

# CS 340 Fall 2007: Homework 5

## 1 Bayes Ball

Here we compute some global independence statements from some directed graphical models. You can use the “Bayes ball” algorithm, the d-separation criterion, or the method of converting to an undirected graph (all should give the same results).

1. Consider the DAG in Figure 1(a). List all variables  $X$  s.t.,  $X \perp A|B$ , i.e., variables that are independent of  $A$  given evidence on  $B$ .
2. Consider the DAG in Figure 1(b). List all variables  $X$  s.t.,  $X \not\perp A|J$ , i.e., variables which depend on  $A$  given evidence on  $J$ .

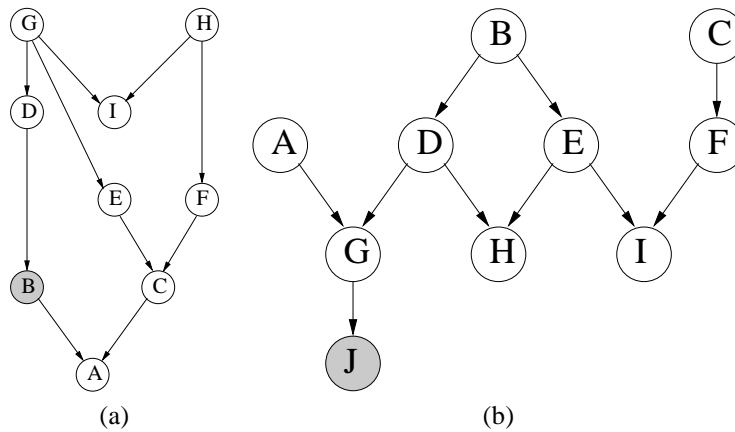
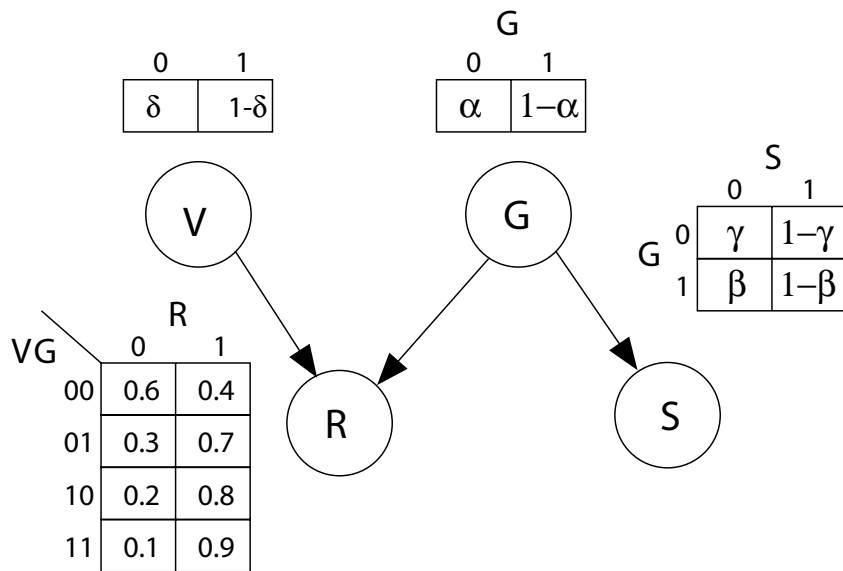


Figure 1: Bayes nets

## 2 Bayes nets for a rainy day

In this question you must model a problem with 4 binary variables:  $G$  = "gray",  $V$  = "Vancouver",  $R$  = "rain" and  $S$  = "sad". You are given the following graphical model describing the relationship between these variables:



1. Write down an expression for  $P(S = 1|V = 1)$  in terms of  $\alpha, \beta, \gamma, \delta$ .
2. Write down an expression for  $P(S = 1|V = 0)$ . Is this the same or different to  $P(S = 1|V = 1)$ ? Explain why.
3. Find maximum likelihood estimates of  $\alpha, \beta, \gamma$  using the following data set, where each row is a training case. (You may state your answers without proof.)

<i>V</i>	<i>G</i>	<i>R</i>	<i>S</i>	
1	1	1	1	(1)
1	1	0	1	
1	0	0	0	

### 3 Fishing nets

Consider the Bayes net shown in Figure 2. Here, the nodes represent the following variables

- $X_1 \in \{\text{winter, spring, summer, autumn}\}$
- $X_2 \in \{\text{salmon, sea bass}\}$
- $X_3 \in \{\text{light, medium, dark}\}$
- $X_4 \in \{\text{wide, thin}\}$

The corresponding conditional probability tables are

$$p(x_1) = ( .25 \ .25 \ .25 \ .25 )$$

$$p(x_2|x_1) = \begin{pmatrix} .9 & .1 \\ .3 & .7 \\ .4 & .6 \\ .8 & .2 \end{pmatrix}$$

$$p(x_3|x_2) = \begin{pmatrix} .33 & .33 & .34 \\ .8 & .1 & .1 \end{pmatrix}$$

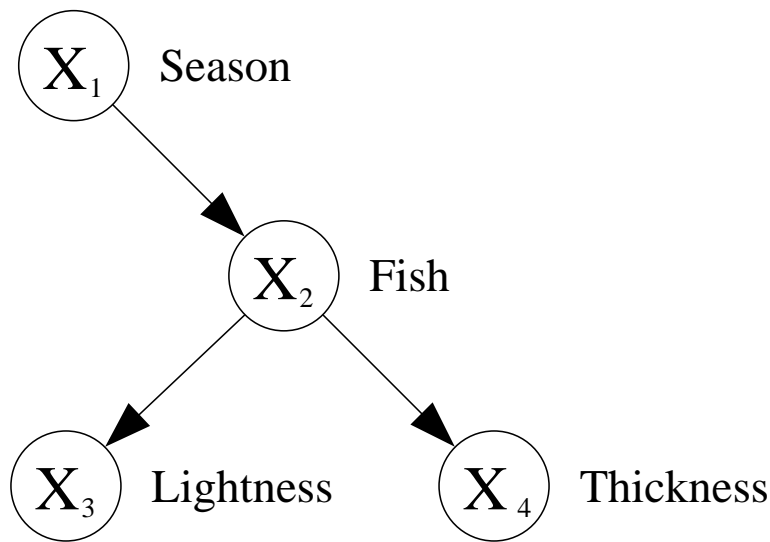


Figure 2: Fish Bayes net

$$p(x_4|x_2) = \begin{pmatrix} .4 & .6 \\ .95 & .05 \end{pmatrix}$$

Note that in  $p(x_4|x_2)$ , the rows represent  $x_2$  and the columns  $x_4$  (so each row sums to one and represents the child of the CPD). Thus  $p(x_4 = \text{thin}|x_2 = \text{sea bass}) = 0.05$ ,  $p(x_4 = \text{thin}|x_2 = \text{salmon}) = 0.6$ , etc.

Answer the following queries. You may use matlab or do it by hand. In either case, show your work.

1. Suppose the fish was caught on December 20 — the end of autumn and the beginning of winter — and thus let  $p(x_1) = (.5, 0, 0, .5)$  instead of the above prior. (This is called **soft evidence**, since we do not know the exact value of  $X_1$ , but we have a distribution over it.) Suppose the lightness has not been measured but it is known that the fish is thin. Classify the fish as salmon or sea bass.
2. Suppose all we know is that the fish is thin and medium lightness. What season is it now, most likely? Use  $p(x_1) = (.25 \ .25 \ .25 \ .25)$

## 4 Learning to use BNT

BNT (the Bayes Net Toolbox) is a Matlab package I wrote that makes it easy to do Bayesian inference about discrete random variables in directed graphical models with fixed parameters. In this exercise, you will learn how to use some of its most basic functions.

1. Download BNT.zip from <http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>.
2. Install according to the instructions at <http://www.cs.ubc.ca/~murphyk/Software/BNT/install.html>
3. Read the manual at <http://www.cs.ubc.ca/~murphyk/Software/BNT/usage.html>. In particular, read the sections entitled “Creating your first Bayes net” and “Inference”.
4. Load the file `BNT/examples/static/sprinkler1.m`, which illustrates how to do inference in the water sprinkler example. Modify this file to compute  $p(S = \text{true}|W = \text{false})$  (hint: should be 0.0621) and  $p(S = \text{true}|W = \text{false}, R = \text{false})$ . (Note: false is state 1, and true is state 2; the value of 0 is not allowed as evidence in BNT.) Turn in your probability estimates and code.

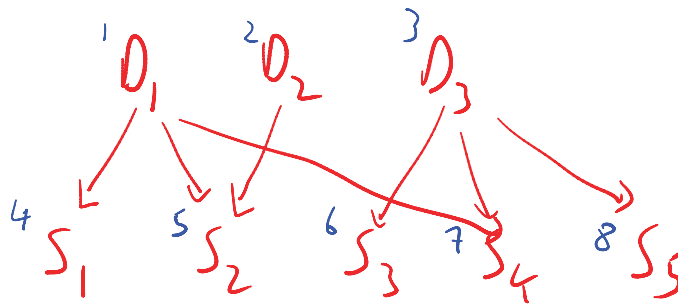


Figure 3: A small QMR Bayes net with 3 diseases and 5 symptoms. The small blue numbers in the top left next to each node are the topological numbers.

## 5 QMR in BNT

As discussed in class, the QMR (Quick Medical Reference) is a Bayes net designed by doctors to model the domain of internal medicine. Here we will consider a highly simplified version with 3 diseases nodes and 5 symptom nodes (see Figure 3). We will let the probability of inhibition be 0 on all arcs, so that the child nodes (symptoms) will be deterministic ORs of their parents. Also, we will initially set the probability of inhibiting the leak nodes to 1.0, thus essentially turning off the leaky parents.

1. Load the file `qmrStub`. This sets up the Bayes net in the way we have just described. `qmrStub` shows how you can create an evidence vector indicating that  $S_1 = +$  (node  $S_1$  is in state 2) but that all the other symptoms/ test results are unobserved/ unknown, perhaps because the test has not been done yet; we denote this by  $S_2 = ?$ ,  $S_3 = ?$ ,  $S_4 = ?$ ,  $S_5 = ?$  in the table below, and insert an empty array as evidence for the corresponding nodes. If the test result was negative, we would write  $S_1 = -$  (node  $S_1$  is in state 1). After calling the inference engine, you need to extract the marginal probabilities on each disease node, and then compute  $p(D_i = +|ev)$ , for  $i = 1 : 3$ , where  $ev$  means all the observed evidence (data). These marginal probabilities are shown in the columns labeled  $D_1$ ,  $D_2$  and  $D_3$  in the first row of the table below. We see that, since  $S_1$  is on, its only parent,  $D_1$ , must also be on; however, the other diseases are at their baseline probabilities (0.1), since we have no evidence about their status. In addition, we show  $\log p(ev)$ , the log likelihood of the data. Note that  $\log(\epsilon) = -36.0437$ , where  $\epsilon$  is the smallest representable number in Matlab, which means  $p(ev) \approx 0$  for the fourth case (see discussion below). Your goal is to finish writing the `qmrStub` function so it computes the remaining  $x$  entries in the table. Turn in your numbers and code.

S1, S2, S3, S4, S5	D1	D2	D3	LogLik
+ ? ? ? ?	1.000	0.100	0.100	-2.303
? + ? ? ?	x	x	x	-1.661
+ + ? + ?	x	x	x	-2.303
+ + ? - ?	0.000	0.000	0.000	-36.044

2. In the example above, we see that in the fourth case,  $S_1$  is on, so its parent  $D_1$  must also be on; however,  $D_1$ 's other child,  $S_4$ , is off. This is a logical contradiction, so  $p(ev) = 0$ . The deterministic model cannot explain why  $D_1$  is on yet  $S_4$  is off. We can handle such noisy data by using the noisy OR model. One option would be to assume the “wires” in the OR gate occasionally flip +’s to -’s; this could explain why  $S_4$  is off even though  $D_1$  is on. An alternative is to add leak nodes, which allow child nodes to turn on even if all their parents are off; this could explain why  $S_1$  is on even if  $D_1$  is off. We will follow the latter strategy here. Modify your code so the probability of inhibiting the leak is 0.999; thus leaky parents are enabled. Now recompute the posterior marginals using the 4 evidence cases above, and fill in the table below. Turn in your numbers and code.

S1,S2,S3,S4,S5	D1	D2	D3	LogLik
+ ? ? ? ?	0.991	0.100	0.100	-2.294
? + ? ? ?	x	x	x	-1.656
+ + ? + ?	x	x	x	-2.302
+ + ? - ?	0.000	0.991	0.000	-9.413

We see that the behavior of this model is similar to the deterministic model, except this stochastic model does not “crash” when it encounters the “logically contradictory” case 4. Instead, we infer that  $D_2$  is (probably) on, causing  $S_2$ , but that  $D_1$  is off, so  $S_1$  was caused by  $S_1$ ’s leak parent. ( $D_1$  cannot be on, since  $S_4$  is off, and there is no mechanism to turn children off if all the inhibition probabilities are zero.)