Programming Assignment 5: Probabilistic reasoning

Problem Description

The uncertainties of the environment comprising of poor road conditions, rain, air pollution, corruption etc., causing injuries, accidents, diseases and death etc. are modeled as Bayesian network given below [Fig.1]. The accidents may occur due to fact that drivers driving in drunk state are not able to control their vehicle speed and direction. Sometimes even a person below the age of 18 years is issued a license without verifying the details. The corruption in the system is responsible for such states of driving as it allows the defaulters to misuse the system and sets them free without punishment. This eventually causes accidents. Also the rainy weather makes the roads condition bad and they become full of pits. The driver may miss the sight and bump onto the pit leading to a collision and sometimes death of people. Also air pollution is causing breathing troubles in people. People are developing diseases such as Asthma and lung cancer which may cause death. Also, the air pollution creates a haze in the atmosphere which is responsible for many accidents due to poor visibility on roads. The environment of the probabilistic reasoning agent is represented by 15 boolean variables. Many of these variables cause an effect on one or more variables. Three variables namely Corruption (O), Rains(R) and Air pollution(P) are considered to be the root causes considered for the given environment. The cause and effect relationships among the environment variables are captured in the Bayesian Network given in Fig. 1. The nodes have associated conditional probabilities given in the adjoining Conditional Probability Tables (CPT).

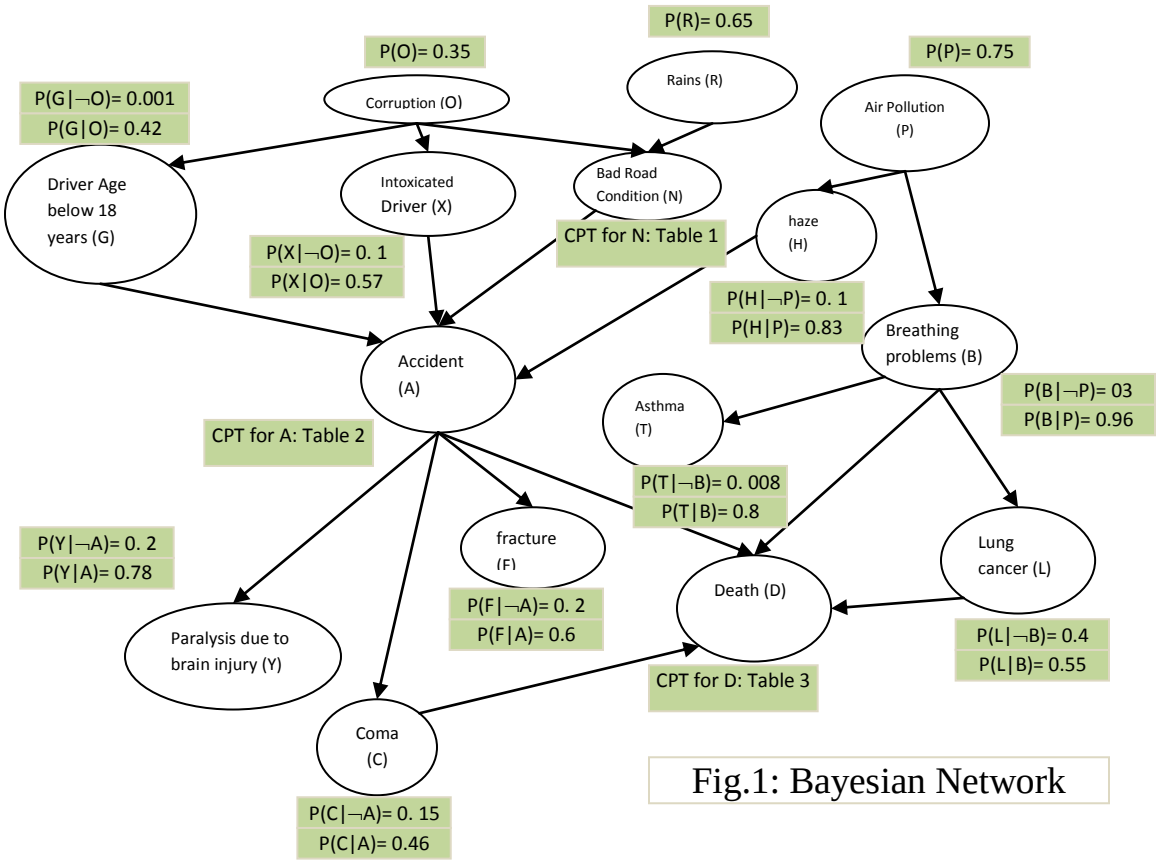


Fig.1: Bayesian Network

Table 1

P(N | ¬O, ¬R)	0.005
P(N | ¬O, R)	0.4
P(N | O, ¬R)	0.7
P(N | O, R)	0.89

Table 2

P(A|¬G, ¬X, ¬N, ¬H)	0.0001
P(A|¬G, ¬X, ¬N, H)	0.05
P(A|¬G, ¬X, N, ¬H)	0.09
P(A|¬G, ¬X, N, H)	0.25
P(A|¬G, X, ¬N, ¬H)	0.15
P(A|¬G, X, ¬N, H)	0.37
P(A|¬G, X, N, ¬H)	0.30
P(A|¬G, X, N, H)	0.65
P(A|G, ¬X, ¬N, ¬H)	0.2
P(A|G, ¬X, ¬N, H)	0.45
P(A|G, ¬X, N, ¬H)	0.30
P(A|G, ¬X, N, H)	0.32
P(A|G, X, ¬N, ¬H)	0.72
P(A|G, X, ¬N, H)	0.8
P(A|G, X, N, ¬H)	0.75
P(A|G, X, N, H)	0.9

Table 3

P(D|¬A, ¬C, ¬B, ¬L)	0.006
P(D|¬A, ¬C, ¬B, L)	0.8
P(D|¬A, ¬C, B, ¬L)	0.3
P(D|¬A, ¬C, B, L)	0.78
P(D|¬A, C, ¬B, ¬L)	0.45
P(D|¬A, C, ¬B, L)	0.1
P(D|¬A, C, B, ¬L)	0.52
P(D|¬A, C, B, L)	0.15
P(D|A, ¬C, ¬B, ¬L)	0.3
P(D|A, ¬C, ¬B, L)	0.25
P(D|A, ¬C, B, ¬L)	0.2
P(D|A, ¬C, B, L)	0.18
P(D|A, C, ¬B, ¬L)	0.02
P(D|A, C, ¬B, L)	0.3
P(D|A, C, B, ¬L)	0.61
P(D|A, C, B, L)	0.999

Input File Format

There are two sets of >> (two consecutive greater than signs) in the format used. The name (say D) immediately before the first pair (i.e. >>) depicts the environmental variable having an effect (Call it effect variable) due to the list of causes (say Cause variables) occurring immediately after the first pair of >. The list of causes is enclosed within the square bracket pair [and]. The nodes (e.g. A, C, B, L) within the pair [] are separated by commas. If the pair [] is empty then the node is itself a root and have not been modeled with any effect of other variables. There comes the second pair of > (i.e. >>) after which is the list of conditional probabilities. These conditional probabilities are separated by a blank and the number of such probabilities measures for one effect variable having effects due to 'n' cause variables is 2^n.

The file contains effect variables and their corresponding details in separate lines.

Example line from the input file input1.txt:

A >> [G, X, N, H] >> 0.0001 0.05 0.09 0.25 0.15 0.37 0.30 0.65 0.20 0.45 0.30 0.32 0.72 0.8 0.75 0.9

The above is understood as follows. The table below places data read from the above line from given input file in each row below and meaning of each character or number is explained in the adjoining columns. [Note: The following table is used only to explain the data format and is not used as a data file itself]

A	Node having an effect on		
>>	>> sign		
[	List open sign		
G	First cause variable	G	This order is important in reading conditional probabilities from file as below
,	Separator		
X	Second cause variable	X	
,	Separator		
N	Third cause variable	N	
,	Separator		
H	Fourth cause variable	H	
]	List Close sign		
>>	>> sign	The boolean variables **G, X, N and H** have respective truth values as follows	
0.0001	Conditional Probability P(A|¬G, ¬X, ¬N, ¬H)	0000	
0.05	Conditional Probability P(A|¬G, ¬X, ¬N, H)	0001	
0.09	Conditional Probability P(A|¬G, ¬X, N, ¬H)	0010	
0.25	Conditional Probability P(A|¬G, ¬X, N, H)	0011	
0.15	Conditional Probability P(A|¬G, X, ¬N, ¬H)	0100	
0.37	Conditional Probability P(A|¬G, X, ¬N, H)	0101	
0.30	Conditional Probability P(A|¬G, X, N, ¬H)	0110	
0.65	Conditional Probability P(A|¬G, X, N, H)	0111	
0.2	Conditional Probability P(A|G, ¬X, ¬N, ¬H)	1000	

0.45	Conditional Probability P(A|G, ¬X, ¬N, H)	1001
0.30	Conditional Probability P(A|G, ¬X, N, ¬H)	1010
0.32	Conditional Probability P(A|G, ¬X, N, H)	1011
0.72	Conditional Probability P(A|G, X, ¬N, ¬H)	1100
0.8	Conditional Probability P(A|G, X, ¬N, H)	1101
0.75	Conditional Probability P(A|G, X, N, ¬H)	1110
0.9	Conditional Probability P(A|G, X, N, H)	1111
$$	End of File marker	

Note: The representation $P(x_1|x_2, x_3, x_4)$ represents the conditional probability of x_1 given the joint occurrence of x_2, x_3 and x_4.

Contents of the file input1.txt

O >> [] >> 0.35
G >> [O] >> 0.001 0.42
X >> [O] >> 0.1 0.57
N >> [O, R] >> 0.005 0.4 0.7 0.89
R >> [] >> 0.65
A >> [G, X, N, H] >> 0.0001 0.05 0.09 0.25 0.15 0.37 0.30 0.65 0.20 0.45 0.30 0.32 0.72 0.8 0.75 0.9
P >> [] >> 0.75
H >> [P] >> 0.1 0.83
B >> [P] >> 0.3 0.96
T >> [B] >> 0.008 0.8
L >> [B] >> 0.4 0.55
D >> [A, C, B, L] >> 0.006 0.8 0.3 0.78 0.45 0.1 0.52 0.15 0.3 0.25 0.2 0.18 0.02 0.3 0.61 0.999
F >> [A] >> 0.2, 0.6
Y >> [A] >> 0.2 0.78
C >> [A] >> 0.15 0.46
$$

Another file named input2.txt to represent the example given in Fig, 4.2 of your text book is

B >> [] >> 0.001
E >> [] >> 0.002
A >> [B, E] >> 0.001 0.29 0.94 0.95
J >> [A] >> 0.05 0.90
M >> [A] >> 0.01 0.70
$$

Similarly you can create input file for the sprinkler example.

These file formats are important as you are going to develop a generic probabilistic reasoning agent which will take as input the file in above format and establish the cause and effect relationship to construct Bayesian network.

Markov Blanket

The Markov blanket of a node refers to its parent, children and children's parents nodes to establish the conditional independence with rest of the nodes in the network. For example, the Markov Blanket of the node B (Breathing problems) is the set of nodes B(itself), P (parent), T, D, L (children) and A, C (children's parents). Given the Markov Blanket of B, it is independent of G, O, X, R, N, H, F and Y [Fig.2]

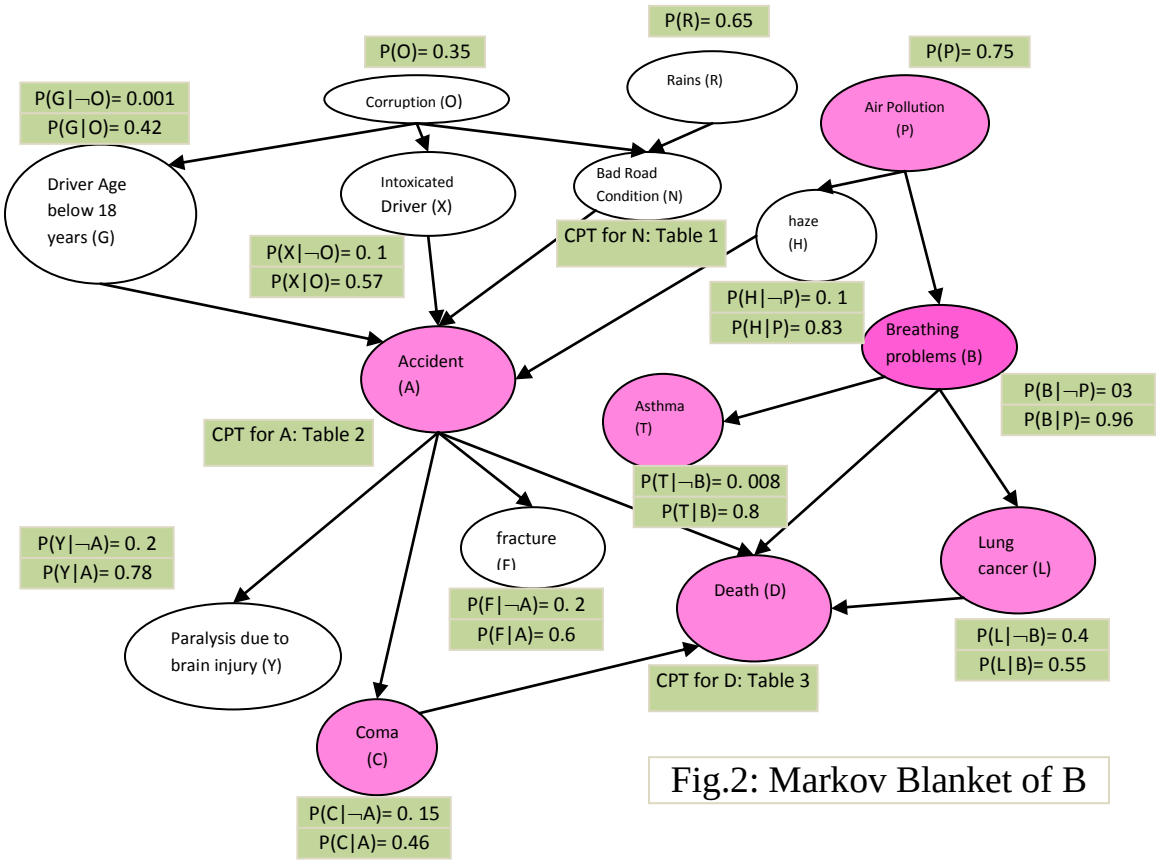


Fig.2: Markov Blanket of B

Modules

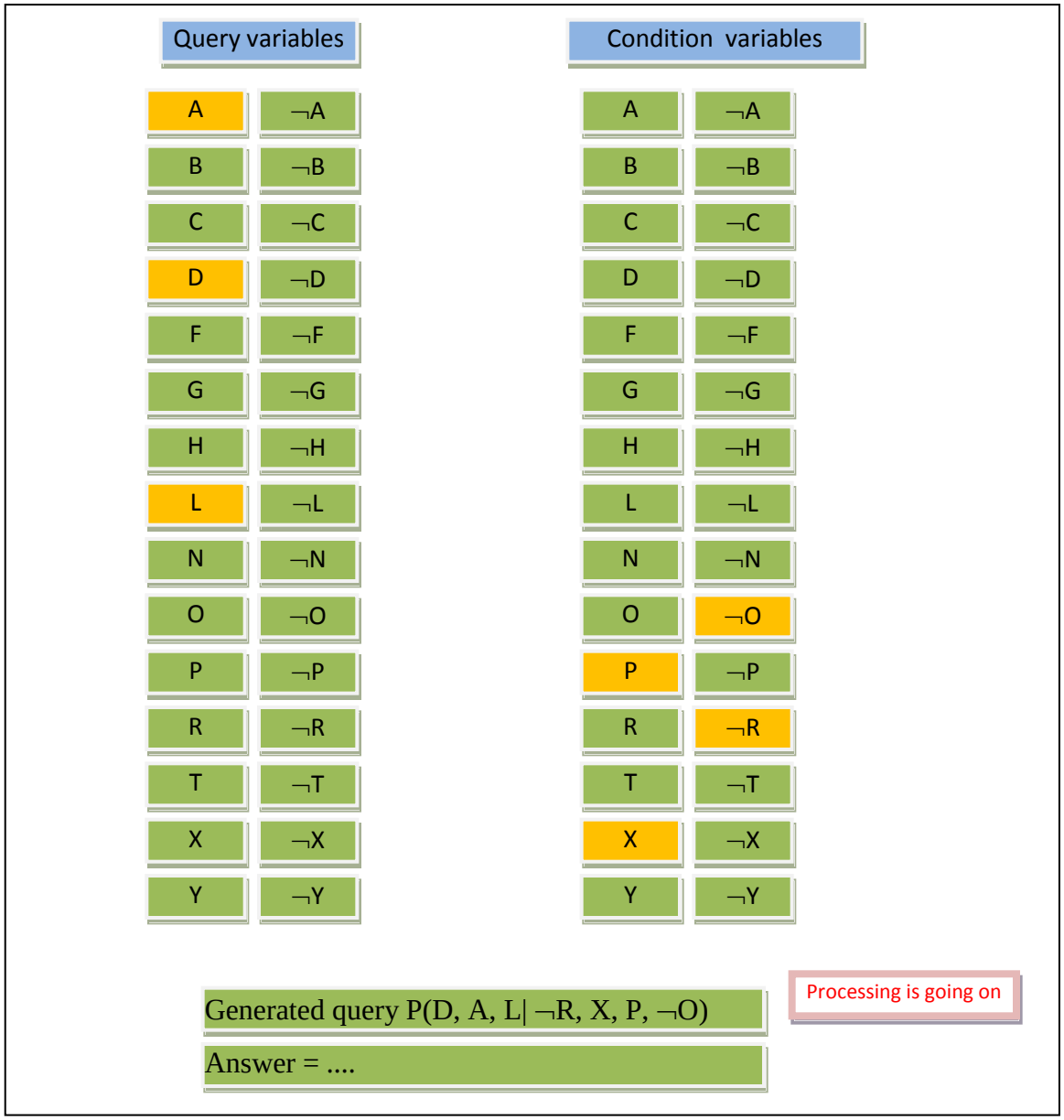
1. **BayesianNetwork createBayesianNetwork(cause_effect_file):** This function reads the data from the given file input#.txt (where # is 1, 2, 3, ..) and establishes the cause and effect relationship between variables. This uses appropriate data structure to define the node of the Bayesian network. The function creates the directed graph and returns it with its nodes and edges populated according to the data read from input file. Each CPT must also be associated with the corresponding node in the graph representing the Bayesian network.
2. **MarkovBlanket computeMarkovBlanket(BayesianNetwork B, node A):** This function returns a list of nodes in the Markov Blanket of node A in Bayesian network B.

The node A is conditionally independent of remaining nodes in B given its Markov blanket.

3. **expression createExpression(query_variables Q, condition_variables C):** This function uses the GUI as described in later text to collect the query variables and condition variables. An example expression for probabilistic query P(D, A, L| ¬R, X, P, ¬O) has the list Q of query variables D, A and L and the list C of condition variables ¬R, X, P and ¬O. An expression can be an object consisting of lists of both types of variables. If the list C is empty, the expression involves a simple joint computation further. If Q is empty, the query is reported as invalid. The minimum number of variables in Q is 1 while that of C is 0. A limit of maximum number of 10 variables in each of Q and C should be maintained right at the time of receiving the query from user through the following GUI.
4. **real_value computeProbability(MarkovBlanket M, expression E):** This function takes as input the query expression and uses its Markov blanket to construct the expressions involving joint probabilities using marginalization, product and Bayes' rules. It uses the Chain rule to compute the joint probabilities in terms of the products of conditional probabilities. The function returns the real value in [0, 1].

Graphics

The graphics involves the list of all variables (in negation as well, but restricted to only one of the two to be selected). The upper cap of the number of variables is 10 as above for each of Q and C. Also report the inability to take the same variable (positive or negative) for both Q and C. If a variable is chosen to be included as a query variable, then it should not be taken as a condition variable (in either of positive or negative form). The boxes of selected variables should be highlighted and the formed query must be presented on the screen in conventional way. The same must be taken as input for processing and the computed result be written in precision upto 20 digits (or in exponent and mantissa form appropriately) after decimal, if the answer is such.



Also provide similar interface to generate the Markov blanket of selected variables.

Driver

The driver must integrate all functionalities and execute the functions appropriately using the selections made through the above GUI.

Writeup, evaluation and submission

Write up details will be made available two days before the submission. Evaluation will be out of 15 marks (5% weight). Students are advised to inform me immediately, if any discrepancy exists in this document. The assignment is due for submission on November 28, 2017 (Tuesday) by

7:00 p.m. The students are expected to read the text book chapters 13 and 14 thoroughly and clarify all their doubts pertaining to the problem specification, explanations given above, conceptual understanding, doubts related to individual problem or the data structures to be used, and the doubts related to other aspects of implementation.

Please feel free to meet me and discuss your doubts.

Vandana
November 17, 2017