1 Problem 1

(a) The joint distribution $p(X, Y)$

$X =$	0	1	2
$Y = 0$	1/16	0	0
1	1/8	1/8	0
2	1/16	1/4	1/16
3	0	1/8	1/8
4	0	0	1/16

(1)

(b) The marginals $p(X)$ and $p(Y)$

$X =$	0	1	2
$p(X)$	1/4	1/2	1/4

$Y =$	0	1	2	3	4
$p(Y)$	1/16	1/4	3/8	1/4	1/16

(2)

(c) The conditionals $p(X|Y)$ and $p(Y|X)$

$X =$	0	1	2
$Y = 0$	1	0	0
1	1/2	1/2	0
2	1/6	2/3	1/6
3	0	1/2	1/2
4	0	0	1

$X =$	0	1	2
$Y = 0$	1/4	0	0
1	1/2	1/4	0
2	1/4	1/2	1/4
3	0	1/4	1/2
4	0	0	1/4

(3)

(d) The distribution of $Z = Y - X$, $p(Z)$

$Z =$	0	1	2
$p(Z)$	1/4	1/2	1/4

(4)

2 Problem 2

The conditional probabilities implied by this situation are as follows:

- The probability of testing positive given that you have the disease is $p(t|d) = 0.99$.
- The probability of testing positive given that you *don't* have the disease is $p(t|\tilde{d}) = 0.01$.
- The marginal probability of having the disease is only $p(d) = 10^{-4}$ and the probability of not having the disease is $p(\tilde{d}) = 1 - 10^{-4}$.
- Therefore, the marginal probability of testing positive is

$$p(t) = p(t|d)\,p(d) + p(t|\tilde{d})\,p(\tilde{d}) = 0.99 \times 10^{-4} + 0.01\,(1 - 10^{-4}) = 100.98 \times 10^{-4} \quad (5)$$

The value that the patient really cares about, though is the probability that they have the disease given that they tested positive $p(d|t)$. This — by Bayes — is

$$p(d|t) = \frac{p(d)\,p(t|d)}{p(t)} = \frac{0.99 \times 10^{-4}}{100.98 \times 10^{-4}} \approx 0.0098 \ll 1. \tag{6}$$

3 Problem 3

The simplest possible cyclic directed graph is shown in Figure 1. This graph implies the factorization

$$p(A, B) = p(A|B)\,p(B|A). \tag{7}$$

It is easy to construct an example that violates this. If A and B are binary valued, valid CPTs for these are something like

$p(B\|A)$	$A=1$	$A=0$
$B=0$	0.75	0.5
$B=1$	0.25	0.5

and

$p(A\|B)$	$A=1$	$A=0$
$B=0$	0.1	0.9
$B=1$	0.8	0.2

(8)

In this example, the factorization in Equation (7) implies the join distribution

$p(A,B)$	$A=1$	$A=0$
$B=0$	0.075	0.45
$B=1$	0.2	0.1

(9)

which is improper, i.e. $\sum_{A,B} p(A, B) = 0.825 < 1$.

Figure 1: The simplest cyclic directed graph.

4 Problem 4

We are given the three statements

1. $p(A, B|C) = p(A|C)\,p(B|C)$
2. $p(A|B, C) = p(A|C)$
3. $p(B|A, C) = p(B|C)$

To see that statement 1 implies statement 2, apply the chain rule to find

$$p(A|C)\,p(B|C) \overset{1}{=} p(A, B|C) = p(B|C)\,p(A|B, C). \tag{10}$$

Cancelling $p(B|C)$ on both sides, we find statement 2. Therefore, it is clear that statement 1 implies statement 2. Also, since we have only used the chain rule, the inverse also applies. Specifically, applying the chain rule to statement 2, we find

$$p(A|B,C) = \frac{p(A,B|C)}{p(B|C)} \stackrel{2}{=} p(A|C) \to [\text{Statement 1}]. \tag{11}$$

Similarly, statement 1 implies statement 3 as follows

$$p(A|C)\,p(B|C) \stackrel{1}{=} p(A,B|C) = p(A|C)\,p(B|A,C) \to [\text{Statement 3}] \tag{12}$$

and the inverse

$$p(B|A,C) = \frac{p(A,B|C)}{p(A|C)} \stackrel{3}{=} p(B|C) \to [\text{Statement 1}]. \tag{13}$$

Finally, since the equivalence holds between 1 and 2 and also between 1 and 3, it is clear that 2 and 3 are also equivalent.

5 Problem 5

(a) By Bayes' Theorem,

$$p(H|E_1,E_2) = \frac{p(E_1,E_2|H)\,p(H)}{p(E_1,E_2)}. \tag{14}$$

Therefore, set (ii) is clearly sufficient for this calculation. Without any conditional independence assumptions, Equation (14) cannot be simplified any further so the other two sets are not sufficient. In particular, $p(E_1,E_2|H) \neq p(E_1|H)\,p(E_2|H)$ unless $E_1 \perp E_2|H$.

(b) Since $E_1 \perp E_2|H$, $p(E_1,E_2|H) = p(E_1|H)\,p(E_2|H)$ and Equation (14) becomes

$$p(H|E_1,E_2) = \frac{p(E_1|H)\,p(E_2|H)\,p(H)}{p(E_1,E_2)}. \tag{15}$$

Therefore, sets (i) and (ii) are now sufficient. Set (iii) is not sufficient because it would require that $E_1 \perp E_2$ but E_1 and E_2 are only *conditionally* independent.

6 Problem 6

Since the variables x, y and z are interchangeable in this distribution, it is sufficient to show that none of the graphs in Figure 2 imply the same set of independences as the distribution. The joint distribution of x, y and z is

$Z=$	0	0	1	1
$X=$	0	1	0	1
$Y=0$	1/12	1/6	1/6	1/12
1	1/6	1/12	1/12	1/6

(16)

The first graph in Figure 2 implies the condition $X \perp Z|Y$ or,

$$p(X,Y,Z) = p(X)\,p(Y|X)\,p(Z|Y). \tag{17}$$

For our example, $p(X = 1) = 1/2$, $p(Y = 1|X = 1) = 1/2$ and $p(Z = 1|Y = 1) = 1/2$ while $p(X = 1, Y = 1, Z = 1) = 1/6$ therefore,

$$\frac{1}{6} \neq \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}. \tag{18}$$

For the next graphs, it is important to note that X and Y are independent, i.e.

$$p(X, Y) \stackrel{?}{=} p(X)\, p(Y) \tag{19}$$

where $p(X = 1) = 1/2$, $p(Y = 1) = 1/2$ and $p(X = 1, Y = 1) = 1/4$ satisfy this equality.

The second graph in Figure 2 also implies that

$$p(X, Y, Z) = p(X)\, p(Y)\, p(Z|X, Y) \tag{20}$$

but $p(Z = 1|X = 1, Y = 1) = 2/3$ and $p(Z = 1, X = 1, Y = 1) = 1/6$ so this condition is not satisfied by the distribution.

For the third graph, the independence condition implied is

$$p(X, Y, Z) = p(X)\, p(Y)\, p(Z|X). \tag{21}$$

We can try, for example, $p(Z = 1|X = 1) = 1/4$ and $p(X = 1, Y = 1, Z = 1) = 1/6$. Therefore, this condition is not satisfied since

$$\frac{1}{6} \neq \frac{1}{2}\frac{1}{2}\frac{1}{4}. \tag{22}$$

Finally, the fourth graph in Figure 2 is the easiest. It implies that this distribution factors as

$$p(X, Y, Z) = p(X)\, p(Y)\, p(Z) \tag{23}$$

but $p(X = 1)$, $p(Y = 1)$ and $p(Z = 1)$ are all equal to $1/2$ and $p(X = 1, Y = 1, Z = 1) = 1/6 \neq (1/2)^3$.

Therefore, since none of these graphs represent the same independences as the distribution, no directed acyclic graph on 3 variables has $I_{\mathrm{d-sep}}(G) = I(p)$.

7 Problem 7

(a) and (b) I will combine both parts for this question because I find it much easier to generate the tables given a physical intuition.

Case 1: $a > c$. The example for this one is that $X = 1$ means that the fire alarm in your house is broken and it has a relatively small prior probability $p(X = 1) = 0.2$. Then, $Y = 1$ indicates that there is currently a fire in your house and $Z = 1$ says that you can hear the fire alarm going off. In this case, $p(X = 1|Z = 1, Y = 1) \sim 0$ because if the fire alarm is going off *and* there is a fire then the probability that the alarm is broken is very small.

Case 2: $a < c < b$. In this case, an example would be that $X = 1$ means that you are very intelligent, $Y = 1$ means that you work very hard and $Z = 1$ means that you pass this class. There is only a small prior probability that you are very intelligent — $p(X = 1) = 0.1$ — but if you pass then there is a higher probability that you are very intelligent — $p(X = 1|Z = 1) = 0.6$. In between, if you pass but you're also a very hard worker, it's slightly less likely that you're doing well based on your intelligence alone: $p(X = 1|Z = 1, Y = 1) = 0.5$.

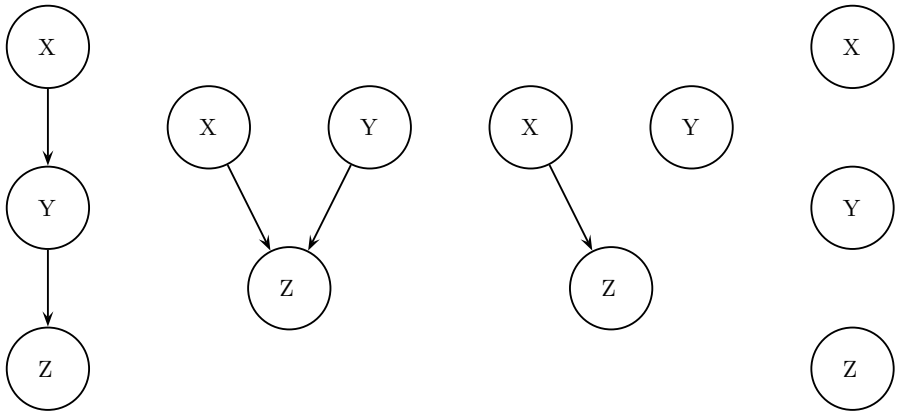

Figure 2: The four different graphs on 3 random variables.

Case 3: $b < a < c$. A simple example satisfying this criterion is that $Z = (X == Y)$. If the prior probability of $X = 1$ is $p(X = 1) = 0.5$ and the prior probability of $Y = 1$ is $p(Y = 1) = 0.01$ then, $b = p(X = 1|Z = 1) = p(X = 1|Z = 1, Y = 1)p(Y = 1) + p(X = 1|Z = 1, Y = 0)p(Y = 0) = 0.01$ and $c = p(X = 1|Z = 1, Y = 1) = 1$. The joint probability table is

$Z =$	0	0	1	1
$X =$	0	1	0	1
$Y = 0$	0	0.495	0.495	0
1	0.005	0	0	0.005

(24)

8 Problem 8

(a) The set A is $\{X_2, X_3, X_4, X_5, X_8\}$. Clearly, all the nodes that are directly connected to X_1 (i.e. $\{X_2, X_3, X_4, X_8\}$) must be included in A because a direct connection always constitutes an active path. The inclusion of X_5 is not immediately obvious but if we just look at the part of the graph containing X_5, we find the V-structure $X_1 \to X_3 \leftarrow X_5$. If we condition on X_3 (which we will do because it is one of the directly connected nodes), it couples its parents X_1 and X_5. Therefore, to satisfy the condition $X_1 \perp \chi - A - \{X_1\}|A$, we must also include X_5 in A. After the inclusion of X_5, there are no other active paths between X_1 and other nodes outside of A — this can be easily seen by trying them all.

(b) This algorithm generalizes as follows:

- Add all the children and parents of X_i to A. As before, these are all required connections because any direct connection in the graph is an active path (i.e. X_i is never independent of these variables).
- This is not yet sufficient, however. If any of the children of X_i (say Y) have other parents then conditioning on Y couples X_i to all of Y's parents since it is a V-structure. Therefore, all of the parents of Y must be added to A.

- This *is* now sufficient because the neighbors of X_i's parents and the children of X_i's children are all conditionally independent of X_i, conditioned on A. This is easily seen by looking at the possible active paths in these directions and remembering that for a cascade structure, the end nodes are conditionally independent of each other conditioned on the middle node. This is also true for a common parent structure.

9 Problem 9

(a) Without any conditioning, the only (non-trivial) active paths in this graph are: $1 \to 6 \to 4$, $8 \to 9 \to 5$, $4 \to 7 \to 9 \to 5$, $4 \to 2 \to 10 \to 3 \to 9 \to 5$, $4 \to 6 \to 2 \to 10 \to 3 \to 9 \to 5$ and $6 \to 2 \to 4$. Therefore, this implies the following set of independences: $X_1 \perp X_2, X_3, X_5, X_7, X_8, X_9, X_{10}$, $X_2 \perp X_1, X_7, X_8$ $X_3 \perp X_1, X_7, X_8$, $X_4 \perp X_8$, $X_5 \perp X_1$, $X_6 \perp X_7, X_8$, $X_7 \perp X_1, X_2, X_3, X_6, X_8, X_{10}$, $X_8 \perp X_1, X_2, X_3, X_4, X_6, X_7, X_{10}$, $X_9 \perp X_1$, and $X_{10} \perp X_1, X_7, X_8$. This can be more clearly summarized in the following table:

	1	2	3	4	5	6	7	8	9	10
1		⊥	⊥		⊥		⊥	⊥	⊥	⊥
2	⊥						⊥	⊥		
3	⊥						⊥	⊥		
4								⊥		
5	⊥									
6							⊥	⊥		
7	⊥	⊥	⊥			⊥		⊥		⊥
8	⊥	⊥	⊥	⊥		⊥	⊥			⊥
9	⊥									
10	⊥						⊥	⊥		

(25)

(b) The conditioning on $\{X_2, X_9\}$ does not actually affect the set of independences implied by the graph structure for X_1. Therefore, the largest set A is $\{X_3, X_5, X_7, X_8, X_{10}\}$.

(c) Using the d-separation algorithm from Koller & Friedman, we find that the only active paths (after conditioning on $\{X_2, X_9\}$) containing the node X_8 are $8 \to 9 \to 3 \to 10 \to 2$ and $8 \to 9 \to 7 \to 4$. Therefore, the set B is $\{X_1, X_5, X_6\}$.

10 Problem 10 — Exercise 3.11

(a) The left panel of Figure 3 shows graph from the problem and the right panel shows the marginalized graph.

(b) The general procedure is as follows: to marginalize over a node A, remove the node A from the graph and draw edges directed from each of the parents of A to each of the children of A.

(c) Applying this procedure to the naive Bayes model and marginalizing out the class variable will cause *all* of the feature nodes to be independent, i.e. $X_i \perp \chi - \{X_i\}$.

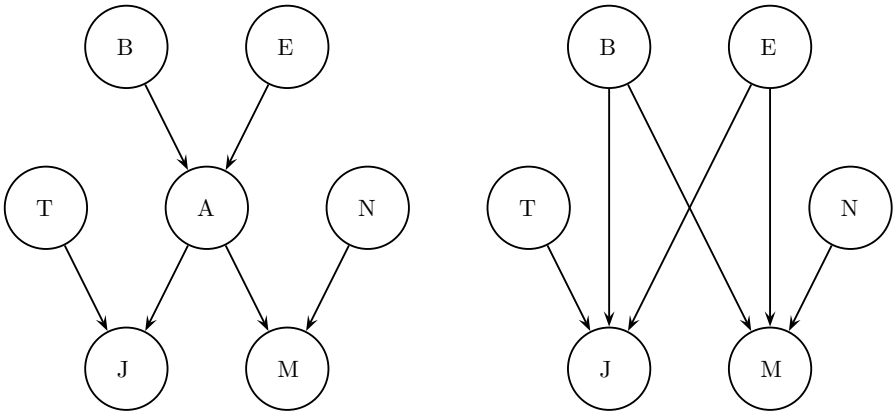

Figure 3: *Left*: the original graph. *Right*: the graph marginalized over A.

11 Problem 11 — Exercise 3.15

The set of independences implied by graph (a) are $D \perp A, C|B$ and $A \perp C$. There are no other I-equivalent graphs. The independences implied by the Bayesian network (b) are $A \perp C, D|B$ and $C \perp D|B$. The four Bayesian networks (including (b) from the exercise) in Figure 4 all imply this same independence structure. Therefore, they are all I-equivalent.

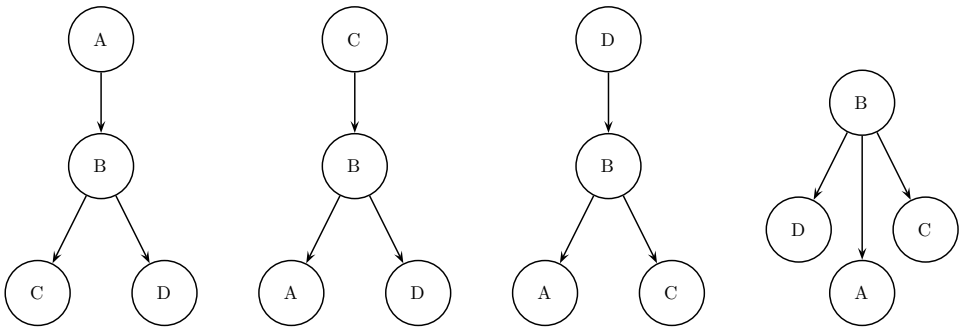

Figure 4: Four I-equivalent Bayesian networks.

12 Problem 12 — Exercise 3.2

(a) The assumption given in Equation (3.6) is that each feature X_i is independent of the other features $\chi - \{X_i\}$ conditioned on the class C. This can be written as $X_i \perp \chi - \{X_i\}|C$. Any joint distribution $p(C, X_1, \ldots, X_n)$ can be factored (using the chain rule) into $p(C)\, p(X_1, \ldots, X_n|C)$. Then, for a particular feature X_i, the conditional independence assumption above implies that

$$p(X_i, \chi - \{X_i\}|C) = p(X_i|C)\, p(\chi - \{X_i\}|C). \tag{26}$$

Applying the conditional independence assumption again, we find

$$p(X_i|C)\,p(X_j, \chi - \{X_i, X_j\}|C) = p(X_i|C)\,p(X_j|C)\,p(\chi - \{X_i, X_j\}|C). \tag{27}$$

We can iterate this procedure for all values of $i = 1, \ldots, n$ to find that

$$p(X_1, \ldots, X_n|C) = \prod_{i=1}^{n} p(X_i|C). \tag{28}$$

Therefore, the conditional independence assumption from Equation (3.6) implies that the join distribution can be factored

$$p(C, X_1, \ldots, X_n) = p(C) \prod_{i=1}^{n} p(X_i|C) \tag{29}$$

which is exactly the result from Equation (3.7).

(b) Using the chain rule, we can rewrite the joint probability above as

$$p(C, \boldsymbol{X}) = p(\boldsymbol{X})\,p(C|\boldsymbol{X}). \tag{30}$$

Therefore, the ratio of joint probabilities can be written (for the observed feature vector $\boldsymbol{x}$)

$$\frac{p(c_1, \boldsymbol{x})}{p(c_2, \boldsymbol{x})} = \frac{p(c_1|\boldsymbol{x})}{p(c_2|\boldsymbol{x})}. \tag{31}$$

Then, using Equation (29), this ratio can also be written

$$\frac{p(c_1, \boldsymbol{x})}{p(c_2, \boldsymbol{x})} = \frac{p(c_1)}{p(c_2)} \prod_{i=1}^{n} \frac{p(x_i|c_1)}{p(x_i|c_2)}. \tag{32}$$

Equating these two expressions, we find the expected Equation (3.8):

$$\frac{p(c_1|\boldsymbol{x})}{p(c_2|\boldsymbol{x})} = \frac{p(c_1)}{p(c_2)} \prod_{i=1}^{n} \frac{p(x_i|c_1)}{p(x_i|c_2)}. \tag{33}$$

(c) Taking the logarithm of Equation (33), we find

$$\log\left[\frac{p(c_1|\boldsymbol{x})}{p(c_2|\boldsymbol{x})}\right] = \log\left[\frac{p(c_1)}{p(c_2)}\right] + \sum_{i=1}^{n} \left[\log p(x_i|c_1) - \log p(x_i|c_2)\right] \tag{34}$$

13 Problem 13

(a) A recursive equation for the marginalized probability $p(X_i = 1)$ is given by

$$p(X_i = 1) = p(X_{i-1} = 1)\;\left[p(X_i = 1|X_{i-1} = 1) - p(X_i = 1|X_{i-1} = 0)\right] + p(X_i = 1|X_{i-1} = 0). \tag{35}$$

Starting with

$$p(X_2 = 1) = p(X_1 = 1)\,p(X_2 = 1|X_1 = 1) + p(X_1 = 0)\,p(X_2 = 1|X_1 = 0), \tag{36}$$

where everything is known, we can iterate using Equation (35) to find $p(X_i = 1)$ for each $i = 1, \ldots, n$ in linear time. The pseudocode for this algorithm is shown here:

```
# given px[i,1,1] = p(X[i] = 1 | X[i-1] = 1)
# and px1[1] = p(X[1] = 1)

pxi1 = []

pxi1.push( px1[1] )

for i in 2...n:
    pxi1.push( pxi1[-1] * (px[i,1,1] - px[i,1,0]) + px[i,1,0] )

# now, pxi1 contains a list of all the p(X[i] = 1)
```

(b) Analogous to Equation (35), the recursive form for the conditional probability is

$$p(X_i = 1|X_1 = 1) = p(X_{i-1} = 1|X_1 = 1)\ [p(X_i = 1|X_{i-1} = 1) - p(X_i = 1|X_{i-1} = 0)] + p(X_i = 1|X_{i-1} = 0).$$

and the algorithm is

```
pxi1 = []

pxi1.push( px[2,1,1] )

for i in 3...n:
    pxi1.push( pxi1[-1] * (px[i,1,1] - px[i,1,0]) + px[i,1,0] )

# now, pxi1 contains a list of all the p(X[i] = 1 | X[1] = 1)
```

(c) By Bayes,

$$p(X_1 = 1|X_i = 1) = \frac{p(X_1 = 1)\, p(X_i = 1|X_1 = 1)}{p(X_i = 1)} \tag{37}$$

Therefore, the algorithm becomes

```
px = []

pxi1  = px1[1] * (px[2,1,1] - px[2,1,0]) + px[2,1,0]
pcond = px[2,1,1]

px.push( px1[1] * pcond / pxi1 )

for i in 3...n:
    pxi1  = pxi1  * (px[i,1,1] - px[i,1,0]) + px[i,1,0]
    pcond = pcond * (px[i,1,1] - px[i,1,0]) + px[i,1,0]
    px.push( px1[1] * pcond / pxi1 )

# now, pxi1 contains a list of all the p(X[1] = 1 | X[i] = 1)
```

This still scales linearly with n.