ¬HC) = 0.1
P(HD|19 < A ≤ 30, G = female, D = none, HBP, HC) = 0.2
P(HD|19 < A ≤ 30, G = female, D = none, HBP, ¬HC) = 0.1
P(HD|19 < A ≤ 30, G = female, D = none, ¬HBP, HC) = 0.1
P(HD|19 < A ≤ 30, G = female, D = none, ¬HBP, ¬HC) = 0.05
P(HD|30 < A ≤ 50, G = male, D = type1, HBP, HC) = 0.6
P(HD|30 < A ≤ 50, G = male, D = type1, HBP, ¬HC) = 0.4
P(HD|30 < A ≤ 50, G = male, D = type1, ¬HBP, HC) = 0.5
P(HD|30 < A ≤ 50, G = male, D = type1, ¬HBP, ¬HC) = 0.3
P(HD|30 < A ≤ 50, G = male, D = type2, HBP, HC) = 0.6
P(HD|30 < A ≤ 50, G = male, D = type2, HBP, ¬HC) = 0.4
P(HD|30 < A ≤ 50, G = male, D = type2, ¬HBP, HC) = 0.5
P(HD|30 < A ≤ 50, G = male, D = type2, ¬HBP, ¬HC) = 0.3
P(HD|30 < A ≤ 50, G = male, D = none, HBP, HC) = 0.5
P(HD|30 < A ≤ 50, G = male, D = none, HBP, ¬HC) = 0.3
P(HD|30 < A ≤ 50, G = male, D = none, ¬HBP, HC) = 0.4
P(HD|30 < A ≤ 50, G = male, D = none, ¬HBP, ¬HC) = 0.2
P(HD|30 < A ≤ 50, G = female, D = type1, HBP, HC) = 0.4
P(HD|30 < A ≤ 50, G = female, D = type1, HBP, ¬HC) = 0.3
P(HD|30 < A ≤ 50, G = female, D = type1, ¬HBP, HC) = 0.3
P(HD|30 < A ≤ 50, G = female, D = type1, ¬HBP, ¬HC) = 0.2
P(HD|30 < A ≤ 50, G = female, D = type2, HBP, HC) = 0.4
P(HD|30 < A ≤ 50, G = female, D = type2, HBP, ¬HC) = 0.3
P(HD|30 < A ≤ 50, G = female, D = type2, ¬HBP, HC) = 0.3
P(HD|30 < A ≤ 50, G = female, D = type2, ¬HBP, ¬HC) = 0.2
P(HD|30 < A ≤ 50, G = female, D = none, HBP, HC) = 0.3
P(HD|30 < A ≤ 50, G = female, D = none, HBP, ¬HC) = 0.2
P(HD|30 < A ≤ 50, G = female, D = none, ¬HBP, HC) = 0.2
P(HD|30 < A ≤ 50, G = female, D = none, ¬HBP, ¬HC) = 0.1
P(HD|A > 50, G = male, D = type1, HBP, HC) = 0.7

P(HD|A > 50, G = male, D = type1, HBP, ¬HC) = 0.5
P(HD|A > 50, G = male, D = type1, ¬HBP, HC) = 0.6
P(HD|A > 50, G = male, D = type1, ¬HBP, ¬HC) = 0.4
P(HD|A > 50, G = male, D = type2, HBP, HC) = 0.7
P(HD|A > 50, G = male, D = type2, HBP, ¬HC) = 0.5
P(HD|A > 50, G = male, D = type2, ¬HBP, HC) = 0.6
P(HD|A > 50, G = male, D = type2, ¬HBP, ¬HC) = 0.4
P(HD|A > 50, G = male, D = none, HBP, HC) = 0.6
P(HD|A > 50, G = male, D = none, HBP, ¬HC) = 0.4
P(HD|A > 50, G = male, D = none, ¬HBP, HC) = 0.5
P(HD|A > 50, G = male, D = none, ¬HBP, ¬HC) = 0.3
P(HD|A > 50, G = female, D = type1, HBP, HC) = 0.5
P(HD|A > 50, G = female, D = type1, HBP, ¬HC) = 0.4
P(HD|A > 50, G = female, D = type1, ¬HBP, HC) = 0.4
P(HD|A > 50, G = female, D = type1, ¬HBP, ¬HC) = 0.3
P(HD|A > 50, G = female, D = type2, HBP, HC) = 0.5
P(HD|A > 50, G = female, D = type2, HBP, ¬HC) = 0.4
P(HD|A > 50, G = female, D = type2, ¬HBP, HC) = 0.4
P(HD|A > 50, G = female, D = type2, ¬HBP, ¬HC) = 0.3
P(HD|A > 50, G = female, D = none, HBP, HC) = 0.4
P(HD|A > 50, G = female, D = none, HBP, ¬HC) = 0.3
P(HD|A > 50, G = female, D = none, ¬HBP, HC) = 0.3
P(HD|A > 50, G = female, D = none, ¬HBP, ¬HC) = 0.2

3.4.3. Method

We mostly used the internet to gather our data, mostly medical related websites and specifically heart related ones. We then formed our common causes, and we determined which one could be used as random variables in our Bayesian network. We researched further to find statistics to model our probability distribution. We also linked these variables and found the following from research various sources on the Internet:

- Physical activity greatly affects if a person is overweight or not.
- Smoking and being overweight can cause a person to have a high blood pressure.
- Being overweight can cause a person to have high cholesterol.
- Being overweight can cause a person to have diabetes.
- Age, gender, high blood pressure, high cholesterol and diabetes are the causes of heart disease.

From here we assigned the probabilities for the variables. Age and gender were based on Australian age and gender demographics. Physically active, overweight and smoking were also taken from Australian statistics found on the Internet. From these defined probabilities we used the Bayes theorem and conditional joint probability to calculate the rest of the probabilities. Some of these probability, however were estimated and weighted.
Finally, we drew up the Bayes network based on our finds and calculations which can see below in Figure 3.

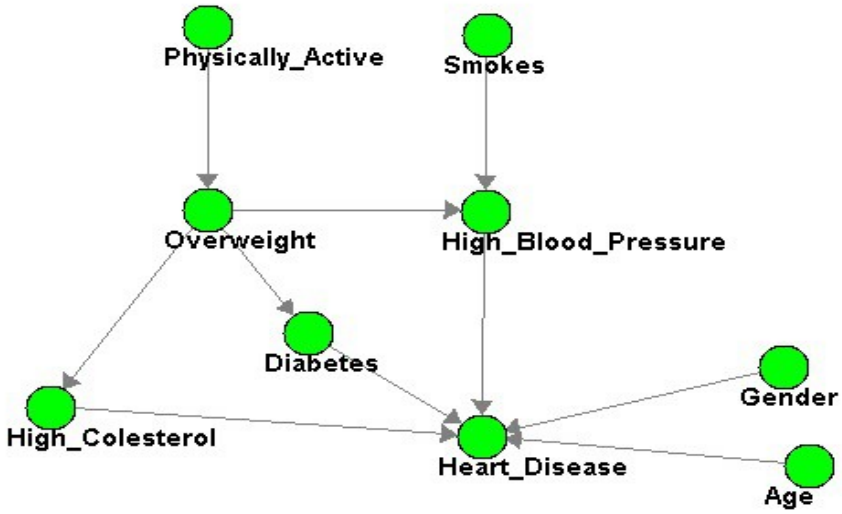

Figure 3– Heart Disease Bayesian Network

4. Conclusions

We examined some health problems and we noticed that it is a serious problem. We learned how to use JavaBayes tools. It makes probabilities computation lot easier. It helps us to build a Bayesian network that gives us a visual insight into the problem we are trying to solve.
Although, these problems we addressed in this paper were not complicated. It is a promising area to work on building more complex and useful Bayesian network for some common problems. They can be medical related areas or other areas like agriculture, engineering, and etc.

5. Reflection

Unfortunately I couldn't do well in this assignment like my first assignment. One reason is that, I worked individually this time as well, but workload from my subjects effected the quality. Another reason is that I didn't want to work n a group based on my previous experience, I have been only team up with bad and lazy students and I end up doing everything myself. Good students want to be with each other, and nobody they don't want to be team up with a person at least 10 years older than them. Having said that, either way good or bad, I learned some new materials and new tools specifically, I learned how to create graphical models.

6. Code

I sadly didn't developed any code for Task 1, so there are no code, and hence no instructions on how to run it. May the force be with me.

7. Reference

[1] http://www.cs.cmu.edu/~javabayes/Home/
[2] http://en.wikipedia.org/wiki/Bayesian_network
[3] Zhang, N.L., Poole, D.: A Simple Approach to Bayesian Network Computations. In:7th Canadian Conference on Artificial Intelligence, pp. 171–178. Springer, New York(1994)
[4] https://facwiki.cs.byu.edu/cs677sp09/index.php/LW-08#Likelihood_Weighting
[5] http://en.wikipedia.org/wiki/Markov_blanket

8. List of Figures

- 3.2.1. Figure 1 – Brain problems graphical network
- 3.3.1. Figure 2 – Naive Bayes classifier network
- 3.4.3. Figure 3– Heart Disease Bayesian Network

9. List of Tables

- 3.1. Table.1: Accuracy results for likelihood weighting
- 3.4.2. Table 2- prior probability table for Age
- 3.4.2. Table 3- prior probability table for Gender
- 3.4.2. Table 4 – prior probability table for Smokes
- 3.4.2. Table 5- prior probability table for Physically Active