Computer Science 384
St. George Campus

Revised March 27, 2014
University of Toronto

Homework Assignment #3: Bayes Networks
Due: Friday April 4, 2014 by 11:59 PM
You may use Grace Days to extend this due date.

Silent Policy: *A silent policy will take effect 12 hours before this assignment is due, i.e. no question about this assignment will be answered, whether it is asked via e-mail, or in person.*

Late Policy: 15% per day after the use of any remaining grace days you may have. Recall that you started the term with 2 grace days and may have received 1 additional grace day from completely Part B of Assignment 2.

Total Marks: This part of the assignment is out of 50 marks and represents 12% of the course grade.

Handing in this Assignment

What to hand in on paper: Nothing.

What to hand in electronically: See individual questions for directions. Please submit your assignment electronically as either a **.pdf**, **.ps** or as a **text** file. **Be sure to write your name and student number at the beginning of your submission.**

To submit your file electronically, use the CDF secure Web site:
`https://www.cdf.utoronto.ca/students`
or use the CDF **submit** command. Type **man submit** for more information.
Warning: marks will be deducted for incorrect submissions.

Clarification Page: Important corrections (hopefully few or none) and clarifications to the assignment will be posted on the Assignment 3 Clarification page, linked from the CSC384 A3 web page. You are responsible for monitoring the A3 Clarification page.

Questions: Questions concerning the assignment should be directed by email to the Assignment 3 TA, Andrew Perrault (t1perrau@cdf.toronto.edu). Please place "384" and "A3" in your email subject header.

1 Credit Card Fraud Detection

For this problem, you will develop a simple Bayes Network that models whether a given credit card transaction is fraudulent. Every time a credit card number is supplied to make a purchase somewhere in the world, all available data is first forwarded to your server, and your system must quickly estimate the probability that the transaction is fraudulent rather than valid. If your system suspects that the card is stolen, then the purchase will be declined and the account suspended.

The data comes in the form of the first seven variables listed below; for a given transaction they might not all be available, but as you know this will not be a problem for your completed Bayes Net. An eighth variable, `Fraud` $\in$ {`yes, no`}, will certainly not be available; the task is to infer its likelihood given the other variables.

- `Signature` $\in$ {`good, bad`}: If someone is using the card fraudulently, they may have trouble reproducing the true user's signature. Vendors equipped with electronic readers can scan the buyer's signature and compare with records, reporting that it is either a good or a bad match.
- `Bulk` $\in$ {`yes, no`}: Fraudulent purchases are more likely to consist of bulk orders of the same item, for instance the purchase of 20 digital cameras for later resale.
- `Purchase` $\in$ {`electronics, groceries, misc`}: The purchase of certain categories of goods are more or less likely based on whether the transaction is fraudulent.
- `Delivery` $\in$ {`yes, no`}: Stolen credit card numbers are actually more likely to be used for online (or phone/catalog) orders. So if a transaction is fraudulent, there is a higher chance that it is for goods that are to be delivered.
- `Location` $\in$ {`far, near`}: We can compare the reported location of the vendor to the registered user's billing address. If the number is being used fraudulently, it is more likely to appear in a transaction far from the user's home. But, the distance also depends on whether the transaction is for delivery; in such cases it is already quite likely that the vendor is located in a distant city. A third possibility is that the card user is merely taking a trip, as described explained below.
- `Journey` $\in$ {`yes, no`}: This variable represents whether the card user is taking a trip, as this would cause any transactions to occur far from home. The value of this variable is not known. We do know from prior experience, though, that 10% of credit card transactions occur while the cardholder is travelling.
- `Transportation` $\in$ {`yes, no`}: One aide in determining whether the legitimate cardholder is travelling is to check whether their previous purchases involve transportation-type expenses. That is, being on a journey would typically cause the previous few transactions to include items like tickets, gas, or lodging.
- `Fraud` $\in$ {`yes, no`}: This is the target variable whose value we would typically like to estimate. One useful fact is that based on prior experience, 1% of all attempted transactions are known to be fraudulent.

Question 1. (20 marks) Constructing the Network.

Construct a Bayes network linking the above eight variables. The structure of the network should reflect the dependencies described above, but you will provide your own probabilities for filling out the conditional probability tables. Note that the specification gives prior probabilities for exactly two of the variables. As such, these are the two "top" variables that will have no parents in your network.

Please submit:

a) A drawing of the network. Make sure that the arcs correspond to the dependencies described above. [5 marks].

b) Conditional probability tables for the various variables. [10 marks].

c) For each variable, a brief (1-3 sentences) description or justification for the selected values in the CPT. [5 marks].

Below is an example conditional probability table for the `Signature` variable, along with its justification. Its single parent in the network is the `Fraud` variable; you will draw the overall network and then provide the same information for the remaining seven variables.

```
P(Signature|Fraud):

            |  S = good  S = bad
--------------------------------
  F = yes   |    .90       .10
  F = no    |    .95       .05

Provided signatures are generally accurate, but they are more likely
to be off when the card is being used fraudulently.
```

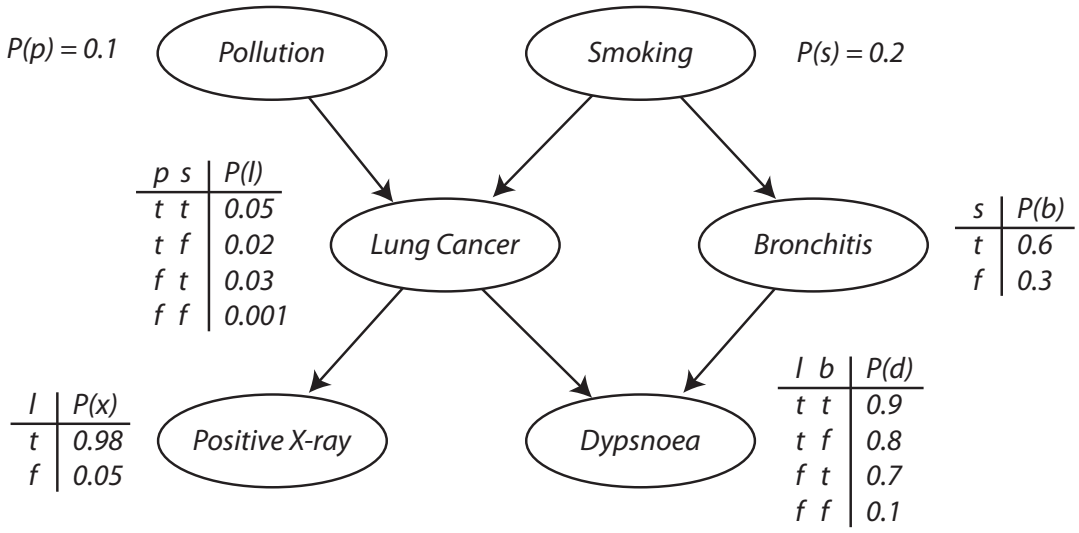
Question 2. (6 marks) Using the Network.

Please write out any numerical calculations by hand, **showing your work**.

a) For each variable, state whether it is an evidence variable (something we observe), query (something we want to know), or neither (something that helps build a good network). [4 marks]

b) Is `Delivery` independent of `Journey`, given no other information? [1 mark].

c) Is `Signature` independent of `Bulk`, given the value of `Fraud`? [1 mark].



2 Medical Diagnosis

The questions in this problem refer to the Bayes Net above, which can estimate the probability that a patient has lung cancer or bronchitis given their patient histories, whether they are exhibiting dypsnoea (shortness of breath) and potentially, whether a chest X-ray showed a tumor.

This network could be used to decide whether it is economical to perform a chest X-ray on patients that exhibit a certain set of symptoms or have a particular background.

Question 3. (14 marks) Using the Network.

Please write out any numerical calculations by hand, **showing your work**.

a) Is `Smoking` independent of `Pollution`, given `Positive X-ray`? [1 mark].

b) Is `Smoking` independent of `Dypsnoea`, given `Bronchitis`? [1 mark].

c) What is the value of `P(Bronchitis=true | Pollution=false, Smoking=true, Dypsnoea=true)`? [2 marks].

d) What is the value of `P(Lung Cancer=true | Pollution=false, Smoking=true, Dypsnoea=true)`? [2 marks].

e) What is the value of `P(Lung Cancer=true | Pollution=false, Smoking=false, Positive X-ray=true, Dypsnoea=false)`? [2 marks].

f) What is the marginal probability that a x-ray is positive? That is, what is the value of `P(Positive X-ray=true)`? [2 marks].

g) Describe a part of this Bayes Net that exhibits "explaining away", i.e., knowing a symptom S increases the probability of its causes $C_1, \ldots, C_k$ but, subsequently, increasing the probability of one of those causes C_i decreases the probability of the other causes as it explains symptom S. [4 marks]

Question 4. (10 marks) Variable Elimination.

Suppose that we do not have an X-ray machine and we are no longer interested in Bronchitis, i.e., the values of `Positive X-ray` and `Bronchitis` are always unknown to us. We want to alter the network so that it no longer contains those variables.

One temptation is to simply erase the two variables from our network. But this loses some of our prior information. Use marginalization (i.e. Variable Elimination) to remove the `Positive X-ray` and `Bronchitis` variables from the network, drawing a new network and updating the CPT of any affected variable. Eliminate the `Positive X-ray` variable first. [10 marks].

GOOD LUCK!