AU 332 Artificial Intelligence: Principles and Techniques

HomeWork: 04

By: Yangqian Wu

Instructor: Yue Gao

November 21, 2019

I. INTRODUCTION

In homework 4, the topic is focus on BayesNetWork.

In Coding part, I implement the code for computing exact inferences in Bayesian networks of discrete random variables using variable elimination. By using function readFactorTable and readFactorTablefromData, we can build conditional probability tables. Then, I implement five functions which are helpful to accomplish variable elimination : joinFactors , marginalizeFactor , marginalizeNetworkVariables , evidenceUpdateNet , inference.

In Written part, I analyze risk factors for certain health problems,including heart disease, stroke, heart attack and diabetes. Answer the five questions based on the RiskNetwork.

II. CODING PART

A. Function Implementation

1. *joinFactors(factor1, factor2)*

Return a factor table that is the join of factor 1 and 2. we can assume that the join of two factors is a valid operation.

```python
def joinFactors(factor1, factor2):
    # your code
    #We use pd.DataFrame.copy() to avoid changing the original factor table
    factor1_copy = pd.DataFrame.copy(factor1)
    factor2_copy = pd.DataFrame.copy(factor2)
    #Figure the same column for factor1 and factor2
    samecolumn = list(column for column in factor1_copy.columns if column in factor2_copy.columns)
    samecolumn.remove('probs')
    #Discuss whether factor1 and factor2 has same column name besides 'probs'
    if len(samecolumn) == 0:
        #Plus 'augxiliary' column to pd.merge
        factor1_copy['augxiliary'] = 0
        factor2_copy['augxiliary'] = 0
        factor = pd.merge(factor1_copy,factor2_copy,how='outer',on=['augxiliary'])
        #Count the new probability
        factor['probs_x'] *= factor['probs_y']
        #Drop the unnecessary columns
        factor = factor.drop(['probs_y','augxiliary'],axis=1)
        factor = factor.rename(columns = {'probs_x':'probs'})
    else:
        #pd.merge by same column
        factor = pd.merge(factor1_copy,factor2_copy,how='outer',on=samecolumn)
        #Count the new probability
        factor['probs_x'] *= factor['probs_y']
        #Drop the unnecessary columns
        factor = factor.drop(['probs_y'],axis=1)
        factor = factor.rename(columns = {'probs_x':'probs'})

    return factor
```

(a) The code for joinFactors

2. *marginalizeFactor(factorTable, hiddenVar)*

This function return a factor table that marginalizes margVar out of it. we can assume that hiddenVar is on the left side of the conditional.

```python
def marginalizeFactor(factorTable, hiddenVar):
    # your code
    #We use pd.DataFrame.copy() to avoid changing the original factor table
    factor = pd.DataFrame.copy(factorTable)
    column = list(factor.columns)
    column.remove('probs')
    column.remove(hiddenVar)
    #The columns we need is remained
    factor = factor.drop(hiddenVar,axis=1)
    #Group by the columns, sum the 'probs' up
    factor = factor[factor.columns].groupby(column,as_index=False).sum()
    #Put the 'probs' column to the first
    probs_content = factor['probs']
    factor = factor.drop('probs',axis=1)
    factor.insert(0,'probs',probs_content,True)

    return factor
```

(b) The code for marginalizeFactor

3. marginalizeNetworkVariables(bayesNet, hiddenVar)

This function takes a Bayesian network, bayesNet, and marginalizes out a list of variables hiddenVar.

```python
def marginalizeNetworkVariables(bayesNet, hiddenVar):
    # your code
    bayesNet_update = bayesNet.copy()
    #Figure out all the column names
    column = []
    for factor in bayesNet_update:
        column.extend(list(factor.columns))
    column = set(column)
    #For every hidden variable, firstly join the factors related to it, then marginalize it
    for var in hiddenVar:
        if var in column:
            bayesNet_left = bayesNet_update.copy()
            factor_update = pd.DataFrame(columns=['probs'])
            #Consider the factor in net, whether it's related to hidden variable
            for factor in bayesNet_left:
                if var in factor.columns:
                    #If the factor is related to hidden variable, delete it in net
                    for j in range(len(bayesNet_update)):
                        if list(bayesNet_update[j].columns) == list(factor.columns):
                            del bayesNet_update[j]
                            break
                    #Join it with updated factor
                    if factor_update.empty:
                        factor_update = factor
                    else:
                        factor_update = joinFactors(factor_update,factor)
            #Marginalize the hidden variable
            factor_update = marginalizeFactor(factor_update,var)
            #Put the new factor into the net
            bayesNet_update.append(factor_update)
```

(c) The code for marginalizeNetworkVariables

4. evidenceUpdateNet(bayesNet, evidenceVars, evidenceVals)

This function takes a Bayesian network, bayesNet, and sets the list of variables, evidenceVars, to the corresponding list of values, evidenceVals. We do not normalize the factors to be proper probabilities.

```python
def evidenceUpdateNet(bayesNet, evidenceVars, evidenceVals):
    # your code
    bayesNet_update = bayesNet.copy()
    #Consider the evidence variable one by one
    for i in range(len(evidenceVars)):
        bayesNet_left = bayesNet_update.copy()
        #For every factor in net, only remain the evidence variable with its instructed value
        for factor in bayesNet_left:
            if evidenceVars[i] in factor.columns:
                for j in range(len(bayesNet_update)):
                    if list(bayesNet_update[j].columns) == list(factor.columns):
                        del bayesNet_update[j]
                        break
                #Figure the new factor, and add it to the net
                factor_update = factor[factor[evidenceVars[i]]==evidenceVals[i]]
                bayesNet_update.append(factor_update)
    return bayesNet_update
```

(d) The code for evidenceUpdateNet

5. inference(bayesNet, hiddenVar, evidenceVars, evidenceVals)

This function takes in a Bayesian network and returns a single joint probability table resulting from the given set of evidence variables and marginalizing a set of hidden variables. We normalize the table to give valid probabilities.The final table should be a proper probability table (entries sum to 1). The hidden variables shown in hiddenVar should not be in the returned table.

```python
def inference(bayesNet, hiddenVar, evidenceVars, evidenceVals):
    # your code
    bayesNet_update = bayesNet.copy()
    #Firstly, filter the evidence value which are not satisfied with the instruction
    bayesNet_update = evidenceUpdateNet(bayesNet_update,evidenceVars,evidenceVals)
    #Secondly, marginalize the hidden variable
    bayesNet_update = marginalizeNetworkVariables(bayesNet_update,hiddenVar)

    #Join the remained factors until there is only one factor, which is the factor we query
    while len(bayesNet_update) != 1:
        factor1 , factor2 = np.random.choice(bayesNet_update,2)
        if len(set(factor1.columns).intersection(set(factor2.columns))) > 1 and list(factor1.columns) != li
            for j in range(len(bayesNet_update)):
                if list(bayesNet_update[j].columns) == list(factor1.columns):
                    del bayesNet_update[j]
                    break
            for j in range(len(bayesNet_update)):
                if list(bayesNet_update[j].columns) == list(factor2.columns):
                    del bayesNet_update[j]
                    break
            factor_update = joinFactors(factor1,factor2)
            bayesNet_update.append(factor_update)
    factor_final = bayesNet_update[0]

    #Normalize
    total = sum(list(factor_final['probs']))
    factor_final['probs'] /= total

    return factor_final
```

(e) The code for inference

B. The Screen-shot for Examples

When I accomplished the coding part and test it on the examples given in BayesNetworkTestScript.py, I found the answer is equal to the screen-shot given in pdf file, which proved the correctness for coding part.

```
inference starts
   probs  gauge
0  0.315      0
1  0.685      1
   probs  fuel  gauge
0   0.81     0      0
1   0.19     0      1
      probs  fuel  gauge
0  0.742857     1      0
1  0.257143     0      0
      probs  battery  fuel  gauge
0  0.888889        0     1      0
1  0.111111        0     0      0
inference ends
income dataframe is
      probs  income
0  0.050848       1
1  0.059429       2
2  0.074042       3
3  0.094414       4
4  0.116356       5
5  0.150725       6
6  0.164430       7
7  0.289755       8
      probs  smoke  exercise  diabetes
0  0.136660      1         2         1
1  0.008915      1         2         2
2  0.837385      1         2         3
3  0.017040      1         2         4
```

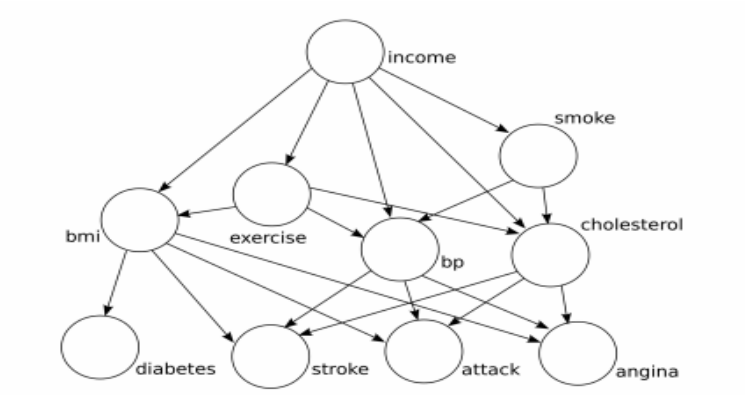
(f) The screen-shot for example

III. WRITTEN PART

In the written part, I analyze risk factors for certain health problems,including heart, disease, stroke, heart attack and diabetes. According to the data from the 2015 Behavioral Risk Factor Surveillance System survey, we consruct the BayesNet and calculate the probability we are interested in.

A. BayesNetwork Description

In the BayesNetwork, the variables and their meanings are as follows:



(g) The visual connection between each variale

- **income** - Annual personal income level.
  1($< \$10,000$) 2($\$10,000 - \$15,000$) 3($\$15,000, - \$20,000$)
  4($\$20,000 - \$25,000$) 5($\$25,000 - \$35,000$) 6($\$35,000 - \$50,000$)
  7($\$50,000 - \$75,000$) 8($> \$75,000$)

- **exercise** - Exercised in past 30 days.
  1 (yes) 2 (no)

- **smoke** - Smoked 100 or more cigarettes in lifetime.
  1 (yes) 2 (no)

- **bmi** - Body mass index (category).
  1 (underweight) 2 (normal) 3 (overweight) 4 (obese)

- **bp** - Has high blood pressure.
  1 (yes) 2 (only when pregnant) 3 (no) 4 (pre-hypertensive)

- **cholesterol** - Has high cholesterol.
  1 (yes) 2 (no)

- **angina** - Had heart disease (angina).
  1 (yes) 2 (no)

- **stroke** - Had a stroke.
  1 (yes) 2 (no)

- **attack** - Had a heart attack.
  1 (yes) 2 (no)

- **diabetes** - Had diabetes.
  1 (yes) 2 (only during pregnancy) 3 (no) 4 (pre-diabetic)

B. Questions

1. *Question1*

Question : What is the size (in terms of the number of probabilities needed) of this network? And what is the total number of probabilities needed to store the full joint distribution?

Answer :

(1)By counting the total number of entries in all of the CPT, we can figure out that the size of this network is 504.

(2)Theoretically, if we want to store the full joint distribution, and there are 10 variables and three variables have 4 possible values, one variable has 8 possible values, others are binary values. Therefore, the total number of probabilities should be $2^6 * 4^3 * 8^1 = 32768$.

```
########################################
The answer for question 1 is :
The size of RiskFactorNet is : 504
########################################
```

2. *Question2*

Question : For each of the four health outcomes (diabetes, stroke, heart attack, angina), answer the following by querying and inferring from your Bayesian network. Please write the results in a table format.

(a) What is the probability of the outcome if I have bad habits (smoke and dont exercise)? How about if I have good habits (dont smoke and do exercise)?

(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)?

Answer :

(a)If I have bad habits(smoke and dont exercise), the probability of the outcome is as below:

```
########################################
The answer for question 2 is :
What is the probability of the diabetes  if I have bad habits (smoke and don't exercise)
      probs  diabetes  smoke  exercise
0  0.150516         1      1         2
1  0.008965         2      1         2
2  0.822423         3      1         2
3  0.018096         4      1         2
What is the probability of the stroke  if I have bad habits (smoke and don't exercise)?
      probs  stroke  smoke  exercise
0  0.049264       1      1         2
1  0.950736       2      1         2
What is the probability of the attack  if I have bad habits (smoke and don't exercise)?
     probs  attack  smoke  exercise
0  0.07433       1      1         2
1  0.92567       2      1         2
What is the probability of the angina  if I have bad habits (smoke and don't exercise)?
      probs  angina  smoke  exercise
0  0.080448       1      1         2
1  0.919552       2      1         2
```

(h) The probability of outcome by given bad habits

If I have good habits(dont smoke and do exercise), the probability of the outcome is as below:

```
What is the probability of the diabetes  if I have good habits (don't smoke and do exercise)?
      probs  diabetes  smoke  exercise
0  0.127119         1      2         1
1  0.008865         2      2         1
2  0.847693         3      2         1
3  0.016323         4      2         1
What is the probability of the stroke  if I have good habits (don't smoke and do exercise)?
     probs  stroke  smoke  exercise
0  0.03611       1      2         1
1  0.96389       2      2         1
What is the probability of the attack  if I have good habits (don't smoke and do exercise)?
      probs  attack  smoke  exercise
0  0.052798       1      2         1
1  0.947202       2      2         1
What is the probability of the angina  if I have good habits (don't smoke and do exercise)?
      probs  angina  smoke  exercise
0  0.054755       1      2         1
1  0.945245       2      2         1
```

(i) The probability of outcome by given good habits

(b)If I have poor health (high blood pressure, high cholesterol, and overweight), the probability of the outcome is as below:

```
What is the probability of the diabetes  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  cholesterol  bp  bmi  diabetes
0  0.115423            1   1    3         1
1  0.007662            1   1    3         2
2  0.860873            1   1    3         3
3  0.016043            1   1    3         4
What is the probability of the stroke  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  cholesterol  bp  bmi  stroke
0  0.082686            1   1    3       1
1  0.917314            1   1    3       2
What is the probability of the attack  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  bp  cholesterol  bmi  attack
0  0.140784   1            1    3       1
1  0.859216   1            1    3       2
What is the probability of the angina  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  cholesterol  bp  bmi  angina
0  0.161608            1   1    3       1
1  0.838392            1   1    3       2
```

(j) The probability of outcome by given poor health

If I have good health (low blood pressure, low cholesterol, and normal weight), the probability of the outcome is as below:

```
What is the probability of the diabetes  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  cholesterol  bp  bmi  diabetes
0  0.057710            2   3    2         1
1  0.009543            2   3    2         2
2  0.922194            2   3    2         3
3  0.010553            2   3    2         4
What is the probability of the stroke  if I have good health (low blood pressure, low cholesterol, and normal weight)?
     probs  bmi  cholesterol  bp  stroke
0  0.01446    2            2   3       1
1  0.98554    2            2   3       2
What is the probability of the attack  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  cholesterol  bp  bmi  attack
0  0.016161            2   3    2       1
1  0.983839            2   3    2       2
What is the probability of the angina  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  cholesterol  bp  bmi  angina
0  0.013326            2   3    2       1
1  0.986674            2   3    2       2
```

(k) The probability of outcome by given good health

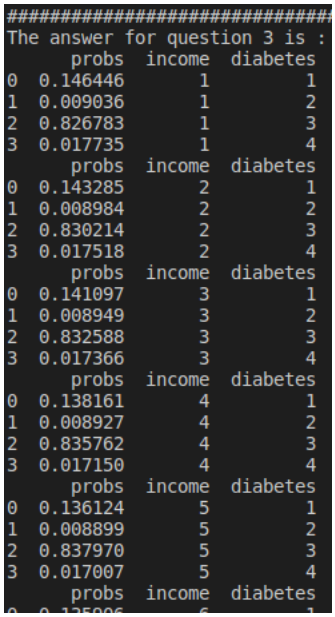
To conclude, the result in table format is as below:

		bad habits	good habits	poor health	good health
diabetes	yes	15.052%	12.712%	11.542%	5.771%
	only during pregncy	0.897%	0.887%	0.766%	0.954%
	no	82.242%	84.769%	86.087%	92.219%
	pre diabetic	1.810%	1.632%	1.604%	1.055%
stroke	yes	4.926%	3.611%	8.269%	1.446%
	no	95.074%	96.389%	91.731%	98.554%
heart attack	yes	7.433%	5.280%	14.078%	1.616%
	no	92.567%	94.720%	85.922%	98.384%
angina	yes	8.045%	5.476%	16.161%	1.333%
	no	91.955%	94.525%	83.839%	98.667%

3. *Question3*

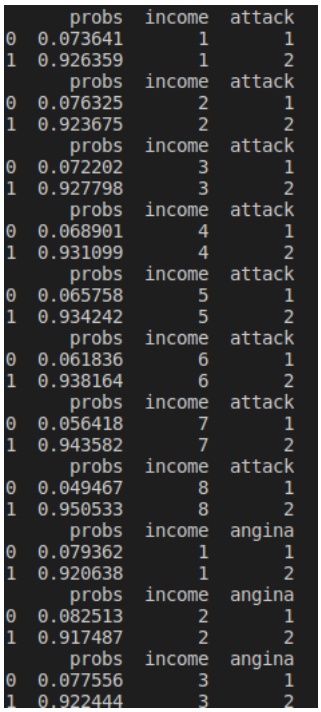
Question : Evaluate the effect a persons income has on their probability of having one of the four health outcomes (diabetes, stroke, heart attack, angina). For each of these four outcomes, plot their probability given income status (your horizontal axis should be i = 1, 2, ..., 8, and your vertical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

Answer :
The screen-shot of queries is as below:



```
##############################
The answer for question 3 is :
      probs  income  diabetes
0  0.146446       1         1
1  0.009036       1         2
2  0.826783       1         3
3  0.017735       1         4
      probs  income  diabetes
0  0.143285       2         1
1  0.008984       2         2
2  0.830214       2         3
3  0.017518       2         4
      probs  income  diabetes
0  0.141097       3         1
1  0.008949       3         2
2  0.832588       3         3
3  0.017366       3         4
      probs  income  diabetes
0  0.138161       4         1
1  0.008927       4         2
2  0.835762       4         3
3  0.017150       4         4
      probs  income  diabetes
0  0.136124       5         1
1  0.008899       5         2
2  0.837970       5         3
3  0.017007       5         4
      probs  income  diabetes
```

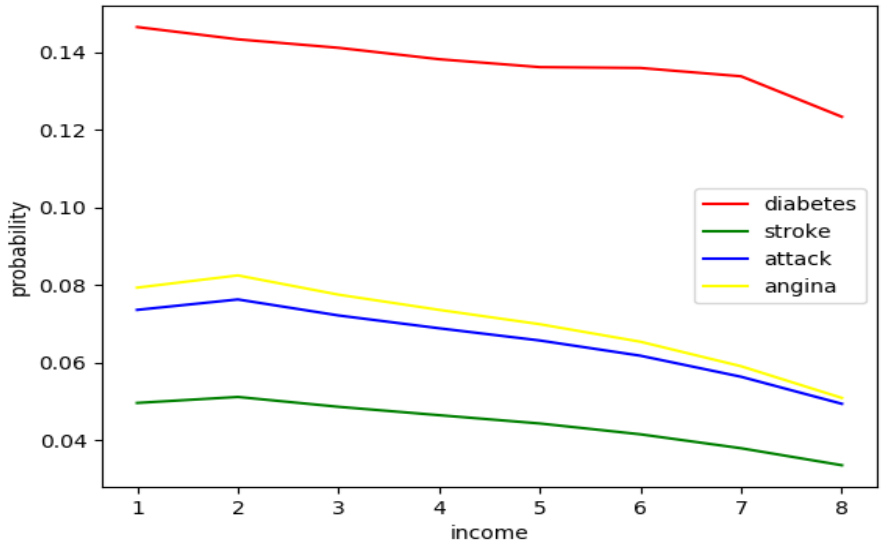
FIG. 1: partial screen-shot



```
      probs  income  attack
0  0.073641       1       1
1  0.926359       1       2
      probs  income  attack
0  0.076325       2       1
1  0.923675       2       2
      probs  income  attack
0  0.072202       3       1
1  0.927798       3       2
      probs  income  attack
0  0.068901       4       1
1  0.931099       4       2
      probs  income  attack
0  0.065758       5       1
1  0.934242       5       2
      probs  income  attack
0  0.061836       6       1
1  0.938164       6       2
      probs  income  attack
0  0.056418       7       1
1  0.943582       7       2
      probs  income  attack
0  0.049467       8       1
1  0.950533       8       2
      probs  income  angina
0  0.079362       1       1
1  0.920638       1       2
      probs  income  angina
0  0.082513       2       1
1  0.917487       2       2
      probs  income  angina
0  0.077556       3       1
1  0.922444       3       2
```

FIG. 2: partial screen-shot

To plot the probability of four outcomes given income status, we can get:



Conclusion :

(1)I found that in the wide range, the probability of each outcome is in order as 'diabetes >angina >attack >stroke'.

(2)When the income of person increases, the probability of four diseases also reduces. That's probably because people with more income could receive better treatment. Except for person whose annual income level is of 2, their probability of having four diseases is higher than those annual income level is of 1. That's might because ther have to work harder to earn more money but increases their risk of having disease while they still have no money to carefully cure.

4. Question4

Question : 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? Lets 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?

Answer :

(1)The assumption is that 'smoking and exercise have no effect on health problem'.

(2)After adding the edges from smoking and exercise to each of the four outcomes, I do the queries in Question2, the screen-shot and table format is as below:

```
########################################
The answer for question 4 is :
After adding the edges, we can get the probability as below :
What is the probability of the diabetes  if I have bad habits (smoke and don't exercise)?
      probs  exercise  smoke  diabetes
0  0.210945         2      1         1
1  0.006915         2      1         2
2  0.760693         2      1         3
3  0.021447         2      1         4
What is the probability of the stroke  if I have bad habits (smoke and don't exercise)?
      probs  exercise  smoke  stroke
0  0.078035         2      1       1
1  0.921965         2      1       2
What is the probability of the attack  if I have bad habits (smoke and don't exercise)?
      probs  exercise  smoke  attack
0  0.121166         2      1       1
1  0.878834         2      1       2
What is the probability of the angina  if I have bad habits (smoke and don't exercise)?
      probs  exercise  smoke  angina
0  0.119007         2      1       1
1  0.880993         2      1       2
```

(a) The probability of outcome by given bad habits

```
What is the probability of the diabetes  if I have good habits (don't smoke and do exercise)?
      probs  exercise  smoke  diabetes
0  0.098552         1      2         1
1  0.009884         1      2         2
2  0.877576         1      2         3
3  0.013988         1      2         4
What is the probability of the stroke  if I have good habits (don't smoke and do exercise)?
      probs  exercise  smoke  stroke
0  0.024311         1      2       1
1  0.975689         1      2       2
What is the probability of the attack  if I have good habits (don't smoke and do exercise)?
      probs  exercise  smoke  attack
0  0.031015         1      2       1
1  0.968985         1      2       2
What is the probability of the angina  if I have good habits (don't smoke and do exercise)?
    probs  exercise  smoke  angina
0  0.0368         1      2       1
1  0.9632         1      2       2
```

(b) The probability of outcome by given good habits

```
What is the probability of the diabetes  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  bmi  diabetes  cholesterol  bp
0  0.123481    3         1            1   1
1  0.007460    3         2            1   1
2  0.852416    3         3            1   1
3  0.016643    3         4            1   1
What is the probability of the stroke  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  bmi  cholesterol  bp  stroke
0  0.084257    3            1   1       1
1  0.915743    3            1   1       2
What is the probability of the attack  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  bmi  cholesterol  bp  attack
0  0.142199    3            1   1       1
1  0.857801    3            1   1       2
What is the probability of the angina  if I have poor health (high blood pressure, high cholesterol, and overweight)?
      probs  bmi  cholesterol  bp  angina
0  0.162972    3            1   1       1
1  0.837028    3            1   1       2
```

(c) The probability of outcome by given poor health

```
What is the probability of the diabetes  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  bmi  diabetes  cholesterol  bp
0  0.054173    2         1            2   3
1  0.009731    2         2            2   3
2  0.925952    2         3            2   3
3  0.010144    2         4            2   3
What is the probability of the stroke  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  bmi  cholesterol  bp  stroke
0  0.013997    2            2   3       1
1  0.986003    2            2   3       2
What is the probability of the attack  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  bmi  cholesterol  bp  attack
0  0.015469    2            2   3       1
1  0.984531    2            2   3       2
What is the probability of the angina  if I have good health (low blood pressure, low cholesterol, and normal weight)?
      probs  bmi  cholesterol  bp  angina
0  0.012944    2            2   3       1
1  0.987056    2            2   3       2
########################################
```

(d) The probability of outcome by given good health

The table is as below:

		bad habits	good habits	poor health	good health
diabetes	yes	21.095%	9.855%	12.348%	5.417%
	only during pregncy	0.692%	0.988%	0.746%	0.973%
	no	76.069%	87.758%	85.242%	92.595%
	pre diabetic	2.145%	1.399%	1.664%	1.014%
stroke	yes	7.804%	2.431%	8.426%	1.400%
	no	92.197%	97.569%	91.574%	98.600%
heart attack	yes	12.117%	3.102%	14.220%	1.547%
	no	87.883%	96.899%	85.780%	98.453%
angina	yes	11.901%	3.680%	16.297%	1.294%
	no	88.100%	96.320%	83.703%	98.706%

(3)Compare the table after adding the edges to the original table, the effect is that the probability of outcomes given habits(smoking and exercise) changed a lot, but the probability of outcomes given health(blood pressure, cholesterol and weight) changed a little. After adding the edge, people with bad habits and poor health are more intended to have diseases, people with good habits and good health are less intended to have diseases.
Therefore I think the assumption of first graph is invalid, for the probability of health problems given habits changes a lot. If the original assumption is valid, then after we add the edge, the probability of outcomes should not change a lot. Also consider that the probability of health problems given health change a little, that's because the factor related to habits is eliminated and marginalized when given health.

5. Question5

Question : 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 I diabetes = 1) and P(stroke = 1 I diabetes = 3) Again, what was the effect, and was the assumption about the interaction between diabetes and stroke valid?

Answer :
(1)The assumption is that 'the four health problems have no effect on each other'.
(2)After adding an edge from diabetes to stroke, I get the probability of stroke by given diabetes.

```
########################################
The answer for question 5 is :
After adding the edge to the BayesNetWork in question 4 , we can get the probability as below :
The factor table for P(stroke |diabetes = 1) is as below :
      probs  stroke  diabetes
0  0.044164       1         1
1  0.955836       2         1
P(stroke = 1 |diabetes = 1) =  0.04416375995893987
The factor table for P(stroke|diabetes = 3) is as below :
      probs  stroke  diabetes
0  0.040478       1         3
1  0.959522       2         3
P(stroke = 1 |diabetes = 3) =  0.04047831460537835
After adding the edge to the BayesNetWork in question 4 , we can get the probability as below :
The factor table for P(stroke |diabetes = 1) is as below :
      probs  diabetes  stroke
0  0.076198         1       1
1  0.923802         1       2
P(stroke = 1 |diabetes = 1) =  0.07619782426264214
The factor table for P(stroke|diabetes = 3) is as below :
      probs  diabetes  stroke
0  0.035015         3       1
1  0.964985         3       2
P(stroke = 1 |diabetes = 3) =  0.03501532629137385
########################################
```

(3)After adding the edge, the effect is that probability of stroke by given diabetes changes a lot. The probability of people with diabetes to have stroke increases from 4.416% to 7.620% , and the probability of people without diabetes to have stroke decreases from 4.048% to 3.502%.
The probability's change means that diatetes actually has effect on stroke, therefore the assumption about the interaction between diabetes and stroke is valid.