AI 3603 Artificial Intelligence: Principles and Techniques

By: Fan Nie (520030910001)

HW#: 3 BayesianNetworks

December 23, 2022

I. INTRODUCTION

A. Purpose

The goal of today's lab is to implement code for computing exact inferences in Bayesian networks of discrete random variables using variable elimination. Our task is to implement the code in BayesianNetworks.py and analyze risk factors for certain health problems (heart disease, stroke, heart attack, diabetes) in the writing part.

- **joinFactors(Factor1,Factor2)**

  Should return a factor table that is the join of factor 1 and 2. You can assume that the join of two factors is a valid operation. Hint: You can look up pd.merge for merging two dataframes.

- **marginalizeFactor(factorTable, hiddenVar)**

  This function should return a factor table that marginalizes margVar out of it. Assume that hiddenVar is on the left side of the conditional. Hint: you can look up pd.groupby.

- **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. You do not need to normalize the factors to be proper probabilities (no need to sum to 1).

- **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. You should 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.

B. Equipment

There is a minimal amount of equipment to be used in this lab. The few requirements are listed below:

- Python 3.8
- NumPy 1.19.5, Pandas, Matplotlib

II. CODING PART

This section consists of the description of the implementation of the functions used to inference probabilities and the result of all of the examples in BayesNetworkTestScript.py.

A. Methodology

1. joinFactors

To join the factors, we can divide the inputs into two categories. If the two factors don't have a column whose column names are the same, then we can use pd.merge(Factor1, Factor2, how='cross'). Otherwise, pd.merge(Factor1, Factor2, on=same_column) will be used. Then we can simply multiply the two probs columns and get the result 'probs' column.

2. marginalizeFactor

To marginalize the factors, we can use factorTable.groupby(column_name).sum(). The column_name is the list of all the column names that exclude 'probs' and hidden variables. Then we can use reset_index() function to reset the indexes and make the form of the dataframe the same as before.

3. evidenceUpdateNet

This function simply traverse the networks and select the network that has variables in the list of evidence variables. Then only the values of evidence variable values will be retained, and the others will be deleted.

4. inference

First, we use evidenceUpdateNet function to update the networks according to the evidence variables and their values. Then, the procedures of inference function are like what we learn in the class. We should repeatedly join the factors according to the hidden variables and marginalize them. Finally, the result should be normalized.

B. Example Results

As is presented in the figure, the results of examples using the functions above are just the same as the correct results.

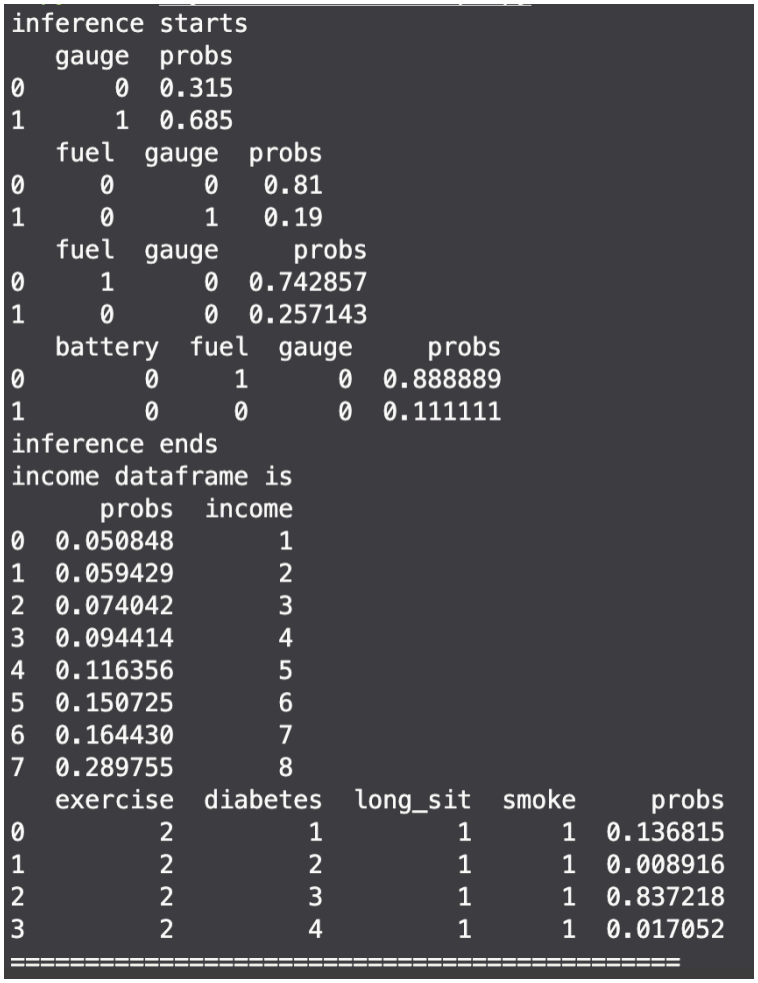

FIG. 1: The results of the examples

III. WRITTEN PART

A. Problem 1

We should first create the neworks according to the Bayesian network in the assignment.

```
# Create factors for the bayesian network
income      = readFactorTablefromData(riskFactorNet, ['income'])
smoke       = readFactorTablefromData(riskFactorNet, ['smoke', 'income'])
exercise    = readFactorTablefromData(riskFactorNet, ['exercise', 'income'])
```

```
long_sit    = readFactorTablefromData(riskFactorNet, ['long_sit', 'income'])
stay_up     = readFactorTablefromData(riskFactorNet, ['stay_up', 'income'])
bmi         = readFactorTablefromData(riskFactorNet, ['bmi', 'exercise', 'income', 'long_sit'])
bp          = readFactorTablefromData(riskFactorNet, ['bp', 'exercise', 'long_sit', 'income', \
'stay_up', 'smoke'])
cholest     = readFactorTablefromData(riskFactorNet, ['cholesterol', 'exercise', 'stay_up', \
'income', 'smoke'])
stoke       = readFactorTablefromData(riskFactorNet, ['stroke', 'bmi', 'bp', 'cholesterol'])
attack      = readFactorTablefromData(riskFactorNet, ['attack', 'bmi', 'bp', 'cholesterol'])
angina      = readFactorTablefromData(riskFactorNet, ['angina', 'bmi', 'bp', 'cholesterol'])
diabetes    = readFactorTablefromData(riskFactorNet, ['diabetes', 'bmi'])

risk_net = [income, smoke, exercise, long_sit, stay_up, bmi, diabetes, bp, cholest, stroke, \
attack, angina]
```

The number of probabilities of each factors are calculated below:
income : 8
smoke: 2*8 = 16
exercise: 2*8 = 16
long_sit: 2*8 = 16
stay_up: 2*8 = 16
bmi: 4*2*8*2 = 128
bp: 4*2*2*8*2*2 = 512
cholest: 2*2*2*8*2 = 128
stoke: 2*4*4*2 = 64
attack: 2*4*4*2 = 64
angina: 2*4*4*2 = 64
diabetes: 4*4 = 16
The total number of probabilities needed is 1048.
Alternatively, the total number of probabilities needed to store the full joint distribution is 8*2*2*2*2*4*4*4*2*2*2*2*2*4 = 131072

B. Problem 2

For each of the four health outcomes (diabetes, stroke, heart attack, angina), answer the following by querying your network (using your infer function)

1. The probability of the outcomes according to habits

1. **diabetes**

The probability of having diabetes if one has bad/good habits:

	long_sit	diabetes	exercise	stay_up	smoke	probs
0	1	1	2	1	1	0.179597
1	1	2	2	1	1	0.008754
2	1	3	2	1	1	0.791160
3	1	4	2	1	1	0.020489

(a) Bad habits

	long_sit	diabetes	exercise	stay_up	smoke	probs
0	2	1	1	2	2	0.075195
1	2	2	1	2	2	0.009409
2	2	3	1	2	2	0.903426
3	2	4	1	2	2	0.011970

(b) Good habits

FIG. 2: The probability of having diabetes if one has bad/good habits

disease	habits	outcomes	probability
diabetes	bad	yes	0.179597
		only during pregnancy	0.008754
		no	0.791160
		pre-diabetic	0.020489
	good	yes	0.075195
		only during pregnancy	0.009409
		no	0.903426
		pre-diabetic	0.011970
stroke	bad	yes	0.053214
		no	0.946786
	good	yes	0.029202
		no	0.970798
heart attack	bad	yes	0.085704
		no	0.914296
	good	yes	0.036655
		no	0.963345
angina	bad	yes	0.09542
		no	0.90458
	good	yes	0.03551
		no	0.96449

TABLE I: Problem2 - The probability of the health outcomes if one has bad habits or good habits.

According to the result, it's apparent that people with bad habits have a higher risk of having diabetes, and those with good habits have a much lower risk of diabetes.

2. **stroke**

	stroke	exercise	long_sit	stay_up	smoke	probs
0	1	2	1	1	1	0.053214
1	2	2	1	1	1	0.946786
	stroke	exercise	long_sit	stay_up	smoke	probs
0	1	1	2	2	2	0.029202
1	2	1	2	2	2	0.970798

FIG. 3: The probability of having stoke if one has bad/good habits

3. **heart attack**

	attack	exercise	long_sit	stay_up	smoke	probs
0	1	2	1	1	1	0.085704
1	2	2	1	1	1	0.914296
	attack	exercise	long_sit	stay_up	smoke	probs
0	1	1	2	2	2	0.036655
1	2	1	2	2	2	0.963345

FIG. 4: The probability of having heart attack if one has bad/good habits

disease	health	outcomes	probability
diabetes	poor	yes	0.115423
		only during pregnancy	0.007662
		no	0.860873
		pre-diabetic	0.016043
	good	yes	0.057710
		only during pregnancy	0.009543
		no	0.922194
		pre-diabetic	0.010553
stroke	poor	yes	0.082686
		no	0.917314
	good	yes	0.01446
		no	0.98554
heart attack	poor	yes	0.140784
		no	0.859216
	good	yes	0.016161
		no	0.983839
angina	poor	yes	0.161608
		no	0.838392
	good	yes	0.013326
		no	0.986674

TABLE II: Problem2 - The probability of the health outcomes if one has poor health or good health.

4. **angina**

```
   exercise  angina  long_sit  stay_up  smoke    probs
0         2       1         1        1      1  0.09542
1         2       2         1        1      1  0.90458
   exercise  angina  long_sit  stay_up  smoke    probs
0         1       1         2        2      2  0.03551
1         1       2         2        2      2  0.96449
```

FIG. 5: The probability of having angina if one has bad/good habits

2. The probability of the outcomes according to health

1. **diabetes**

	bmi	diabetes	cholesterol	bp	probs
0	3	1	1	1	0.115423
1	3	2	1	1	0.007662
2	3	3	1	1	0.860873
3	3	4	1	1	0.016043
	bmi	diabetes	cholesterol	bp	probs
0	2	1	2	3	0.057710
1	2	2	2	3	0.009543
2	2	3	2	3	0.922194
3	2	4	2	3	0.010553

FIG. 6: The probability of having diabetes if one has poor/good health

It's clear that people with good health have a much lower risk of having diabetes.

2. **stroke**

	cholesterol	bp	bmi	stroke	probs
0	1	1	3	1	0.082686
1	1	1	3	2	0.917314
	cholesterol	bp	bmi	stroke	probs
0	2	3	2	1	0.01446
1	2	3	2	2	0.98554

FIG. 7: The probability of having stoke if one has poor/good health

3. **heart attack**

	cholesterol	bp	bmi	attack	probs
0	1	1	3	1	0.140784
1	1	1	3	2	0.859216
	cholesterol	bp	bmi	attack	probs
0	2	3	2	1	0.016161
1	2	3	2	2	0.983839

FIG. 8: The probability of having heart attack if one has poor/good health

4. **angina**

	cholesterol	bp	bmi	angina	probs
0	1	1	3	1	0.161608
1	1	1	3	2	0.838392
	cholesterol	bp	bmi	angina	probs
0	2	3	2	1	0.013326
1	2	3	2	2	0.986674

FIG. 9: The probability of having angina if one has poor/good health

C. Problem 3

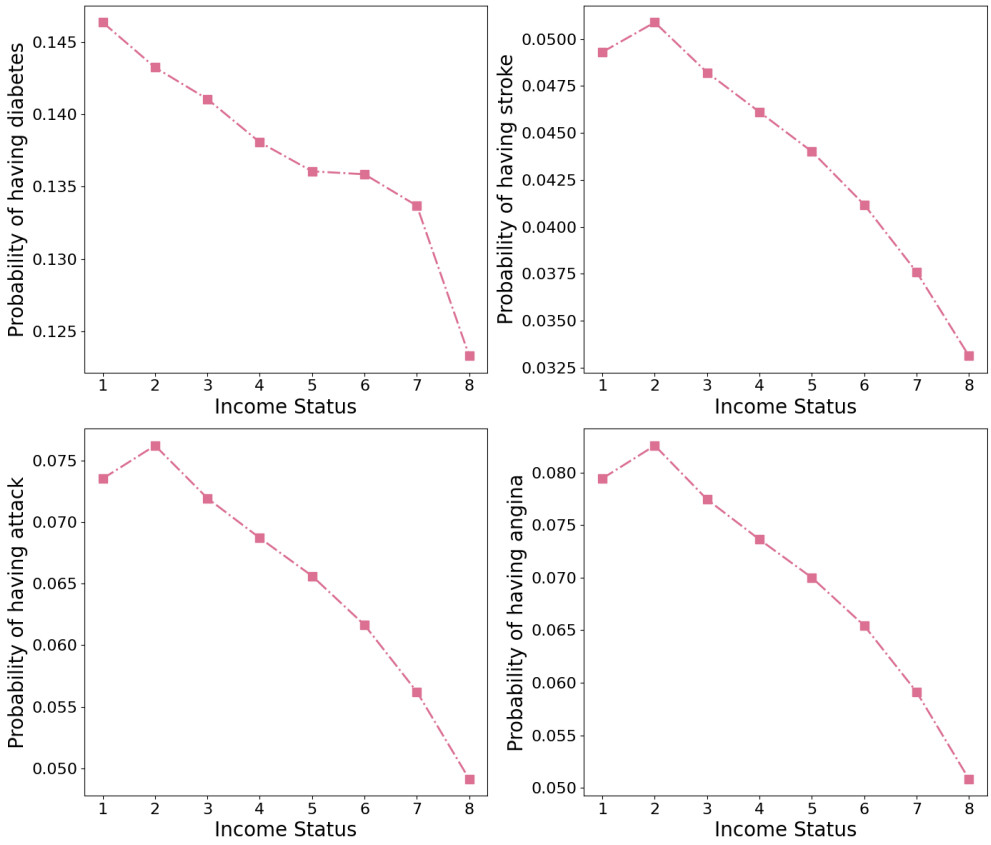

FIG. 10: The chart showing the change of a person's probability of having one of the four health outcomes with income.

According to the figures above, it can be seen that a person's probability of having the health problems, namely diabetes, stroke, heart attack and angina will decrease with the increase of income, so richer people will encounter less health problems, which conforms to common sense.

D. Problem 4

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.

```
income2      = readFactorTablefromData(riskFactorNet, ['income'])
stroke2      = readFactorTablefromData(riskFactorNet, ['stroke', 'bmi', 'bp', \
'cholesterol', 'exercise', 'smoke'])
attack2      = readFactorTablefromData(riskFactorNet, ['attack', 'bmi', 'bp', \
'cholesterol', 'exercise', 'smoke'])
angina2      = readFactorTablefromData(riskFactorNet, ['angina', 'bmi', 'bp', \
'cholesterol', 'exercise', 'smoke'])
diabetes2    = readFactorTablefromData(riskFactorNet, ['diabetes', 'bmi', \
'exercise', 'smoke'])
```

```
# build the new risk net
risk_net = [income2, smoke, exercise, long_sit, stay_up, bmi, bp, cholest, \
stroke2, attack2, angina2, diabetes2]
```

1. The probability of the outcomes according to habits

1. **diabetes**

   The probability of having diabetes if one has bad/good habits:

	stay_up	smoke	diabetes	exercise	long_sit	probs
0	1	1	1	2	1	0.245992
1	1	1	2	2	1	0.006928
2	1	1	3	2	1	0.723721
3	1	1	4	2	1	0.023359
	stay_up	smoke	diabetes	exercise	long_sit	probs
0	2	2	1	1	2	0.056227
1	2	2	2	1	2	0.010160
2	2	2	3	1	2	0.923710
3	2	2	4	1	2	0.009903

FIG. 11: The probability of having diabetes if one has bad/good habits with the edges from smoking and exercise to the outcome.

   According to the result, it's clear that people with bad habits have a higher risk of having diabetes, and those with good habits have a much lower risk of diabetes. And the probability of having diabetes conditioned on bad habits rises considerably compared to the first network.

2. **stroke**

	stay_up	stroke	smoke	exercise	long_sit	probs
0	1	1	1	2	1	0.080488
1	1	2	1	2	1	0.919512
	stay_up	stroke	smoke	exercise	long_sit	probs
0	2	1	2	1	2	0.019464
1	2	2	2	1	2	0.980536

FIG. 12: The probability of having stoke if one has bad/good habits with the edges from smoking and exercise to the outcome.

3. **heart attack**

	stay_up	attack	smoke	exercise	long_sit	probs
0	1	1	1	2	1	0.135301
1	1	2	1	2	1	0.864699
	stay_up	attack	smoke	exercise	long_sit	probs
0	2	1	2	1	2	0.021213
1	2	2	2	1	2	0.978787

FIG. 13: The probability of having heart attack if one has bad/good habits with the edges from smoking and exercise to the outcome.

4. **angina**

	angina	stay_up	smoke	exercise	long_sit	probs
0	1	1	1	2	1	0.138072
1	2	1	1	2	1	0.861928
	angina	stay_up	smoke	exercise	long_sit	probs
0	1	2	2	1	2	0.023948
1	2	2	2	1	2	0.976052

FIG. 14: The probability of having angina if one has bad/good habits with the edges from smoking and exercise to the outcome.

2. The probability of the outcomes according to health

1. **diabetes**

	bp	bmi	diabetes	cholesterol	probs
0	1	3	1	1	0.121241
1	1	3	2	1	0.007492
2	1	3	3	1	0.854769
3	1	3	4	1	0.016498
	bp	bmi	diabetes	cholesterol	probs
0	3	2	1	2	0.055937
1	3	2	2	2	0.009697
2	3	2	3	2	0.924042
3	3	2	4	2	0.010323

FIG. 15: The probability of having diabetes if one has poor/good health with the edges from smoking and exercise to the outcome.

It's clear that people with good health have a much lower risk of having diabetes.

2. **stroke**

	bp	bmi	stroke	cholesterol	probs
0	1	3	1	1	0.082697
1	1	3	2	1	0.917303
	bp	bmi	stroke	cholesterol	probs
0	3	2	1	2	0.014544
1	3	2	2	2	0.985456

FIG. 16: The probability of having stoke if one has poor/good health with the edges from smoking and exercise to the outcome.

3. **heart attack**

	bp	bmi	cholesterol	attack	probs
0	1	3	1	1	0.140083
1	1	3	1	2	0.859917
	bp	bmi	cholesterol	attack	probs
0	3	2	2	1	0.016183
1	3	2	2	2	0.983817

FIG. 17: The probability of having heart attack if one has poor/good health with the edges from smoking and exercise to the outcome.

4. **angina**

	bp	bmi	cholesterol	angina	probs
0	1	3	1	1	0.161096
1	1	3	1	2	0.838904
	bp	bmi	cholesterol	angina	probs
0	3	2	2	1	0.013328
1	3	2	2	2	0.986672

FIG. 18: The probability of having agina if one has poor/good health with the edges from smoking and exercise to the outcome.

As is presented in the figures above, the possibility of health outcomes conditioned on having poor or good health doesn't change so much and the possibility of health outcomes conditioned on having good habits doesn't change so much, either. However, the possibility of health outcomes conditioned on having bad habits rise significantly.

Therefore, due to the changeof possibilities of health outcomes based on bad habits, in my opinion, the assumption of the first graph is not valid, and the edges between habits and health outcomes are reasonable.

E. Problem 5

There are no edges between the four outcomes before. The assumption this makes is that there're no interactions between health problems

Problem 5 (without edges)

	stroke	diabetes	probs
0	1	1	0.044417
1	2	1	0.955583
	stroke	diabetes	probs
0	1	3	0.039955
1	2	3	0.960045

FIG. 19: The probabilities of $P(stroke|diabetes = 1)$ and $P(stroke|diabetes = 3)$ without an edge from diabetes to stroke.

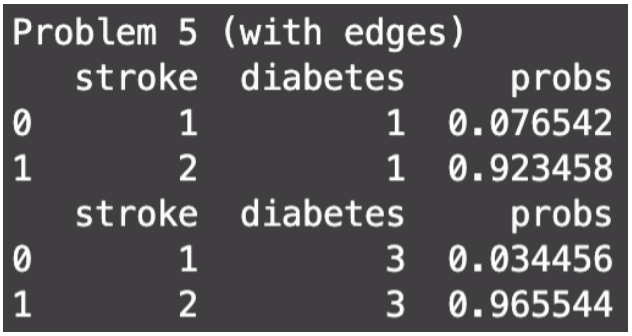

FIG. 20: The probabilities of $P(stroke|diabetes = 1)$ and $P(stroke|diabetes = 3)$ with an edge from diabetes to stroke.

As is shown in the figures above, we can see that when there's no edge between diabetes and stroke, the $P(stroke = 1|diabetes = 1)$ is 0.044417, and $P(stroke = 1|diabetes = 3)$ is 0.039955, which is very similar. However, when adding an edge between diabetes and stroke, the $P(stroke = 1|diabetes = 1)$ becomes 0.076542, and $P(stroke = 1|diabetes = 3)$ is 0.034456. We can find that in the latter situation, having diabetes will increase the probability of having stroke, which is more reasonable. Therefore, in my opinion, the assumption about the interaction between diabetes and stroke is valid.