CS6957: Probabilistic Modeling **Partial Code Due Tu 2/25**
Final Assignment Due Th 3/6

Homework 2: Bayesian Networks

Instructions: Be sure to electronically submit your answers in PDF format for the written part and as an R file for the coding part. Include all of the output of your code, plots, and discussion of the results in your written part. You may work together and discuss the problems with your classmates, but write up your final answers entirely on your own.

R Coding Part

In this assignment you will implement code for computing inferences in Bayesian networks of discrete random variables. Start by downloading the template code `BayesianNetworks.r` from the course website. Two functions are already implemented for you, `createCPT` and `createCPT.fromData`, which build conditional probability tables, represented as factors. See the source file `BayesNetExamples.r` for a demonstration of how these work. Your job is to implement the following functions:

- `productFactor(A, B)`

  This function should compute the product between two factors, `A` and `B`, and return the resulting factor. You can assume that the product of `A` and `B` is a valid operation.

- `marginalizeFactor(A, margVar)`

  This function should take a single factor `A` and marginalize the variable `margVar` from it. You can assume that marginalization of the given variable is a valid operation, i.e., that `margVar` appears on the left side of the conditional.

Turn in your source file for these two functions, `productFactor` and `marginalizeFactor`, by Tuesday, 2/25. These will be graded for correctness, but there is no need for a write-up at this stage.

- `marginalize(bayesNet, margVars)`

  This function takes a Bayesian network, `bayesNet`, and marginalizes out a list of variables, `margVars`. Do this using variable elimination, i.e., you will need to first take products of all factors involving a variable before marginalizing it.

- `observe(bayesNet, obsVars, obsVals)`

  This function takes a Bayesian network, `bayesNet`, and sets the list of variables, `obsVars`, to the corresponding list of values, `obsVals`. You do not need to normalize the factors to be probabilities.

- `infer(bayesNet, margVars, obsVars, obsVals)`

  Put it all together! This function takes in a Bayesian network and returns a single joint probability table resulting from observing a set of variables and marginalizing a set of variables. You should normalize the table to give valid probabilities.

Written Part

In this part you will be analyzing risk factors for certain health problems (heart disease, stroke, heart attack, diabetes). The data is from the 2011 Behavioral Risk Factor Surveillance System (BRFSS) survey, which is run by the Centers for Disease Control (CDC). The distilled data is in the spreadsheet `RiskFactorData.csv`. Not required, but if you are interested in the original raw data, it is here: http://www.cdc.gov/brfss/technical_infodata/surveydata/2011.htm. The variables and their meanings are as follows:

- `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.

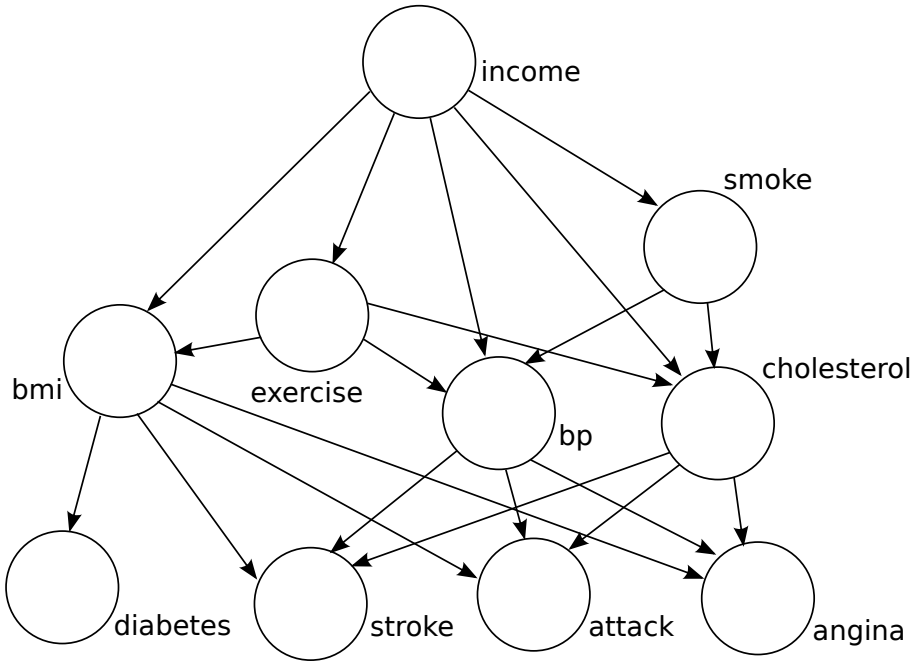
  1 (yes) 2 (no)

- `diabetes` - Had diabetes.

  1 (yes) 2 (only during pregnancy) 3 (no) 4 (pre-diabetic)

Do the following. **Make sure to turn in all of your R code and to clearly comment the code that answers each questions!** Your code should run smoothly by simply calling `source("yourfilename.r", echo = TRUE)` at the R prompt.

1. Create the following Bayesian network to analyze the survey results. You will want to use the provided function `createCPT.fromData`.



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

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

   (a) What is the 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)?

   (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)?

   Organize these results in an easy-to-read format (e.g., tables) in your write-up.

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 $P(y = 1 \,|\, \text{income} = i)$, where y is the outcome). What can you conclude?

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?

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:

$$P(\text{stroke} = 1 \,|\, \text{diabetes} = 1) \quad \text{and} \quad P(\text{stroke} = 1 \,|\, \text{diabetes} = 3)$$

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

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.