STA 247 — Assignment #1, Due in class on October 23

Late assignments will be accepted only with a valid medical or other excuse.

This assignment is to be done by each student individually. You may discuss it in general terms with other students, but the work you hand in should be your own. Handing in work that is not your own is a serious academic offense. Fabricating results, such handing in fake output that was not actually produced by your program, is also an academic offense.

Part I

For the first two questions, you are to prove some laws of probability using only the basic axioms of probability (on page 8), the definitions of conditional probability, independence, and mutual independence, and the properties of sets. You may also use Probability Law 1 (that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$), but not other laws given in the textbook.

1) Suppose that A, B and C are events in some sample space, S.

   a) Prove that if $P(A \cap C) > 0$ and $P(B \cap C) > 0$, then

   $$P(A|B \cap C) \quad = \quad \frac{P(B|A \cap C)P(A|C)}{P(B|C)}$$

   b) Prove that if $P(C) > 0$, then $P(A|C) = P(A \cap B|C) + P(A \cap B^c|C)$.

2) Suppose that A, B, and C are mutually independent events in some sample space, S.

   a) Prove that $A \cap B$ and C are independent.

   b) Prove that $A \cup B$ and C are independent.

   Hint: Remember the distributive laws for union and intersection: For any sets E, F, and G, $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$ and $E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$.

For the next two questions, you must produce an actual numerical answer (a simple fraction or a decimal number), not just a formula for the answer. You must also justify your answer, using the basic axioms of probability, or any of the laws of probability stated in Chapter 1 of the textbook. You may also use the results of questions (1) and (2) above (even if you haven't solved them),

3) You have three urns containing red and black balls. Urn 1 initially contains 3 red and 4 black balls. Urn 2 intially contains 3 red and 2 black balls. Urn 3 initially contains 2 red and 2 black balls. You randomly select a ball from Urn 1 and place it in Urn 2. You then select a ball at random from Urn 2 and place it in Urn 3. Finally, you select a ball at random from Urn 3. You see that this ball is black, but you didn't look at the balls you selected earlier. Given this, what is the probability that the ball you took from Urn 1 was black?

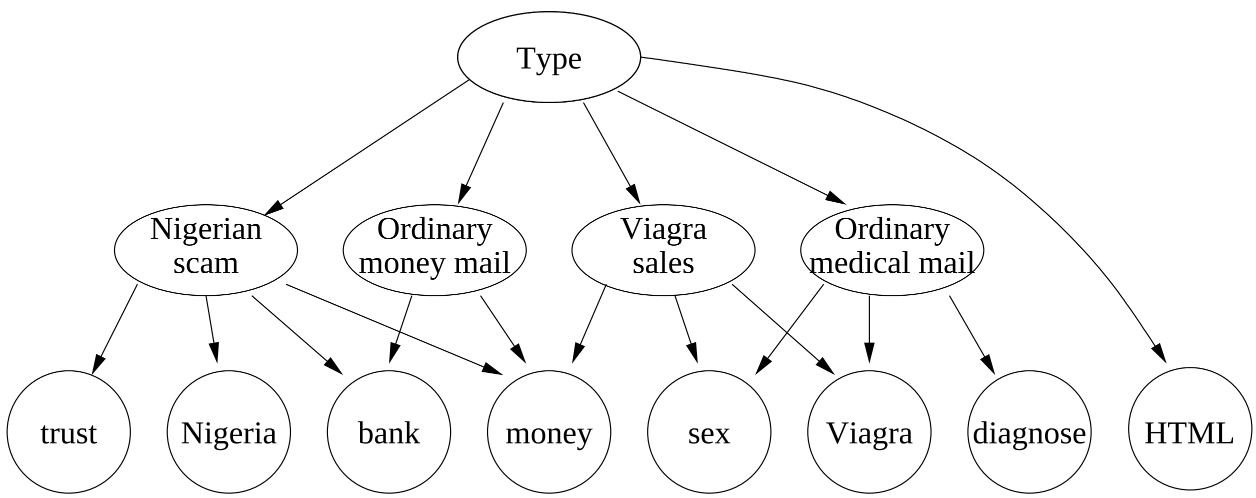
4) A computer has been set up with two redundant disk drives, called drive 1 and drive 2, with all data being "mirrored" on both drives, so that if one fails the computer can continue to operate without interruption and without loss of data. Computers can also fail for other reasons, of course. Suppose that drive 1 working correctly, drive 2 working correctly, and the rest of the computer working correctly are mutually independent events. Suppose also that, on this day, the probability of drive 1 failing is 0.01, the probability of drive 2 failing is 0.02, and the probability of some other failure is 0.03. Find the probability that the computer will fail to operate this day (because either both drives fail or something else goes wring).

Part II

Unwanted e-mail — commonly called "spam" — is now a major problem. Many people use "spam filters" to automatically discard e-mail that looks like spam. Some of these filters use a probabilistic model to determine how likely a piece of e-mail is to be spam, and then discard it if the probability of it being spam is sufficiently high (eg, greater than 0.99, or whatever level the user thinks is appropriate).

Here, we'll consider a simple probabilistic model for e-mail, that incorporates some of the common characteristics of spam. (A real spam filter would need a more complex model.) Your task will be to write an R program to simulate generation of e-mail according to this model, and to use this simulation to determine how likely it is that a piece of e-mail with certain characteristics is spam.

The model is based on a "causal network" (also called a "belief network") that shows how each random variable depends on its "parent" random variables. Here is a picture of the model for an e-mail message:



The top node represents the **Type** random variable, which has six possible values:

Type $= -1$	Spam about Nigerian money transfers
Type $= -2$	Spam offering to sell Viagra cheaply
Type $= -3$	Some other spam message
Type $= 1$	Ordinary (non-spam) e-mail about money
Type $= 2$	Ordinary (non-spam) e-mail on a medical topic
Type $= 3$	Some other message that isn't spam

The four nodes below this have possible values of 0 and 1, representing whether or not the message is about each of four possible topics (1 means it's about this topic, 0 that it isn't about this topic). The **Nigeria scam** random variable is 1 if the message is about offers to make money from a scam involving money in Nigeria; it is 1 if Type$= -1$, and otherwise is 1 with only a small probability. The **Viagra sales** random variables is 1 if the message is about sales of Viagra; it is 1 if Type$=-2$, and otherwise has only a small probability of being 1. The **Ordinary money mail** and **Ordinary medical mail** variables are 1 if Type=1 or Type=2, respectively, and are 1 with only a small probability for other types.

The bottom row of random variables represent things we can find out about the message. They all have values of 0 or 1. The first seven random variables indicate whether each of the words "trust", "Nigeria", "bank", "money", "sex", "Viagra", and "diagnose" appear in the message, with the value 1 indicatating that the word does appear. The **HTML** random variable is 1 if the message is in HTML format, rather than plain text. (HTML is used more often in spam messages.)

The model needs to specify the joint probability mass function for all possible combinations of values for these thirteen random variables. There are $6 \times 2^{12} = 24576$ such possible combinations, so we'd rather not specify the probability of each combination of values separately. Instead, we use the fact that we can always write the joint probability for a set of random variables —eg, W, X, Y, and Z — as the following product, once we've chosen some order for them:

$$
\begin{aligned}
&P(W=w,\, X=x,\, Y=y,\, Z=z) \\
&\quad = \;\; P(W=w)\,P(X=x\,|\,W=w)\,P(Y=y\,|\,W=w,\, X=x)\,P(Z=z\,|\,W=w,\, X=x,\, Y=y)
\end{aligned}
$$

If we can furthermore simplify some of the factors on the right, we may be able to specify the model using many fewer numbers.

When the model is specified using a causal network, we order the variables so that all the arrows go forward (top to bottom in this case), and then use conditional probabilities that are conditional only on the "parents" of a variable. A variable X is a parent of variable Y if there is an arrow from X to Y. For instance, in the network above, we can do the following simplification:

$$
\begin{aligned}
&P(\text{bank}=1\,|\,\text{Type}=1,\ \text{Nigerian\_scam}=0,\ \text{Ordinary\_money\_mail}=1, \\
&\qquad\qquad \text{Viagra\_sales}=0,\ \text{Ordinary\_medical\_mail}=0,\ \text{trust}=0,\ \text{Nigeria}=0) \\
&\qquad\qquad\qquad = \;\; P(\text{bank}=1\,|\,\text{Nigerian\_scam}=0,\ \text{Ordinary\_money\_mail}=1)
\end{aligned}
$$

On the last page of this assignment, all the probabilities needed to specify the joint distribution for the variables in this causal network are given. You are to use the network and these probabilities to write an R function called `generate_email`, which takes no arguments, and which returns a randomly generated set of values for all the random variables in the network. These values should be returned as an R "list" with thirteen elements, named `Type`, `Nigerian_scam`, etc.

You should then write an R function called `spam_probability`, which takes as arguments the values of the eight random variables in the bottom row of the network, and returns an estimate of the conditional probability that the message is spam (ie, that **Type** is less than zero) given that the bottom eight random variables have the values specified. This is the procedure that the spam filter would use to decide whether to discard the e-mail, by comparing this probability with the threshold set by the user (eg, 0.99).

You should implement the `spam_probability` function by randomly generating many e-mail messages using `generate_email`. You should count how many of these messages match the values of the eight known random variables, and of these matching messages, how many are spam. The ratio of these numbers gives an estimate of the conditional probability that the actual message is spam. You should continue generating messages until *either* you have generated 10000 messages in total, *or* you have generated 1000 messages that match the known values of the eight random variables.

You should hand in (on paper) the following:

a) A listing of your R program, with suitable (but not excessive) comments.

b) The output of five calls of your `generate_email` function.

c) The output of your `spam_probability` function when called with the following six sets of arguments (values in order from left to right):

    0 0 0 0 0 0 0 0
    0 1 0 0 0 0 0 0
    1 1 0 1 0 0 0 1
    0 0 1 1 1 0 0 0
    0 0 0 1 1 1 0 1
    0 0 0 0 1 1 1 0

d) A brief discussion of how much the answers you got for (c) vary when you run your function again (without resetting the random number seed).

Here are the probabilities you will need:

$P(Type = -1) = 0.11, \quad P(Type = -2) = 0.14, \quad P(Type = -3) = 0.4$
$P(Type = 1) = 0.02, \quad P(Type = 2) = 0.02, \quad P(Type = 3) = 0.31$

$P(Nigerian\_scam = 1 \,|\, Type = -1) = 1$
For $v \neq -1$, $P(Nigerian\_scam = 1 \,|\, Type = v) = 0.001$

$P(Ordinary\_money\_mail = 1 \,|\, Type = 1) = 1$
For $v \neq 1$, $P(Ordinary\_money\_mail = 1 \,|\, Type = v) = 0.001$

$P(Viagra\_sales = 1 \,|\, Type = -2) = 1$
For $v \neq -2$, $P(Viagra\_sales = 1 \,|\, Type = v) = 0.001$

$P(Ordinary\_medical\_mail = 1 \,|\, Type = 2) = 1$
For $v \neq 2$, $P(Ordinary\_medical\_mail = 1 \,|\, Type = v) = 0.001$

$P(trust = 1 \,|\, Nigerian\_scam = 0) = 0.03, \quad P(trust = 1 \,|\, Nigerian\_scam = 1) = 0.4$

$P(Nigeria = 1 \,|\, Nigerian\_scam = 0) = 0.02, \quad P(Nigeria = 1 \,|\, Nigerian\_scam = 1) = 0.9$

$P(bank = 1 \,|\, Nigerian\_scam = 0, Ordinary\_money\_mail = 0) = 0.1$
$P(bank = 1 \,|\, Nigerian\_scam = 1, Ordinary\_money\_mail = 0) = 0.5$
$P(bank = 1 \,|\, Nigerian\_scam = 0, Ordinary\_money\_mail = 1) = 0.5$
$P(bank = 1 \,|\, Nigerian\_scam = 1, Ordinary\_money\_mail = 1) = 0.5$

$P(money = 1 \,|\, Nigerian\_scam = 0, Ordinary\_money\_mail = 0, Viagra\_sales = 0) = 0.1$
$P(money = 1 \,|\, Nigerian\_scam = 0, Ordinary\_money\_mail = 0, Viagra\_sales = 1) = 0.7$
$P(money = 1 \,|\, Nigerian\_scam = 0, Ordinary\_money\_mail = 1, Viagra\_sales = 0) = 0.7$
$P(money = 1 \,|\, Nigerian\_scam = 0, Ordinary\_money\_mail = 1, Viagra\_sales = 1) = 0.7$
$P(money = 1 \,|\, Nigerian\_scam = 1, Ordinary\_money\_mail = 0, Viagra\_sales = 0) = 0.7$
$P(money = 1 \,|\, Nigerian\_scam = 1, Ordinary\_money\_mail = 0, Viagra\_sales = 1) = 0.7$
$P(money = 1 \,|\, Nigerian\_scam = 1, Ordinary\_money\_mail = 1, Viagra\_sales = 0) = 0.7$
$P(money = 1 \,|\, Nigerian\_scam = 1, Ordinary\_money\_mail = 1, Viagra\_sales = 1) = 0.7$

$P(sex = 1 \,|\, Viagra\_sales = 0, Ordinary\_medical\_mail = 0) = 0.02$
$P(sex = 1 \,|\, Viagra\_sales = 1, Ordinary\_medical\_mail = 0) = 0.6$
$P(sex = 1 \,|\, Viagra\_sales = 0, Ordinary\_medical\_mail = 1) = 0.03$
$P(sex = 1 \,|\, Viagra\_sales = 1, Ordinary\_medical\_mail = 1) = 0.6$

$P(Viagra = 1 \,|\, Viagra\_sales = 0, Ordinary\_medical\_mail = 0) = 0.001$
$P(Viagra = 1 \,|\, Viagra\_sales = 1, Ordinary\_medical\_mail = 0) = 0.99$
$P(Viagra = 1 \,|\, Viagra\_sales = 0, Ordinary\_medical\_mail = 1) = 0.03$
$P(Viagra = 1 \,|\, Viagra\_sales = 1, Ordinary\_medical\_mail = 1) = 0.99$

$P(diagnose = 1 \,|\, Ordinary\_medical\_mail = 0) = 0.01$
$P(diagnose = 1 \,|\, Ordinary\_medical\_mail = 1) = 0.4$

For $v < 0$, $P(HTML = 1 \,|\, Type = v) = 0.6$
For $v > 0$, $P(HTML = 1 \,|\, Type = v) = 0.2$