CS6190: Probabilistic Modeling Homework 2 Bayesian Networks

Gopal Menon

10 March, 2018

Written Part

1. The total probability of the network is given by

$$\begin{aligned} p(Beysian\ Network) &= p(income)p(smoke|income)p(exercise|income)p(bmi|income, exercise) \\ &\times p(blood\ pressure|exercise, income, smoking)p(cholesterol|exercise, income, smoking) \\ &\times p(diabetes|bmi)p(stroke|bmi, bp, cholesterol)p(attack|bmi, bp, cholesterol) \\ &\times p(angina|bmi, bp, cholesterol) \end{aligned}$$

The total probability distribution requires the product of 10 random variables. If the probabilities are multiplied as shown above, the probability table will have 10 variables. Number of allowed states for the random variables are:

Table 1: CDC Survey Variable states

Survey Variable	Number of states
income	8
exercise	2
smoke	2
bmi	4
bp	4
cholesterol	2
angina	2
stroke	2
attack	2
diabetes	4

The maximum number of rows in the probability distribution table for the above variables will be the product of the number of allowed states - $8 \times 2 \times 2 \times 4 \times 4 \times 2 \times 2 \times 2 \times 2 \times 4 = 32768$. This is the number of probabilities that are needed to store the full joint distribution.

However it is easier and cheaper in terms of memory requirements to store probabilities corresponding to each node in the Bayes Net. For example the number of probabilities that need to be stored for the node exercise corresponds to $p(exercise|income)$ which will have $2 \times 8 = 16$ different probabilities that need to be stored. The following table shows the number of probabilities that need to be stored at the level of each node.

Table 2: Number of Probabilities to be stored at node level

Survey Variable	Number of Probabilities
income	8
exercise	16

Survey Variable	Number of Probabilities
smoke	16
bmi	64
bp	128
cholesterol	64
angina	64
stroke	64
attack	64
diabetes	16
Total	504

So it is more efficient to store probabilities at the level of a node and based on the query, marginalize variables that we are not interested in after observing the variables that are given (on the right side of the conditional probability).

2. For each of the four health outcomes (diabetes, stroke, heart attack, angina)

   (a) Probability of the outcome if I have bad habits (smoke and don't exercise)? How about if I have good habits (don't smoke and do exercise)?

   Table 3: Health Outcome Probabilities based on Habits

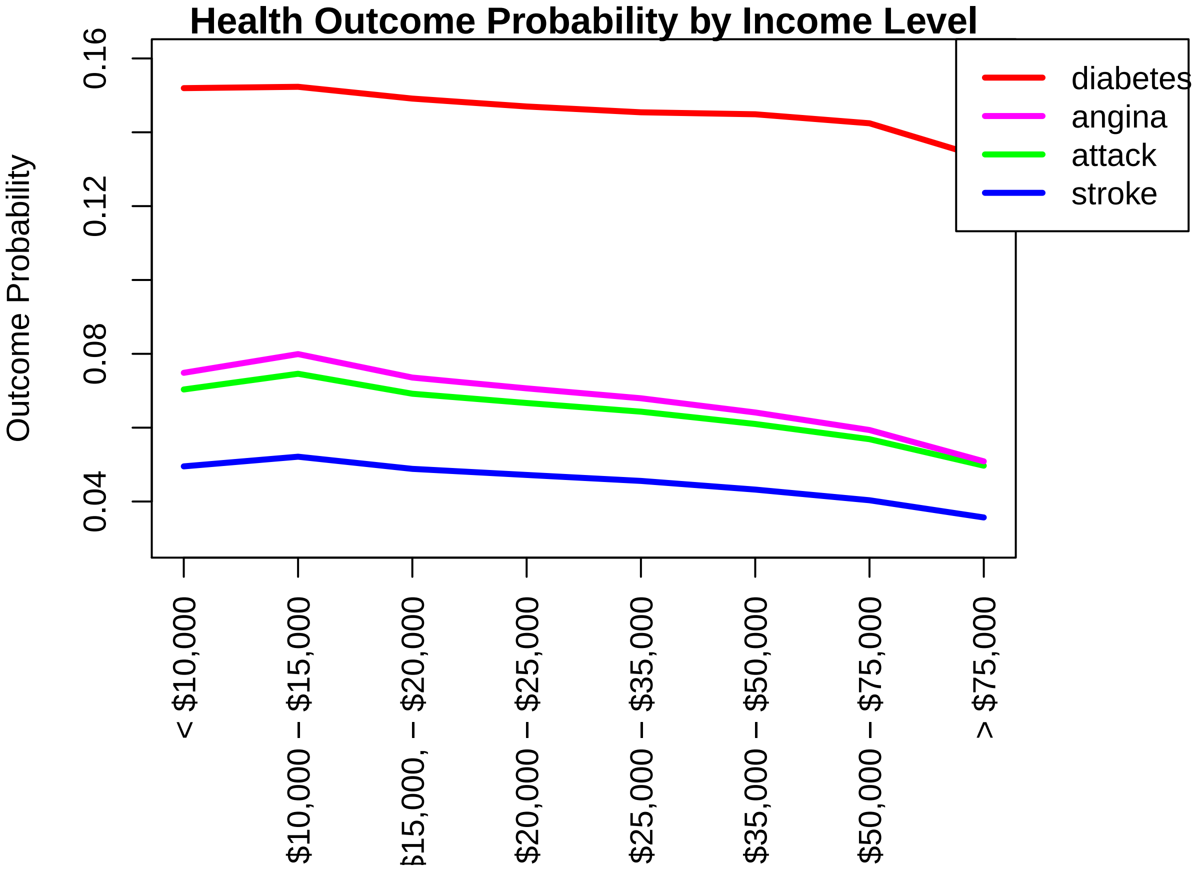
Health Outcome	Bad Habits	Good Habits
diabetes	0.1593304	0.1350624
stroke	0.0501267	0.0368082
attack	0.0724154	0.0510272
angina	0.0777561	0.0523189

   (b) What is the probability of the outcome if I have poor health (high blood pressure, high cholesterol, and overweight)? What if I have good health (low blood pressure, low cholesterol, and normal weight)?

   Table 4: Health Outcome Probabilities based on Health

Health Outcome	Poor Health	Good Health
diabetes	0.1222794	0.0616345
stroke	0.0839749	0.0138704
attack	0.1343300	0.0158879
angina	0.1531853	0.0128387

3. Evaluate the effect a person's income has on their probability of having one of the four health outcomes (diabetes, stroke, heart attack, angina). For each of these four outcomes, plot their probability given income status (your horizontal axis should be $i = 1, 2, \ldots, 8$, and your vertical axis should be $P(y = 1|income = i)$, where y is the outcome). What can you conclude?



It looks like overall the probability of a bad health outcome reduces with increasing income. However, it looks like the probability first increases from the lowest to the next highest income and then reduces.

4. Notice there are no links in the graph between the habits (smoking and exercise) and the outcomes. What assumption is this making about the effects of smoking and exercise on health problems? Let's test the validity of these assumptions. Create a second Bayesian network as above, but add edges from smoking to each of the four outcomes and edges from exercise to each of the four outcomes. Now redo the queries in Question 2. What was the effect, and do you think the assumptions of the first graph were valid or not?

   I think the assumption made in the graph is that the habits (smoking and exercise) have an impact on the health (bmi, bp and cholesterol) of the person and thus indirectly impact the outcomes (diabetes, stroke, attack and angina).

Table 5: Health Outcome Probabilities based on Habits

Health Outcome	Bad Habits	Good Habits
diabetes	0.2267429	0.1025276
stroke	0.0790277	0.0253386
attack	0.1175417	0.0303845
angina	0.1142626	0.0359644

Table 6: Health Outcome Probabilities based on Health

Health Outcome	Poor Health	Good Health
diabetes	0.1310737	0.0576544

Health Outcome	Poor Health	Good Health
stroke	0.0857540	0.0133544
attack	0.1361617	0.0152652
angina	0.1548693	0.0124368

Comparing tables for health outcomes based on habits, we can see that when we have links from habits to outcomes,the probability for the outcome is higher for bad habits and lower for good habits in the latter case. This suggests that habits do directly impact the outcomes and so the second graph is a more accurate representation of the model. Its not that the first graph was invalid, it was less accurate.

Comparing tables for health outcomes based on health, we can see that when we have links from health to outcomes,the probability for the outcome is higher for bad health and lower for good health in the latter case. Again, this suggests that habits do directly impact the outcomes and so the second graph is a more accurate representation of the model.

5. Also notice there are no edges between the four outcomes. What assumption is this making about the interactions between health problems? Make a third network, starting from the network in Question 4, but adding an edge from diabetes to stroke. For both networks, evaluate the following probabilities:

   P(stroke = 1|diabetes = 1) and P(stroke = 1|diabetes = 3)

   Again, what was the effect, and was the assumption about the interaction between diabetes and stroke valid?

Table 7: Probability for stroke when diabetes is present - no edge from diabetes to stroke

Probability	stroke
0.0451013	Will have stroke
0.9548987	Will not have stroke

Table 8: Probability for stroke when diabetes is not present - no edge from diabetes to stroke

Probability	stroke
0.0412291	Will have stroke
0.9587709	Will not have stroke

Table 9: Probability for stroke when diabetes is present - edge from diabetes to stroke

Probability	stroke
0.0764278	Will have stroke
0.9235722	Will not have stroke

Table 10: Probability for stroke when diabetes is not present - edge from diabetes to stroke

Probability	stroke
0.0358626	Will have stroke

Probability	stroke
0.9641374	Will not have stroke

The assumption being made when there are no edges between the four outcomes is that only the health conditions (and the habits when there is a link from them to the outcomes), affect the outcomes and that the presence of one outcome does not impact the probability of another outcome.

When there is no edge in the bayes net from diabetes to stroke, the probability of stroke does not vary much whether diabetes is present or not present. However when there is an edge, probability of stroke is much higher when diabetes is present compared to when it is not. So this suggests that there is an impact of diabetes on the probability of having a stroke. So this suggests that the graph with the edge from diabetes to stroke is a more accurate representation of the model.

6. Finally, make sure that your code runs correctly on all of the examples in BayesNetExamples.r. Your code will be graded for correctness on these also.

Table 11: Bishop 8.30 - p(G = 0)

probs	gauge
0.685	1
0.315	0

Table 12: Bishop 8.31 - p(G = 0|F = 0)

probs	gauge	fuel
0.19	1	0
0.81	0	0

Table 13: Bishop 8.32 - p(F = 0|G = 0)

probs	gauge	fuel
0.7428571	0	1
0.2571429	0	0

Table 14: Bishop 8.33 - p(F = 0|G = 0, B = 0)

probs	gauge	battery	fuel
0.8888889	0	0	1
0.1111111	0	0	0

Table 15: Kevin Murphy - Pr(S=1|W=1)

probs	wet	sprinkler
0.5702364	T	F
0.4297636	T	T

Table 16: Kevin Murphy - Pr(R=1|W=1)

probs	wet	rain
0.2920723	T	F
0.7079277	T	T

Table 17: Kevin Murphy - Pr(W=1)

probs	wet
0.3529	F
0.6471	T

Table 18: Kevin Murphy - Pr(S=1|W=1,R=1)

probs	wet	sprinkler	rain
0.805501	T	F	T
0.194499	T	T	T