1. (Modified version of Exercise 3.1 in the textbook) Suppose you have three Boolean random variables $X_1, X_2, X_3$. Write out a discrete joint probability distribution $P(X_1, X_2, X_3)$ where for each $i \neq j$, we have that $(X_i \perp X_j) \in I(P)$, but we also have that $((X_1, X_2) \perp X_3) \notin I(P)$. This should be a table with $2^3$ rows (one for each combination of values for the three random variables). Explain your solution. [4 points]

You will need to create a probability distribution where $P(X_i, X_j) = P(X_i)P(X_j)$ and $P(X_1, X_2, X_3) \neq P(X_1, X_2)P(X_3)$. There are many possible solutions. One example is the XOR function in which the joint distribution has probability $1/4$ if there are an odd number of “true”s and 0 otherwise. Note that in the XOR example, knowing all three random variables gives you the value of the joint distribution but knowing only two of them doesn’t.

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$P(X_1, X_2, X_3)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>false</td>
<td>false</td>
<td>false</td>
<td>0.25</td>
</tr>
<tr>
<td>false</td>
<td>false</td>
<td>true</td>
<td>0.25</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>false</td>
<td>0.25</td>
</tr>
<tr>
<td>false</td>
<td>true</td>
<td>true</td>
<td>0</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>false</td>
<td>0</td>
</tr>
<tr>
<td>true</td>
<td>false</td>
<td>true</td>
<td>0.25</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>false</td>
<td>0</td>
</tr>
<tr>
<td>true</td>
<td>true</td>
<td>true</td>
<td>0.25</td>
</tr>
</tbody>
</table>

To see that $(X_i \perp X_j) \in I(P)$

$P(X_1 = \text{true}, X_2 = \text{true}) = P(X_1 = \text{true})P(X_2 = \text{true}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$P(X_2 = \text{true}, X_3 = \text{true}) = P(X_2 = \text{true})P(X_3 = \text{true}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$P(X_1 = \text{true}, X_3 = \text{true}) = P(X_1 = \text{true})P(X_3 = \text{true}) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

To see that $((X_1, X_2) \perp X_3) \notin I(P)$.

$P(X_1 = \text{true}, X_2 = \text{true}, X_3 = \text{true}) \neq P(X_1 = \text{true}, X_2 = \text{true})P(X_3 = \text{true})$

$= \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$
2. (Modified version of 3.6) Consider a set of random variables \( X_1, X_2, \ldots, X_n \) where each \( X_i \) has \( |Val(X_i)| = l \) i.e. \( X_i \) takes on \( l \) different values.

a) Assume that we have a Bayesian network over \( X_1, X_2, \ldots, X_n \) such that each node has at most \( k \) parents. What is a simple upper bound on the number of independent parameters in the Bayesian network? How many independent parameters are in the full joint distribution over \( X_1, X_2, \ldots, X_n \)? [2 points]

If each node has at most \( k \) parents, it has \((l - 1)l^k\) independent parameters.
With \( n \) nodes, the upper bound is \( n(l - 1)l^k \) independent parameters

There are \( l^n - 1 \) independent parameters in the full joint distribution.

b) Now, assume that each variable \( X_i \) has the parents \( X_1, X_2, \ldots, X_{i-1} \). How many independent parameters are there in the Bayesian network? What can you conclude about the expressive power of this type of network? [4 points]

To compute the number of independent parameters, consider it node by node:
\( X_1 \) has \((l - 1)\)
\( X_2 \) has \((l - 1)l\)
\( X_3 \) has \((l - 1)l^2\)

More generally, we have: \( \sum_{t=1}^{n} (l - 1)l^{t-1} = (l - 1) \sum_{t=1}^{n} l^{t-1} \)

This network is basically the chain rule and it can express any joint probability distribution e.g.
\[ P(X_1, X_2, X_3) = P(X_1|X_2, X_3)P(X_2|X_3)P(X_3) \]

To compute the number of independent parameters, consider it node by node:
\( X_1 \) has \((l - 1)\)
\( X_2 \) has \((l - 1)l\)
\( X_3 \) has \((l - 1)l^2\)

More generally, we have: \( \sum_{t=1}^{n} (l - 1)l^{t-1} = (l - 1) \sum_{t=1}^{n} l^{t-1} \)

This network is basically the chain rule and it can express any joint probability distribution e.g.
\[ P(X_1, X_2, X_3) = P(X_1|X_2, X_3)P(X_2|X_3)P(X_3) \]

To compute the number of independent parameters, consider it node by node:
\( X_1 \) has \((l - 1)\)
\( X_2 \) has \((l - 1)l\)
\( X_3 \) has \((l - 1)l^2\)

More generally, we have: \( \sum_{t=1}^{n} (l - 1)l^{t-1} = (l - 1) \sum_{t=1}^{n} l^{t-1} \)

This network is basically the chain rule and it can express any joint probability distribution e.g.
\[ P(X_1, X_2, X_3) = P(X_1|X_2, X_3)P(X_2|X_3)P(X_3) \]

c) Now, consider a naïve Bayes model where \( X_1, X_2, \ldots, X_n \) are evidence variables (i.e. observed features), and we have an additional class variable \( C \) which has \( k \) possible values \( c_1, \ldots, c_k \). How many independent parameters are required to specify the naïve Bayes model? How many independent parameters are required for an explicit representation of this joint distribution? [4 points]

The class variable node \( C \) has \( k - 1 \) independent parameters.
Each evidence variable \( P(X_i|C) \) has \((l - 1)k\) independent parameters. There are \( n \) evidence variables and thus a grand total of \( nk(l - 1) + (k - 1) \) independent parameters.

The joint distribution has \( k l^n - 1 \) independent parameters

3. Answer true or false to the following conditional independence statements using the graph below. For partial credit, show the paths that are blocked or not blocked. [9 points]
a) \( A \perp D \mid \{C, F\} \)

True:
- \( A \rightarrow B \rightarrow C \rightarrow D \) blocked by \( C \)
- \( A \rightarrow E \rightarrow C \rightarrow D \) blocked by \( C \)
- \( A \rightarrow F \rightarrow D \) blocked by \( F \)
- \( A \rightarrow G \leftarrow D \) blocked by not observing \( G \) (v-structure)

Note that there are other paths but they will include the paths above as subpaths

b) \( B \perp F \mid A \)

True
- \( B \rightarrow C \rightarrow D \leftarrow F \) blocked by not observing \( D \) (v-structure)
- \( B \leftarrow A \rightarrow F \) blocked by \( A \)
- \( B \rightarrow C \leftarrow E \leftarrow A \rightarrow F \) blocked by \( C \) not being observed and by \( A \)
- \( B \leftarrow A \rightarrow G \leftarrow D \leftarrow F \) blocked by \( A \) being observed and \( G \) not observed
- \( B \rightarrow C \leftarrow E \leftarrow A \rightarrow F \) blocked by \( A \) being observed and \( C \) not observed

In general, all paths through the v-structure involving \( C \) or the v-structure involving \( D \) or the v-structure involving \( G \) will be blocked

c) \( A \perp C \mid \{B, E, F, G\} \)

False
- \( A \rightarrow G \leftarrow D \leftarrow C \) not blocked (\( G \) is observed and descendant of v-structure at \( D \))

4. (Exercise 3.7 in the textbook) Show how you could efficiently compute the distribution over a variable \( X_i \) given some assignment to all the other variables in the network: \( P(X_i \mid x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \). Your procedure should not require the construction of the entire joint distribution \( P(x_1, ..., x_n) \). Specify the computational complexity of your procedure using Big-O notation. [8 points]
\[
P(X_l|x_1, \ldots, x_{i-1}, x_{i+1}, x_n) \]
\[
= \frac{P(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, x_n)}{P(x_1, \ldots, x_{i-1}, x_{i+1}, x_n)} \sum_{x_i} P(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, x_n)
\]
\[
= \frac{\prod_{j=1}^n P(x_j | \text{Parents}(x_j))}{\sum_{x_i} \prod_{j=1}^n P(x_j | \text{Parents}(x_j))} \cdot \frac{\prod_{x_j \notin \{x_i, \text{Children}(x_i)\}} P(x_j | \text{Parents}(x_j)) P(x_i | \text{Parents}(x_i)) \prod_{x_k \in \text{Children}(x_i)} P(x_k | x_i, \text{other parents})}{\prod_{x_j \notin \{x_i, \text{Children}(x_i)\}} P(x_j | \text{Parents}(x_j)) \sum_{x_i} P(x_i | \text{Parents}(x_i)) \prod_{x_k \in \text{Children}(x_i)} P(x_k | x_i, \text{other parents})}
\]

Assuming each node in the Bayesian network takes k values and that there are C children of node \( x_i \) then the numerator takes \( O(C) \) to compute and the denominator takes \( O(kC) \) to compute.

5. (Exercise 3.10) Prove that the global independencies, derived from d-separation, imply the local independencies (i.e. a node is conditionally independent of its non-descendants given its parents). In other words, prove that a node is d-separated from its non-descendants given its parents. [9 points]

Let X be the node of interest, Y be its parent, D be a descendant and ND a non-descendant. To complete this proof, you need to go through the cases of d-separation that involve non-descendants.

1) ND->Y->X->D
   Here, the non-descendant is an ancestor of X. From the causal trail case (or evidential trail case) of d-separation, observing Y renders ND and X independent. Thus \( X \perp ND | Y \)

2) ND <- Y -> X -> D
   From the common cause case of d-separation, observing Y renders ND and X independent. Thus \( X \perp ND | Y \)

3) Y->X->D<-ND
   From the common effect ( v-structure ) case of d-separation, not observing D renders X and ND independent. Thus \( X \perp ND | Y \)

QED