## Homework 2: Bayesian Networks

Due date: Oct/19

1) (50 pts) Let the following random variables be associated with two common respiratory conditions: nasal congestion (N) and flu (F). A common hypothesis is that nasal congestion may be caused by allergy (A) or the flu (F). Let's consider three symptoms that can be observed related to these conditions: headache (H), runny nose (R), and fever (V). Assume that fever is directly caused only by having the flu, while runny nose and headache are directly caused only by nasal congestion. For example, having an allergy may only cause a runny nose indirectly. The random variables can be either true or false (boolean).

(i)   (ii)   (iii)   (iv)
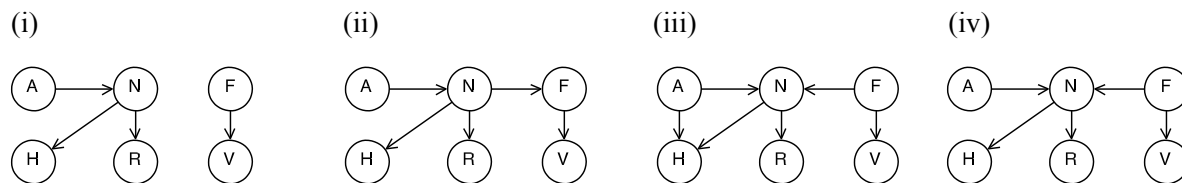
a) Which of the four networks is the best description of the problem domain as described above? Defend.

b) How many parameters are needed to define the conditional probability distribution for each node of the best Bayes network (i.e. the number of rows in its conditional probability tables CPTs)?

c) Consider the conditional independence properties for each network above. Does it have too many or too few conditional independence properties than it should have? State at least one independence that is "extra" or "missing". For example, "X should (not) be independent of Y given Z".

d) Consider the best Bayes network and that you need to compute P(A| +h, -v) using Variable Elimination. Taking the evidence into account, list the factors you will start out with. Starting out by eliminating F, what factors will you need to join and what factor will be created after performing the join, but before summing out F?
    Starting factors:
    Factors to join:
    Resulting factor:

    Tip: The answer format for the factors should include the factor number, arguments, elimination sum, and joined factors, e.g.:

$$f_1(X, +y) = \sum_z p(z| + y)p(X|z)$$

.

2) (50 pts) The Bayesian network has four variables: A, B, C, D and E, each variable has 2 states.

With the following assumptions:

- A and B are independent variables (they have no parents).

- A is the parent of C.

- A and B are the parents of D.

- C and D are the parents of E.

a) Draw the corresponding Bayesian network.

b) Express the joint probability distribution in terms of conditional probabilities in the network, i.e., write the expression for P(A, B, C, D, E).

c) Write the expression for the conditional probability of A given that you observe B=+b, C=-c, D=+d and E=-e. Just rewriting the expression found in part (b) with the states instead of variables.
P(A | +b, -c, +d, -e) =?

d) Using the BNET toolbox (https://github.com/bayesnet/bnt), implement the above Bayesian network in Matlab. There are example networks in the toolbox. Choose and assign the values in the CPTs and independent probabilities in the network. Make sure to go through the manual for BNET toolbox first.

e) Calculate P(A | +b, -c, +d, -e) based on the probabilities you assigned in part (d) using the expression you created in part (c).

f) Write an m-file to generate data using the Bayesian network you have created by sampling the network. Generate four sets of data with 50, 500, 2500, 5000, 10000 cases respectively. Compute and verify the conditional probability value you have calculated in part (e) using the probabilities you entered. Note: There is an m-file to sample a Bayesian network in the BNET toolbox.

g) Learn the structure of the network (i.e. the graph, not the parameters of the CPTs) based on 10000 data cases you have generated in part (f) using one of the structural learning algorithms available in the toolbox.

h) Describe your results implementing a Bayesian network and explain why your results are the way they are. Add any additional graphs that demonstrate your findings.

What to hand in

Upload to myCourses' dropbox two separate files: your write-up and responses to both questions in PDF, and your Matlab scripts (zipped if more than one file). If any special instructions are needed for executing your code, please include this in a readme.txt file.

Show your intermediate formulas and justify your answers. You can scan your handwritten response (or take a photo with your cellphone) instead of typing in the computer. Include your name as part of the filename.