Deasises

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)

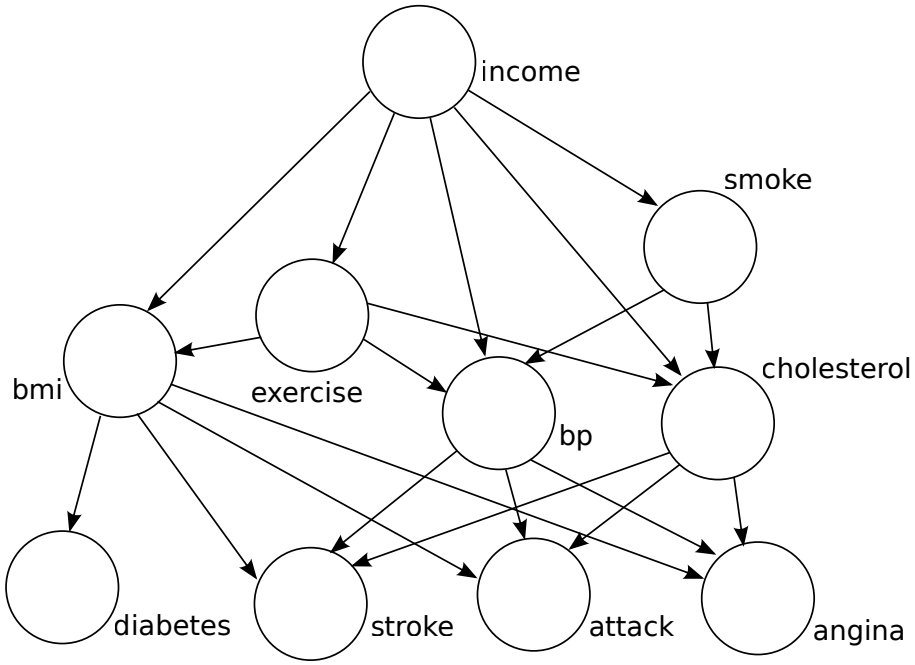
- `diabetes` - Had diabetes.

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

Do the following.

1. Create the following Bayesian network to analyze the survey results

.



   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?

.