Recap of Probability Theory, Bayesian Networks, Conditional Independence in Bayesian Networks
Module 17.0: Recap of Probability Theory
We will start with a quick recap of some basic concepts from probability
Axioms of Probability

For any event $A$,
$$P(A) \geq 0$$

If $A_1, A_2, A_3, \ldots, A_n$ are disjoint events (i.e., $A_i \cap A_j = \emptyset \forall i \neq j$),
$$P(\bigcup A_i) = \sum_i P(A_i)$$

If $\Omega$ is the universal set containing all events then
$$P(\Omega) = 1$$
Axioms of Probability

- For any event $A$,

$$P(A) \geq 0$$
Axioms of Probability

- For any event $A$, 
  
  $$P(A) \geq 0$$

- If $A_1, A_2, A_3, \ldots, A_n$ are disjoint events (i.e., $A_i \cap A_j = \emptyset \ \forall i \neq j$) then 
  
  $$P(\cup A_i) = \sum_i P(A_i)$$
Axioms of Probability

- For any event $A$,
  \[ P(A) \geq 0 \]

- If $A_1, A_2, A_3, \ldots, A_n$ are disjoint events (i.e., $A_i \cap A_j = \emptyset \ \forall i \neq j$) then
  \[ P(\cup A_i) = \sum_i P(A_i) \]

- If $\Omega$ is the universal set containing all events then
  \[ P(\Omega) = 1 \]
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: $A, B, C$

$$\Omega$$

A

B

C
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: $A, B, C$
- One way of interpreting this is that there are 3 possible events here.

Suppose a student can get one of 3 possible grades in a course: $A, B, C$. One way of interpreting this is that there are 3 possible events here.
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: \(A, B, C\)
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a random variable \(G\) which each student to one of the 3 possible values
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: $A, B, C$
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a random variable $G$ which each student to one of the 3 possible values
- And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$
Random Variable (intuition)

- Suppose a student can get one of 3 possible grades in a course: $A, B, C$
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a random variable $G$ which each student to one of the 3 possible values
- And we are interested in $P(G = g)$ where $g \in \{A, B, C\}$
- Of course, both interpretations are conceptually equivalent
Random Variable (intuition)

- But the second one (using random variables) is more compact
Random Variable (intuition)

- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - grade, height, age, etc.
Random Variable (intuition)

- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - grade, height, age, etc.
- We could have one random variable corresponding to each attribute
Random Variable (intuition)

- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - grade, height, age, etc.
- We could have one random variable corresponding to each attribute
- And then ask for outcomes (or students) where \( \text{Grade} = g, \text{Height} = h, \text{Age} = a \) and so on
Random Variable (formal)

A random variable is a function which maps each outcome in \( \Omega \) to a value. In the previous example, \( G \) (or \( f \) grade) maps each student in \( \Omega \) to a value: A, B, or C.

The event \( \text{Grade} = A \) is a shorthand for the event \( \{ \omega \in \Omega : f \text{Grade} = A \} \).
Random Variable (formal)

- A random variable is a \textit{function} which maps each outcome in $\Omega$ to a value.

A random variable is a function which maps each outcome in $\Omega$ to a value.

In the previous example, $G$ (or $f$ grade) maps each student in $\Omega$ to a value: A, B or C. The event $\text{Grade} = A$ is a shorthand for the event $\{\omega \in \Omega : f \text{Grade} = A\}$.
Random Variable (formal)

- A random variable is a \textit{function} which maps each outcome in $\Omega$ to a value.
- In the previous example, $G$ (or $f_{\text{grade}}$) maps each student in $\Omega$ to a value: $A$, $B$ or $C$. 

\[ \Omega \]

\[ \begin{array}{c}
\text{Grades} \\
A \\
B \\
C \\
\text{Height} \\
\text{Short} \\
\text{Tall} \\
\text{Age} \\
\text{Adult} \\
\text{Young} \\
\end{array} \]
Random Variable (formal)

- A random variable is a function which maps each outcome in $\Omega$ to a value.
- In the previous example, $G$ (or $f_{grade}$) maps each student in $\Omega$ to a value: $A$, $B$ or $C$.
- The event $Grade = A$ is a shorthand for the event $\{\omega \in \Omega : f_{Grade} = A\}$.
Random Variable (continuous v/s discrete)

- A random variable can either take continuous values (for example, weight, height)
Random Variable (continuous v/s discrete)

- A random variable can either take continuous values (for example, weight, height)
- Or discrete values (for example, grade, nationality)
Random Variable (continuous v/s discrete)

- A random variable can either take continuous values (for example, weight, height)
- Or discrete values (for example, grade, nationality)
- For this discussion we will mainly focus on discrete random variables
Marginal Distribution

What do we mean by marginal distribution over a random variable?

Specifying the marginal distribution over $G$ means specifying $P(G = g)$ for all $g \in A, B, C$.

We denote this marginal distribution compactly by $P(G)$.
Marginal Distribution

- What do we mean by *marginal distribution* over a random variable?
- Consider our random variable $G$ for grades

Specifying the marginal distribution over $G$ means specifying $\Pr(G = g) \forall g \in A, B, C$.

We denote this marginal distribution compactly by $\Pr(G)$. 

Mitesh M. Khapra  
CS7015 (Deep Learning) : Lecture 17
Marginal Distribution

- What do we mean by marginal distribution over a random variable?
- Consider our random variable $G$ for grades.
- Specifying the marginal distribution over $G$ means specifying

$$P(G = g) \quad \forall g \in A, B, C$$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$P(G = g)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Marginal Distribution

- What do we mean by *marginal distribution* over a random variable?
- Consider our random variable $G$ for grades
- Specifying the marginal distribution over $G$ means specifying

$$P(G = g) \quad \forall g \in A, B, C$$

- We denote this marginal distribution compactly by $P(G)$

<table>
<thead>
<tr>
<th>$G$</th>
<th>$P(G = g)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1</td>
</tr>
<tr>
<td>B</td>
<td>0.2</td>
</tr>
<tr>
<td>C</td>
<td>0.7</td>
</tr>
</tbody>
</table>
Joint Distribution

- Consider two random variable $G$ (grade) and $I$ (intelligence $\in \{\text{High, Low}\}$)
Joint Distribution

- Consider two random variable $G$ (grade) and $I$ (intelligence $\in \{\text{High, Low}\}$)
- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

$$P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}$$
Joint Distribution

- Consider two random variable $G$ (grade) and $I$ (intelligence $\in \{\text{High, Low}\}$)

- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

\[
P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}
\]

<table>
<thead>
<tr>
<th>$G$</th>
<th>$I$</th>
<th>$P(G = g, I = i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>Low</td>
<td>0.1</td>
</tr>
<tr>
<td>B</td>
<td>High</td>
<td>0.15</td>
</tr>
<tr>
<td>B</td>
<td>Low</td>
<td>0.15</td>
</tr>
<tr>
<td>C</td>
<td>High</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>Low</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Joint Distribution

- Consider two random variables \( G \) (grade) and \( I \) (intelligence \( \in \{\text{High, Low}\} \))

- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

\[
P(G = g, I = i) \quad \forall (g, i) \in \{A, B, C\} \times \{H, L\}
\]

- We denote this joint distribution compactly by \( P(G, I) \)

<table>
<thead>
<tr>
<th>( G )</th>
<th>( I )</th>
<th>( P(G = g, I = i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>A</td>
<td>Low</td>
<td>0.1</td>
</tr>
<tr>
<td>B</td>
<td>High</td>
<td>0.15</td>
</tr>
<tr>
<td>B</td>
<td>Low</td>
<td>0.15</td>
</tr>
<tr>
<td>C</td>
<td>High</td>
<td>0.1</td>
</tr>
<tr>
<td>C</td>
<td>Low</td>
<td>0.2</td>
</tr>
</tbody>
</table>
**Conditional Distribution**

- Consider two random variables $G$ (grade) and $I$ (intelligence)

| $G$ | $P(G|I = H)$ |
|-----|--------------|
| A   | 0.6          |
| B   | 0.3          |
| C   | 0.1          |

| $G$ | $P(G|I = L)$ |
|-----|--------------|
| A   | 0.3          |
| B   | 0.4          |
| C   | 0.3          |
Conditional Distribution

- Consider two random variable $G$ (grade) and $I$ (intelligence)
- Suppose we are given the value of $I$ (say, $I = H$) then the conditional distribution $P(G|I)$ is defined as

$$P(G = g | I = H) = \frac{P(G = g, I = H)}{P(I = H)} \quad \forall g \in \{A, B, C\}$$

| $G$ | $P(G|I = H)$ |
|-----|-------------|
| A   | 0.6         |
| B   | 0.3         |
| C   | 0.1         |

| $G$ | $P(G|I = L)$ |
|-----|-------------|
| A   | 0.3         |
| B   | 0.4         |
| C   | 0.3         |
Conditional Distribution

- Consider two random variable $G$ (grade) and $I$ (intelligence)

- Suppose we are given the value of $I$ (say, $I = H$) then the conditional distribution $P(G|I)$ is defined as

$$P(G = g|I = H) = \frac{P(G = g, I = H)}{P(I = H)} \quad \forall g \in \{A, B, C\}$$

- More compactly defined as

$$P(G|I) = \frac{P(G, I)}{P(I)}$$

or

$$P(G, I) = P(G|I) \ast P(I)$$

| $G$ | $P(G|I = H)$ |
|-----|--------------|
| A   | 0.6          |
| B   | 0.3          |
| C   | 0.1          |

| $G$ | $P(G|I = L)$ |
|-----|--------------|
| A   | 0.3          |
| B   | 0.4          |
| C   | 0.3          |
Joint Distribution \((n\ \text{random variables})\)

- The joint distribution of \(n\) random variables assigns probabilities to all events involving the \(n\) random variables,

\[
\begin{array}{|c|c|c|c|}
\hline
X_1 & \ldots & X_n & P(X_1, X_2, \ldots, X_n) \\
\hline
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots \\
\hline
\end{array}
\]

\[
\sum = 1
\]
Joint Distribution \((n\ \text{random variables})\)

- The joint distribution of \(n\) random variables assigns probabilities to all events involving the \(n\) random variables,
- In other words it assigns

\[
P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]
for all possible values that variable \(X_i\) can take

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(\ldots)</th>
<th>(X_n)</th>
<th>(P(X_1, X_2, \ldots, X_n))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
<td>(\ldots)</td>
</tr>
</tbody>
</table>

\[\sum = 1\]
Joint Distribution (*n* random variables)

- The joint distribution of *n* random variables assigns probabilities to all events involving the *n* random variables,

- In other words it assigns

\[
P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)
\]

*for all possible values that variable* \(X_i\) *can take*

- If each random variable \(X_i\) can take two values then the joint distribution will assign probabilities to the \(2^n\) possible events
Joint Distribution \((n\) random variables\)

- The joint distribution over two random variables \(X_1\) and \(X_2\) can be written as,

\[
P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)
\]

\[
\begin{array}{|c|c|c|}
\hline
X_1 & \ldots & X_n \\
\hline
\ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots \\
\hline
\end{array}
\]

\[
P(X_1, X_2, \ldots, X_n) = \prod_{i=2}^{n} P(X_i|X_{i-1})\] (chain rule)
Joint Distribution \((n\text{ random variables})\)

- The joint distribution over two random variables \(X_1\) and \(X_2\) can be written as,
  \[
P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)
  \]

- Similarly for \(n\) random variables
  \[
P(X_1, X_2, \ldots, X_n)
  \]
Joint Distribution \((n\ \text{random variables})\)

- The joint distribution over two random variables \(X_1\) and \(X_2\) can be written as,

\[
P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)
\]

- Similarly for \(n\ \text{random variables}\)

\[
P(X_1, X_2, \ldots, X_n)
\]

\[
= P(X_2, \ldots, X_n|X_1)P(X_1)
\]
Joint Distribution \((n \text{ random variables})\)

- The joint distribution over two random variables \(X_1\) and \(X_2\) can be written as,

\[
P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)
\]

- Similarly for \(n\) random variables

\[
P(X_1, X_2, ..., X_n) \\
= P(X_2, ..., X_n|X_1)P(X_1) \\
= P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1)
\]
Joint Distribution \((n\ \text{random variables})\)

- The joint distribution over two random variables \(X_1\) and \(X_2\) can be written as,

\[
P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)
\]

- Similarly for \(n\ \text{random variables}\)

\[
P(X_1, X_2, ..., X_n)
= P(X_2, ..., X_n|X_1)P(X_1)
= P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1)
= P(X_4, ..., X_n|X_1, X_2, X_3)P(X_3|X_2, X_1)P(X_1)
= \cdots
= P(X_n|X_1, X_2, ..., X_{n-1})P(X_{n-1}|X_{n-2}, ..., X_1)P(X_1)
\]

\[
P(X_1, X_2, ..., X_n) = \prod_{i=2}^{n} P(X_i|X_{i-1})
\]

\(\text{(chain rule)}\)
Joint Distribution \((n\text{ random variables})\)

- The joint distribution over two random variables \(X_1\) and \(X_2\) can be written as,
  \[
P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)
  \]

- Similarly for \(n\) random variables
  \[
P(X_1, X_2, ..., X_n) = P(X_2, ..., X_n|X_1)P(X_1) = P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1)
  \]
  \[
  = P(X_4, ..., X_n|X_1, X_2, X_3)P(X_3|X_2, X_1)P(X_2|X_1)P(X_1)
  \]
  \[
  = P(X_1) \prod_{i=2}^{n} P(X_i|X_{i-1}) \quad (\text{chain rule})
  \]
## From Joint Distributions to Marginal Distributions

Suppose we are given a joint distribution over two random variables $A, B$.

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$P(A = a, B = b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>0.25</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>0.35</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$P(A = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.55</td>
</tr>
<tr>
<td>Low</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$B$</th>
<th>$P(B = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.65</td>
</tr>
<tr>
<td>Low</td>
<td>0.35</td>
</tr>
</tbody>
</table>
From Joint Distributions to Marginal Distributions

- Suppose we are given a joint distribution over two random variables $A$, $B$.
- The marginal distributions of $A$ and $B$ can be computed as:

$$P(A = a) = \sum_{\forall b} P(A = a, B = b)$$

$$P(B = b) = \sum_{\forall a} P(A = a, B = b)$$
### From Joint Distributions to Marginal Distributions

Suppose we are given a joint distribution over two random variables $A$, $B$.

The marginal distributions of $A$ and $B$ can be computed as

$$P(A = a) = \sum_{\forall b} P(A = a, B = b)$$

$$P(B = b) = \sum_{\forall a} P(A = a, B = b)$$

More compactly written as

$$P(A) = \sum_{B} P(A, B)$$

$$P(B) = \sum_{A} P(A, B)$$

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$P(A = a, B = b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>0.25</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>0.35</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$P(A = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.55</td>
</tr>
<tr>
<td>Low</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$B$</th>
<th>$P(B = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.65</td>
</tr>
<tr>
<td>Low</td>
<td>0.35</td>
</tr>
</tbody>
</table>
What if there are \( n \) random variables?

- Suppose we are given a joint distribution over \( n \) random variables \( X_1, X_2, \ldots, X_n \)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>( P(A = a, B = b) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>0.25</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>0.35</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( P(A = a) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.55</td>
</tr>
<tr>
<td>Low</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( P(B = a) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.65</td>
</tr>
<tr>
<td>Low</td>
<td>0.35</td>
</tr>
</tbody>
</table>
What if there are $n$ random variables?

- Suppose we are given a joint distribution over $n$ random variables $X_1, X_2, ..., X_n$
- The marginal distributions over $X_1$ can be computed as

$$P(X_1 = x_1) = \sum_{\forall x_2, x_3, ..., x_n} P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$P(A = a, B = b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>0.25</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>0.35</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$P(A = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.55</td>
</tr>
<tr>
<td>Low</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$B$</th>
<th>$P(B = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.65</td>
</tr>
<tr>
<td>Low</td>
<td>0.35</td>
</tr>
</tbody>
</table>
What if there are $n$ random variables?

- Suppose we are given a joint distribution over $n$ random variables $X_1, X_2, \ldots, X_n$.
- The marginal distributions over $X_1$ can be computed as

\[
P(X_1 = x_1) = \sum_{\forall x_2, x_3, \ldots, x_n} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]

- More compactly written as

\[
P(X_1) = \sum_{X_2, X_3, \ldots, X_n} P(X_1, X_2, \ldots, X_n)
\]

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$P(A = a, B = b)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>0.3</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>0.25</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>0.35</td>
</tr>
<tr>
<td>Low</td>
<td>Low</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A$</th>
<th>$P(A = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.55</td>
</tr>
<tr>
<td>Low</td>
<td>0.45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$B$</th>
<th>$P(B = a)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.65</td>
</tr>
<tr>
<td>Low</td>
<td>0.35</td>
</tr>
</tbody>
</table>
Conditional Independence

- Two random variables $X$ and $Y$ are said to be independent if

$$P(X|Y) = P(X)$$
Conditional Independence

- Two random variables $X$ and $Y$ are said to be independent if

$$P(X|Y) = P(X)$$

- We denote this as $X \perp \perp Y$
Conditional Independence

- Two random variables $X$ and $Y$ are said to be independent if
  \[ P(X|Y) = P(X) \]

- We denote this as $X \perp \perp Y$

- In other words, knowing the value of $Y$ does not change our belief about $X$
Conditional Independence

- Two random variables $X$ and $Y$ are said to be independent if

\[ P(X|Y) = P(X) \]

- We denote this as $X \perp \perp Y$
- In other words, knowing the value of $Y$ does not change our belief about $X$
- We would expect *Grade* to be dependent on *Intelligence* but independent of *Weight*
Recall that by Chain Rule of Probability

\[ P(X, Y) = P(X)P(Y|X) \]

However, if \( X \) and \( Y \) are independent, then

\[ P(X,Y) = P(X)P(Y) \]

We denote this as \( X \perp \perp Y \)

In other words, knowing the value of \( Y \) does not change our belief about \( X \)

We would expect \textit{Grade} to be dependent on \textit{Intelligence} but independent of \textit{Weight}
Recall that by Chain Rule of Probability

\[ P(X, Y) = P(X)P(Y|X) \]

However, if \( X \) and \( Y \) are independent, then

\[ P(X, Y) = P(X)P(Y) \]

**Conditional Independence**

- Two random variables \( X \) and \( Y \) are said to be independent if

\[ P(X|Y) = P(X) \]

- We denote this as \( X \perp \perp Y \)

- In other words, knowing the value of \( Y \) does not change our belief about \( X \)

- We would expect *Grade* to be dependent on *Intelligence* but independent of *Weight*
Okay, we are now ready to move on to Bayesian Networks or Directed Graphical Models
Module 17.1: Why are we interested in Joint Distributions
In many real world applications, we have to deal with a large number of random variables.
In many real world applications, we have to deal with a large number of random variables.

For example, an oil company may be interested in computing the probability of finding oil at a particular location.
In many real world applications, we have to deal with a large number of random variables.

For example, an oil company may be interested in computing the probability of finding oil at a particular location.

This may depend on various (random) variables.

\[ P(Y, X_1, X_2, X_3, X_4, X_5, X_6) \]
In many real world applications, we have to deal with a large number of random variables.

For example, an oil company may be interested in computing the probability of finding oil at a particular location.

This may depend on various (random) variables.

The company is interested in knowing the joint distribution.

\[ P(Y, X_1, X_2, X_3, X_4, X_5, X_6) \]
But why joint distribution?

If we know the joint distribution, we can find answers to a bunch of interesting questions. Let us see some such questions of interest.

$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$
But why joint distribution?

Aren’t we just interested in $P(Y|X_1, X_2, ..., X_n)$?

$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$
But why joint distribution?

Aren’t we just interested in $P(Y|X_1, X_2, ..., X_n)$?

Well, if we know the joint distribution, we can find answers to a bunch of interesting questions
- But why joint distribution?
- Aren’t we just interested in \( P(Y|X_1, X_2, ..., X_n) \)?
- Well, if we know the joint distribution, we can find answers to a bunch of interesting questions
- Let us see some such questions of interest

\[ P(Y, X_1, X_2, X_3, X_4, X_5, X_6) \]
We can find the conditional distribution

\[ P(Y|X_1, \ldots, X_n) = \frac{P(Y, X_1, \ldots, X_n)}{\sum_{X_1, \ldots, X_n} P(Y, X_1, \ldots, X_n)} \]

We can find the marginal distribution,

\[ P(Y) = \sum_{X_1, \ldots, X_n} P(Y, X_1, \ldots, X_n) \]

We can find the conditional independencies,

\[ P(Y, X_1) = P(Y) P(X_1) \]

\[ P(Y, X_1, X_2, X_3, X_4, X_5, X_6) \]
We can find the conditional distribution

\[
P(Y|X_1, ..., X_n) = \frac{P(Y, X_1, ..., X_n)}{\sum_{X_1, ..., X_n} P(Y, X_1, ..., X_n)}
\]

We can find the marginal distribution,

\[
P(Y) = \sum_{X_1, ..., X_n} P(Y, X_1, X_2, ..., X_n)
\]
We can find the conditional distribution

$$P(Y|X_1,\ldots,X_n) = \frac{P(Y,X_1,\ldots,X_n)}{\sum_{X_1,\ldots,X_n} P(Y,X_1,\ldots,X_n)}$$

We can find the marginal distribution,

$$P(Y) = \sum_{X_1,\ldots,X_n} P(Y,X_1,X_2,\ldots,X_n)$$

We can find the conditional independencies,

$$P(Y,X_1) = P(Y)P(X_1)$$
Module 17.2: How do we represent a joint distribution
Let us return to the case of \( n \) random variables

\[
P(Y, X_1, X_2, X_3, X_4, X_5, X_6)
\]
Let us return to the case of $n$ random variables.

For simplicity assume each of these variables can take binary values.
Let us return to the case of $n$ random variables.

For simplicity assume each of these variables can take binary values.

To specify the joint distribution, we need to specify $2^n - 1$ values. Why not $(2^n)$?
Let us return to the case of $n$ random variables.

For simplicity assume each of these variables can take binary values.

To specify the joint distribution, we need to specify $2^n - 1$ values. Why not $2^n$?

If we specify these $2^n - 1$ values, we have an explicit representation for the joint distribution.
<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>...</th>
<th>$X_n$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>0.002</td>
</tr>
</tbody>
</table>

(Once the first $2^n - 1$ values are specified the last value is deterministic as the values need to sum to 1)
Challenges with explicit representation

- **Computational**: Expensive to manipulate and too large to store

(Once the first $2^n - 1$ values are specified the last value is deterministic as the values need to sum to 1)
Challenges with explicit representation

- **Computational**: Expensive to manipulate and too large to store
- **Cognitive**: Impossible to acquire so many numbers from a human

(Once the first $2^n - 1$ values are specified the last value is deterministic as the values need to sum to 1)
### Challenges with explicit representation

- **Computational:** Expensive to manipulate and too large to store

- **Cognitive:** Impossible to acquire so many numbers from a human

- **Statistical:** Need huge amounts of data to learn the parameters

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>...</th>
<th>$X_n$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>0.002</td>
</tr>
</tbody>
</table>

(Once the first $2^n - 1$ values are specified the last value is deterministic as the values need to sum to 1)
Module 17.3: Can we represent the joint distribution more compactly?
Consider the case of two random variables, Intelligence \((I)\) and SAT Scores \((S)\)
Consider the case of two random variables, Intelligence ($I$) and SAT Scores ($S$)

Assume that both are binary and take values from High(1), Low(0)
Consider the case of two random variables, Intelligence ($I$) and SAT Scores ($S$).

Assume that both are binary and take values from High(1), Low(0).

Here is one way of specifying the joint distribution:

<table>
<thead>
<tr>
<th>$I$</th>
<th>$S$</th>
<th>$P(I, S)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.665</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.035</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Consider the case of two random variables, Intelligence ($I$) and SAT Scores ($S$)

Assume that both are binary and take values from High(1), Low(0)

Here is one way of specifying the joint distribution

Of course, there are many such joint distributions possible

<table>
<thead>
<tr>
<th>$I$</th>
<th>$S$</th>
<th>$P(I, S)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.665</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.035</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.24</td>
</tr>
</tbody>
</table>
Consider the case of two random variables, Intelligence ($I$) and SAT Scores ($S$).

Assume that both are binary and take values from High(1), Low(0).

Here is one way of specifying the joint distribution.

Of course, there are many such joint distributions possible.
Consider the case of two random variables, Intelligence \((I)\) and SAT Scores \((S)\)

- Assume that both are binary and take values from High(1), Low(0)
- Here is one way of specifying the joint distribution
- Of course, there are many such joint distributions possible

<table>
<thead>
<tr>
<th>(I)</th>
<th>(S)</th>
<th>(P(I, S))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.665</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.035</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.24</td>
</tr>
</tbody>
</table>

This distribution has \((2^2 - 1 = 3)\) parameters.

Alternatively, the table has 4 rows but the last row is deterministic given the first 3 rows (or parameters)
- Note that there is a natural ordering in these two random variables.
Note that there is a natural ordering in these two random variables.

The SAT Score ($S$) presumably depends upon the Intelligence ($I$). An alternate and even more natural way to represent the same distribution is

$$P(I, S) = P(I) \times P(S|I)$$
<table>
<thead>
<tr>
<th></th>
<th>$i = 0$</th>
<th>$i = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$s = 0$</th>
<th>$s = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I = 0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I = 1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

- Note that there is a natural ordering in these two random variables.

- The SAT Score ($S$) presumably depends upon the Intelligence ($I$). An alternate and even more natural way to represent the same distribution is

$$P(I, S) = P(I) \times P(S|I)$$

- Instead of specifying the 4 entries in $P(I, S)$, we can specify 2 entries for $P(I)$ and 4 entries for $P(S|I)$.
\begin{align*}
\begin{array}{|c|c|c|}
\hline
 & i = 0 & i = 1 \\
\hline
P(I) & 0.7 & 0.3 \\
\hline
\end{array}
\end{align*}

\begin{align*}
\begin{array}{|c|c|c|}
\hline
 & s = 0 & s = 1 \\
\hline
P(S|I = 0) & 0.95 & 0.05 \\
P(S|I = 1) & 0.2 & 0.8 \\
\hline
\end{array}
\end{align*}

- What! So from 3 parameters we have gone to 6 parameters?
- Note that there is a natural ordering in these two random variables
- The SAT Score ($S$) presumably depends upon the Intelligence ($I$). An alternate and even more natural way to represent the same distribution is

$$P(I, S) = P(I) \times P(S|I)$$

- Instead of specifying the 4 entries in $P(I, S)$, we can specify 2 entries for $P(I)$ and 4 entries for $P(S|I)$
\[ \begin{array}{cc} i = 0 & i = 1 \\ \hline P(I) & 0.7 & 0.3 \\ \hline \end{array} \]

**no. of parameters = 1**

\[ \begin{array}{cc} s = 0 & s = 1 \\ \hline P(S|I = 0) & 0.95 & 0.05 \\ P(S|I = 1) & 0.2 & 0.8 \\ \hline \end{array} \]

**no. of parameters = 2**

- What! So from 3 parameters we have gone to 6 parameters?
- Well, not really! (remember sum for each row in the above table has to be 1)

Note that there is a natural ordering in these two random variables.

The SAT Score \((S)\) presumably depends upon the Intelligence \((I)\). An alternate and even more natural way to represent the same distribution is

\[ P(I, S) = P(I) \times P(S|I) \]

Instead of specifying the 4 entries in \( P(I, S) \), we can specify 2 entries for \( P(I) \) and 4 entries for \( P(S|I) \).
<table>
<thead>
<tr>
<th>$i$</th>
<th>$P(I)$</th>
<th>no. of parameters=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

| $s$ | $P(S|I=0)$ | $P(S|I=1)$ | no. of parameters=2 |
|-----|------------|------------|----------------------|
| 0   | 0.95       | 0.2        |                      |
| 1   | 0.05       | 0.8        |                      |

Note that there is a natural ordering in these two random variables.

The SAT Score ($S$) presumably depends upon the Intelligence ($I$). An alternate and even more natural way to represent the same distribution is:

$$P(I, S) = P(I) \times P(S|I)$$

Instead of specifying the 4 entries in $P(I, S)$, we can specify 2 entries for $P(I)$ and 4 entries for $P(S|I)$.
What have we achieved so far?

<table>
<thead>
<tr>
<th></th>
<th>i=0</th>
<th>i=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**no.of parameters=1**

<table>
<thead>
<tr>
<th></th>
<th>s=0</th>
<th>s=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I = 0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I = 1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**no.of parameters=2**
<table>
<thead>
<tr>
<th>P(I)</th>
<th>i=0</th>
<th>i=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

**no.of parameters=1**

| P(S|I = 0) | s=0 | s=1 |
|----------|-----|-----|
| 0.95     | 0.05|     |
| P(S|I = 1) | 0.2 | 0.8 |

**no.of parameters=2**

- What have we achieved so far?
- We were not able to reduce the number of parameters

This is known as conditional parameterization.
What have we achieved so far?

- We were not able to reduce the number of parameters
- But, we have a more natural way of representing the distribution

<table>
<thead>
<tr>
<th></th>
<th>$i=0$</th>
<th>$i=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

no.of parameters = 1

<table>
<thead>
<tr>
<th></th>
<th>$s=0$</th>
<th>$s=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I=0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I=1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

no.of parameters = 2
### What have we achieved so far?
- We were not able to reduce the number of parameters.
- But, we have a more natural way of representing the distribution.
- This is known as conditional parameterization.

---

<table>
<thead>
<tr>
<th></th>
<th>$i=0$</th>
<th>$i=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

No. of parameters = 1

<table>
<thead>
<tr>
<th></th>
<th>$s=0$</th>
<th>$s=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I=0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I=1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

No. of parameters = 2
Now consider a third random variable Grade ($G$).
Now consider a third random variable Grade ($G$)

Notice that none of these 3 variables are independent of each other
Now consider a third random variable Grade \((G)\)  

- Notice that none of these 3 variables are independent of each other  
- Grade and SAT Score are clearly correlated with Intelligence
Now consider a third random variable Grade \((G)\)

Notice that none of these 3 variables are independent of each other

Grade and SAT Score are clearly correlated with Intelligence

Grade and SAT Score are also correlated because we would expect

\[
P(G = 1 | S = 1) > P(G = 1 | S = 0)
\]
However, it is possible that the distribution satisfies a conditional independence.

Formally, we assume that \((S \perp G | I)\). Note that this is just an assumption.
- However, it is possible that the distribution satisfies a conditional independence.

- If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining $G$.

In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score.

Formally, we assume that $(S \perp G | I)$.

Note that this is just an assumption.
However, it is possible that the distribution satisfies a conditional independence.

If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining $G$.

In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score.
However, it is possible that the distribution satisfies a conditional independence.

If we know that \( I = H \), then it is possible that \( S = H \) does not give any extra information for determining \( G \).

In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score.

Formally, we assume that \( (S \perp G | I) \)
However, it is possible that the distribution satisfies a conditional independence.

If we know that \( I = H \), then it is possible that \( S = H \) does not give any extra information for determining \( G \).

In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score.

Formally, we assume that \((S \perp G | I)\).

Note that this is just an assumption.
We could argue that in many cases $S \not\perp G | I$.
- We could argue that in many cases $S \not\perp G|I$

- For example, a student might be intelligent, but we also have to factor in his/her ability to write in time bound exams
We could argue that in many cases $S \not\perp G|I$

For example, a student might be intelligent, but we also have to factor in his/her ability to write in time bound exams.

In which case $S$ and $G$ are not independent given $I$ (because the SAT score tells us about the ability to write time bound exams).
We could argue that in many cases $S \not\perp G|I$

For example, a student might be intelligent, but we also have to factor in his/her ability to write in time bound exams.

In which case $S$ and $G$ are not independent given $I$ (because the SAT score tells us about the ability to write time bound exams).

But, for this discussion, we will assume $S \perp G|I$.
Question

- Now let’s see the implication of this assumption
Question

- Now let’s see the implication of this assumption
- Does it simplify things in any way?
How many parameters do we need to specify $P(I, G, S)$?

$$\text{(2 \times 2 \times 3 - 1 = 11)}$$
• How many parameters do we need to specify $P(I, G, S)$?

$$(2 \times 2 \times 3 - 1 = 11)$$

• What if we use conditional parameterization by following the chain rule?


since $(S \perp G | I)$

We need the following distributions to fully specify the joint distribution.
• How many parameters do we need to specify \( P(I, G, S) \)?

\[
(2 \times 2 \times 3 - 1 = 11)
\]

• What if we use conditional parameterization by following the chain rule?

\[
P(I, G, S) = P(S, G|I)P(I)
\]
How many parameters do we need to specify $P(I, G, S)$?

$$\left(2 \times 2 \times 3 - 1 = 11\right)$$

What if we use conditional parameterization by following the chain rule?

$$P(I, G, S) = P(S, G|I)P(I)$$

$$= P(S|G, I)P(G|I)P(I)$$
- How many parameters do we need to specify \( P(I, G, S) \)?

\[
(2 \times 2 \times 3 - 1 = 11)
\]

- What if we use conditional parameterization by following the chain rule?

\[
P(I, G, S) = P(S, G|I)P(I) \\
= P(S|G, I)P(G|I)P(I) \\
= P(S|I)P(G|I)P(I)
\]
- How many parameters do we need to specify \( P(I, G, S) \)?

\[
(2 \times 2 \times 3 - 1 = 11)
\]

- What if we use conditional parameterization by following the chain rule?

\[
\]

since \((S \perp G | I)\)
• How many parameters do we need to specify $P(I, G, S)$?

$$(2 \times 2 \times 3 - 1 = 11)$$

• What if we use conditional parameterization by following the chain rule?


since $(S \perp G|I)$

• We need the following distributions to fully specify the joint distribution
- How many parameters do we need to specify $P(I, G, S)$?
  
  $(2 \times 2 \times 3 - 1 = 11)$

- What if we use conditional parameterization by following the chain rule?

  \[
  P(I, G, S) = P(S, G|I)P(I) \\
  = P(S|G, I)P(G|I)P(I) \\
  = P(S|I)P(G|I)P(I)
  \]

  since $(S \perp G|I)$

- We need the following distributions to fully specify the joint distribution:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$i = 0$</th>
<th>$i = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>$i = 0$</td>
<td>$i = 1$</td>
</tr>
<tr>
<td>---</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

no. of parameters = 1

- How many parameters do we need to specify $P(I, G, S)$?

$$2 \times 2 \times 3 - 1 = 11$$

- What if we use conditional parameterization by following the chain rule?

$$P(I, G, S) = P(S, G|I)P(I)$$

$$= P(S|G, I)P(G|I)P(I)$$

$$= P(S|I)P(G|I)P(I)$$

since $(S \perp G|I)$

- We need the following distributions to fully specify the joint distribution
<table>
<thead>
<tr>
<th></th>
<th>$i=0$</th>
<th>$i=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

no. of parameters = 1

<table>
<thead>
<tr>
<th></th>
<th>$s=0$</th>
<th>$s=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I=0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I=1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

- How many parameters do we need to specify $P(I, G, S)$?
  \[
  (2 \times 2 \times 3 - 1 = 11)
  \]

- What if we use conditional parameterization by following the chain rule?

\[
P(I, G, S) = P(S, G|I)P(I) \\
= P(S|G, I)P(G|I)P(I) \\
= P(S|I)P(G|I)P(I)
\]

since $(S \perp G|I)$

- We need the following distributions to fully specify the joint distribution
How many parameters do we need to specify $P(I,G,S)$?

$$2 \times 2 \times 3 - 1 = 11$$

What if we use conditional parameterization by following the chain rule?


since ($S \perp G|I$)

We need the following distributions to fully specify the joint distribution:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$i = 0$</th>
<th>$i = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

no.of parameters = 1

<table>
<thead>
<tr>
<th>$s$</th>
<th>$s = 0$</th>
<th>$s = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I = 0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I = 1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

no.of parameters = 2

<table>
<thead>
<tr>
<th>$g$</th>
<th>$g = A$</th>
<th>$g = B$</th>
<th>$g = C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(G</td>
<td>I=0)$</td>
<td>0.24</td>
<td>0.34</td>
</tr>
<tr>
<td>$P(G</td>
<td>I=1)$</td>
<td>0.74</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>$i = 0$</td>
<td>$i = 1$</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
</tbody>
</table>

**no. of parameters = 1**

<table>
<thead>
<tr>
<th></th>
<th>$s = 0$</th>
<th>$s = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I = 0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I = 1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**no. of parameters = 2**

<table>
<thead>
<tr>
<th></th>
<th>$g = A$</th>
<th>$g = B$</th>
<th>$g = C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(G</td>
<td>I = 0)$</td>
<td>0.2</td>
<td>0.34</td>
</tr>
<tr>
<td>$P(G</td>
<td>I = 1)$</td>
<td>0.74</td>
<td>0.17</td>
</tr>
</tbody>
</table>

**no. of parameters = 4**

**total no. of parameters = 7**

- How many parameters do we need to specify $P(I, G, S)$?

  $$(2 \times 2 \times 3 - 1 = 11)$$

- What if we use conditional parameterization by following the chain rule?


  since $(S \perp G|I)$

- We need the following distributions to fully specify the joint distribution
How many parameters do we need to specify $P(I, G, S)$?

$$2 \times 2 \times 3 - 1 = 11$$

What if we use conditional parameterization by following the chain rule?

$$P(I, G, S) = P(S, G|I)P(I)$$

$$= P(S|G, I)P(G|I)P(I)$$

$$= P(S|I)P(G|I)P(I)$$

since $(S \perp G|I)$

We need the following distributions to fully specify the joint distribution:

<table>
<thead>
<tr>
<th>$i = 0$</th>
<th>$i = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
</tr>
<tr>
<td>$P(S</td>
<td>I = 0)$</td>
</tr>
<tr>
<td>$P(S</td>
<td>I = 1)$</td>
</tr>
</tbody>
</table>

no. of parameters = 1

<table>
<thead>
<tr>
<th>$s=0$</th>
<th>$s=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(G—I=0)$</td>
<td>0.2</td>
</tr>
<tr>
<td>$P(G—I=1)$</td>
<td>0.74</td>
</tr>
</tbody>
</table>

no. of parameters = 4
\[
\begin{array}{|c|c|c|}
\hline
i & i = 0 & i = 1 \\
\hline
P(I) & 0.7 & 0.3 \\
\hline
\end{array}
\]

**no.of parameters=1**

\[
\begin{array}{|c|c|c|}
\hline
s & s=0 & s=1 \\
\hline
P(S|I = 0) & 0.95 & 0.05 \\
P(S|I = 1) & 0.2 & 0.8 \\
\hline
\end{array}
\]

**no.of parameters=2**

\[
\begin{array}{|c|c|c|}
\hline
g & g=A & g=B & g=C \\
\hline
P(G—I=0) & 0.2 & 0.34 & 0.46 \\
P(G—I=1) & 0.74 & 0.17 & 0.09 \\
\hline
\end{array}
\]

**no.of parameters=4**

**total no.of parameters=7**

- How many parameters do we need to specify \(P(I, G, S)\)?

\[
(2 \times 2 \times 3 - 1 = 11)
\]

- What if we use conditional parameterization by following the chain rule?

\[
\]

\[
= P(S|I)P(G|I)P(I)
\]

since \((S \perp G|I)\)

- We need the following distributions to fully specify the joint distribution
<table>
<thead>
<tr>
<th>$i$</th>
<th>$P(I)$</th>
<th>$i = 0$</th>
<th>$i = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

no. of parameters = 1

| $s$ | $P(S|I = 0)$ | $s = 0$ | $s = 1$ |
|-----|-------------|--------|--------|
|     |             | 0.95   | 0.05   |
|     |             | 0.2    | 0.8    |

no. of parameters = 2

<table>
<thead>
<tr>
<th>$g$</th>
<th>$P(G—I=0)$</th>
<th>$g = A$</th>
<th>$g = B$</th>
<th>$g = C$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.34</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.74</td>
<td>0.17</td>
<td>0.09</td>
</tr>
</tbody>
</table>

no. of parameters = 4

**total no. of parameters = 7**

- The alternate parameterization is more **natural** than that of the joint distribution
The alternate parameterization is more **natural** than that of the joint distribution

- The alternate parameterization is more **compact** than that of the joint distribution

<table>
<thead>
<tr>
<th></th>
<th>$i = 0$</th>
<th>$i = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(I)$</td>
<td>0.7</td>
<td>0.3</td>
</tr>
</tbody>
</table>

**no.of parameters**=1

<table>
<thead>
<tr>
<th></th>
<th>$s=0$</th>
<th>$s=1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(S</td>
<td>I=0)$</td>
<td>0.95</td>
</tr>
<tr>
<td>$P(S</td>
<td>I=1)$</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**no.of parameters**=2

<table>
<thead>
<tr>
<th></th>
<th>$g=A$</th>
<th>$g=B$</th>
<th>$g=C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(G</td>
<td>I=0)$</td>
<td>0.2</td>
<td>0.34</td>
</tr>
<tr>
<td>$P(G</td>
<td>I=1)$</td>
<td>0.74</td>
<td>0.17</td>
</tr>
</tbody>
</table>

**no.of parameters**=4

**total no.of parameters**=7
<table>
<thead>
<tr>
<th>i = 0</th>
<th>i = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(I) )</td>
<td>0.7</td>
</tr>
</tbody>
</table>

no. of parameters = 1

<table>
<thead>
<tr>
<th>s=0</th>
<th>s=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(S</td>
<td>I=0) )</td>
</tr>
<tr>
<td>( P(S</td>
<td>I=1) )</td>
</tr>
</tbody>
</table>

no. of parameters = 2

<table>
<thead>
<tr>
<th>g=A</th>
<th>g=B</th>
<th>g=C</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(G—I=0) )</td>
<td>0.2</td>
<td>0.34</td>
</tr>
<tr>
<td>( P(G—I=1) )</td>
<td>0.74</td>
<td>0.17</td>
</tr>
</tbody>
</table>

no. of parameters = 4

**total no. of parameters = 7**

- The alternate parameterization is more **natural** than that of the joint distribution
- The alternate parameterization is more **compact** than that of the joint distribution
- The alternate parameterization is more **modular**. (When we added \( G \), we could just reuse the tables for \( P(I) \) and \( P(S|I) \))
Module 17.4: Can we use a graph to represent a joint distribution?
Suppose we have $n$ random variables, all of which are independent given another random variable $C$.

\[
P(C, X_1, X_2, X_3, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i | C)
\]

This is called the Naive Bayes model. It makes the Naive assumption that $n$ pairs are independent given $C$.
Suppose we have $n$ random variables, all of which are independent given another random variable $C$.

The joint distribution factorizes as,

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

since $X_i \perp X_j | C$
• Suppose we have \( n \) random variables, all of which are independent given another random variable \( C \)

• The joint distribution factorizes as,

\[
P(C, X_1, ..., X_n) = P(C)P(X_1|C)P(X_2|X_1, C)P(X_3|X_2, X_1, C)...\]

since \( X_i \perp X_j | C \)
Suppose we have $n$ random variables, all of which are independent given another random variable $C$.

The joint distribution factorizes as,

$$P(C, X_1, ..., X_n) = P(C)P(X_1|C)$$
$$P(X_2|X_1, C)$$
$$P(X_3|X_2, X_1, C)$$...

$$= P(C) \prod_{i=1}^{n} P(X_i|C)$$

since $X_i \perp X_j | C$
This is called the Naive Bayes model

Suppose we have $n$ random variables, all of which are independent given another random variable $C$.

The joint distribution factorizes as,

$$P(C, X_1, \ldots, X_n) = P(C)P(X_1|C)$$
$$P(X_2|X_1, C)$$
$$P(X_3|X_2, X_1, C)\ldots$$

$$= P(C)\prod_{i=1}^{n} P(X_i|C)$$

since $X_i \perp X_j | C$
This is called the Naive Bayes model.

It makes the Naive assumption that \( {n \choose 2} \) pairs are independent given \( C \).

Suppose we have \( n \) random variables, all of which are independent given another random variable \( C \).

The joint distribution factorizes as,

\[
P(C, X_1, ..., X_n) = P(C) P(X_1|C) P(X_2|X_1, C) P(X_3|X_2, X_1, C) ... = P(C) \prod_{i=1}^{n} P(X_i|C)
\]

since \( X_i \perp X_j | C \).
Bayesian networks build on the intuitions that we developed for the Naive Bayes model.
• Bayesian networks build on the intuitions that we developed for the Naive Bayes model
• But they are not restricted to strong (naive) independence assumptions
- Bayesian networks build on the intuitions that we developed for the Naive Bayes model
- But they are not restricted to strong (naive) independence assumptions
- We use graphs to represent the joint distribution
- Bayesian networks build on the intuitions that we developed for the Naive Bayes model
- But they are not restricted to strong (naive) independence assumptions
- We use graphs to represent the joint distribution
- **Nodes:** Random Variables
Bayesian networks build on the intuitions that we developed for the Naive Bayes model.

But they are not restricted to strong (naive) independence assumptions.

We use graphs to represent the joint distribution.

**Nodes:** Random Variables

**Edges:** Indicate dependence
Let’s revisit the student example

- The grade now depends on student's Intelligence & exam’s Difficulty level.
- The SAT score depends on Intelligence.
- The recommendation Letter from the course instructor depends on the Grade.
Let’s revisit the student example

- We will introduce a few more random variables and independence assumptions
Let’s revisit the student example

- We will introduce a few more random variables and independence assumptions
- The grade now depends on student’s Intelligence & exam’s Difficulty level
Let’s revisit the student example

- We will introduce a few more random variables and independence assumptions
- The grade now depends on student’s Intelligence & exam’s Difficulty level
- The SAT score depends on Intelligence
Let’s revisit the student example

- We will introduce a few more random variables and independence assumptions
- The grade now depends on student’s Intelligence & exam’s Difficulty level
- The SAT score depends on Intelligence
- The recommendation Letter from the course instructor depends on the Grade
The Bayesian network contains a node for each random variable.
The Bayesian network contains a node for each random variable.

The edges denote the dependencies between the random variables.
The Bayesian network contains a node for each random variable.
The edges denote the dependencies between the random variables.
Each variable depends directly on its parents in the network.
The Bayesian network can be viewed as a data structure.
The Bayesian network can be viewed as a data structure

It provides a skeleton for representing a joint distribution compactly by factorization
The Bayesian network can be viewed as a data structure.

It provides a skeleton for representing a joint distribution compactly by factorization.

Let us see what this means.
- Each node is associated with a local probability model.
Each node is associated with a local probability model.

Local, because it represents the dependencies of each variable on its parents.
Each node is associated with a local probability model.

- Local, because it represents the dependencies of each variable on its parents.
- There are 5 such local probability models associated with the graph.

Each node is associated with a local probability model.

- Local, because it represents the dependencies of each variable on its parents.
- There are 5 such local probability models associated with the graph.
Each node is associated with a local probability model.

Local, because it represents the dependencies of each variable on its parents.

There are 5 such local probability models associated with the graph.

Each variable (in general) is associated with a conditional probability distribution (conditional on its parents).
The graph gives us a natural factorization for the joint distribution.

\[
P(I, D, G, S, L) = P(I) \cdot P(D) \cdot P(G | I, D) \cdot P(S | I) \cdot P(L | G)
\]

For example, \(P(I = 1, D = 0, G = B, S = 1, L = 0) = 0.3 \times 0.6 \times 0.08 \times 0.8 \times 0.4\).
The graph gives us a natural factorization for the joint distribution

In this case,

\[ P(I, D, G, S, L) = P(I)P(D) \]

\[ P(G|I, D)P(S|I)P(L|G) \]
The graph gives us a natural factorization for the joint distribution:

\[ P(I, D, G, S, L) = P(I)P(D) \]
\[ P(G|I, D)P(S|I)P(L|G) \]

For example,

\[ P(I = 1, D = 0, G = B, S = 1, L = 0) = 0.3 \times 0.6 \times 0.08 \times 0.8 \times 0.4 \]
The graph gives us a natural factorization for the joint distribution.

In this case,

\[
P(I, D, G, S, L) = P(I)P(D)P(G|I, D)P(S|I)P(L|G)
\]

For example,

\[
P(I = 1, D = 0, G = B, S = 1, L = 0) = 0.3 \times 0.6 \times 0.08 \times 0.8 \times 0.4
\]

The graph structure (nodes, edges) along with the conditional probability distribution is called a Bayesian Network.
Module 17.5: Different types of reasoning in a Bayesian network
New Notations

- We will denote $P(I = 0)$ by $P(i^0)$
New Notations

- We will denote $P(I = 0)$ by $P(i^0)$
- In general, we will denote $P(I = 0, D = 1, G = B, S = 1, L = 0)$ by $P(i^0, d^1, g^b, s^1, l^0)$
Causal Reasoning

- Here, we try to predict downstream effects of various factors

\[ P(l_1) = \sum_{I \epsilon (0,1)} \sum_{D \epsilon (0,1)} \sum_{S \epsilon (0,1)} \sum_{G \epsilon (A,B,C)} P(I,D,G,S,l_1) \]

\[ P(l_1) = 50.2\% \]

Mitesh M. Khapra
CS7015 (Deep Learning) : Lecture 17
Causal Reasoning

- Here, we try to predict downstream effects of various factors
- Let us consider an example
### Causal Reasoning

- Here, we try to predict downstream effects of various factors.
- Let us consider an example.
- What is the probability that a student will get a good recommendation letter, \( P(l^1) \)?

#### Graphical Representation

- **Difficulty** \( D \): Difficulty levels \( d^0, d^1 \) with probabilities \( 0.6, 0.4 \), respectively.
- **Intelligence** \( I \): Intelligence levels \( i^0, i^1 \) with probabilities \( 0.7, 0.3 \), respectively.
- **Grade** \( G \): Grade levels \( g^1, g^2, g^3 \) with conditional probabilities:
  - \( i^0, d^0 \): \( 0.3 \) \( g^1 \), \( 0.4 \) \( g^2 \), \( 0.3 \) \( g^3 \)
  - \( i^0, d^1 \): \( 0.05 \) \( g^1 \), \( 0.25 \) \( g^2 \), \( 0.7 \) \( g^3 \)
  - \( i^1, d^0 \): \( 0.9 \) \( g^1 \), \( 0.08 \) \( g^2 \), \( 0.02 \) \( g^3 \)
  - \( i^1, d^1 \): \( 0.5 \) \( g^1 \), \( 0.3 \) \( g^2 \), \( 0.2 \) \( g^3 \)
- **Letter** \( L \): Letter grades \( l^0, l^1 \) with conditional probabilities:
  - \( i^0 \): \( 0.95 \) \( l^0 \), \( 0.05 \) \( l^1 \)
  - \( i^1 \): \( 0.2 \) \( l^0 \), \( 0.8 \) \( l^1 \)
- **SAT** \( S \): SAT scores \( s^0, s^1 \) with conditional probabilities:
  - \( i^0 \): \( 0.9 \) \( s^0 \), \( 0.1 \) \( s^1 \)
  - \( i^1 \): \( 0.4 \) \( s^0 \), \( 0.6 \) \( s^1 \)
  - \( i^1 \): \( 0.99 \) \( s^0 \), \( 0.01 \) \( s^1 \)
Causal Reasoning

- Here, we try to predict downstream effects of various factors
- Let us consider an example
- What is the probability that a student will get a good recommendation letter, \( P(l^1) \)?

\[
P(l^1) = \sum_{I \in \{0,1\}} \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(I, D, G, S, l^1)
\]
\[ P(l^1) = \sum_{I \in \{0,1\}} \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(I, D, G, S, l^1) \]
\[ P(l^1) = \sum_{I \in \{0,1\}} \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(I, D, G, S, l^1) \]

\[ = \sum_{I \in \{0,1\}} P(I) \sum_{D \in \{0,1\}} P(D|I) \sum_{S \in \{0,1\}} P(S|I, D) \sum_{G \in \{A,B,C\}} P(G|I, D, S).P(l^1|G, I, D, S) \]
\[ P(l^1) = \sum_{I \in \{0,1\}} \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(I, D, G, S, l^1) \]

\[ = \sum_{I \in \{0,1\}} P(I) \sum_{D \in \{0,1\}} P(D|I) \sum_{S \in \{0,1\}} P(S|I, D) \sum_{G \in \{A,B,C\}} P(G|I, D, S).P(l^1|G, I, D, S) \]

\[ = \sum_{I \in \{0,1\}} P(I) \sum_{D \in \{0,1\}} P(D) \sum_{S \in \{0,1\}} P(S|I) \sum_{G \in \{A,B,C\}} P(G|I, D).P(l^1|G) \]
\[ P(l^1) = \sum_{I \in (0,1)} P(I) \sum_{D \in (0,1)} P(D) \sum_{S \in (0,1)} P(S|I) \sum_{G \in (A,B,C)} P(G|I,D) P(l^1|G) \]

Similarly using the other tables, we can evaluate this equation:

\[ P(l^1) = 0.502 \]

<table>
<thead>
<tr>
<th>( l^0 )</th>
<th>( l^1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g^1 )</td>
<td>0.1</td>
</tr>
<tr>
<td>( g^2 )</td>
<td>0.4</td>
</tr>
<tr>
<td>( g^3 )</td>
<td>0.99</td>
</tr>
</tbody>
</table>
\[ P(l^1) = \sum_{I \epsilon (0,1)} P(I) \sum_{D \epsilon (0,1)} P(D) \sum_{S \epsilon (0,1)} P(S|I) \sum_{G \epsilon (A,B,C)} P(G|I,D)P(l^1|G) \]

\[
= \sum_{I \epsilon (0,1)} P(I) \sum_{D \epsilon (0,1)} P(D) \sum_{S \epsilon (0,1)} P(S|I) 0.9(P(g^1|I,D)) + 0.6(P(g^2|I,D)) + 0.01(P(g^3|I,D))
\]
\[ P(l^1) = \sum_{I \in (0,1)} P(I) \sum_{D \in (0,1)} P(D) \sum_{S \in (0,1)} P(S|I) \sum_{G \in (A,B,C)} P(G|I,D)P(l^1|G) \]

\[ = \sum_{I \in (0,1)} P(I) \sum_{D \in (0,1)} P(D) \sum_{S \in (0,1)} P(S|I)0.9(P(g^1|I,D)) + 0.6(P(g^2|I,D)) + 0.01(P(g^3|I,D)) \]

Similarly using the other tables, we can evaluate this equation.
\[ P(l^1) = \sum_{I \in (0,1)} P(I) \sum_{D \in (0,1)} P(D) \sum_{S \in (0,1)} P(S|I) \sum_{G \in (A,B,C)} P(G|I,D) P(l^1|G) \]

\[ = \sum_{I \in (0,1)} P(I) \sum_{D \in (0,1)} P(D) \sum_{S \in (0,1)} P(S|I) 0.9(P(g^1|I,D)) + 0.6(P(g^2|I,D)) + 0.01(P(g^3|I,D)) \]

- Similarly using the other tables, we can evaluate this equation

\[ P(l^1) = 0.502 \]
Causal Reasoning

- Now what if we start adding information about the factors that could influence $l^1$?
Causal Reasoning

- Now what if we start adding information about the factors that could influence $l^1$?
- What if someone reveals that the student is not intelligent?
Causal Reasoning

- Now what if we start adding information about the factors that could influence $l^1$?
- What if someone reveals that the student is not intelligent?
- Intelligence will affect the score and hence the grade
\[
P(l^1 | i^0) = \frac{P(l^1, i^0)}{P(i^0)}
\]
\[ P(l^1|i^0) = \frac{P(l^1,i^0)}{P(i^0)} \]

\[ P(l^1,i^0) = \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(i^0, D, G, S, l^1) \]
\[ P(l^1|i^0) = \frac{P(l^1, i^0)}{P(i^0)} \]

\[ P(l^1, i^0) = \sum_{D \in \{0, 1\}} \sum_{S \in \{0, 1\}} \sum_{G \in \{A, B, C\}} P(i^0, D, G, S, l^1) \]

\[ = \sum_{D \in \{0, 1\}} P(D) \sum_{S \in \{0, 1\}} P(S|i^0) \sum_{G \in \{A, B, C\}} P(G|D, i^0)P(l^1|G) \]
\[ P(l^1|i^0) = \frac{P(l^1,i^0)}{P(i^0)} \]

\[ P(l^1,i^0) = \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(i^0, D, G, S, l^1) \]

\[ = \sum_{D \in \{0,1\}} P(D) \sum_{S \in \{0,1\}} P(S|i^0) \sum_{G \in \{A,B,C\}} P(G|D, i^0) P(l^1|G) \]

\[ = \sum_{D \in \{0,1\}} P(D) \sum_{S \in \{0,1\}} P(S|i^0) \sum_{G \in \{A,B,C\}} 0.9P(g^1|D, i^0) + 0.6P(g^2|D, i^0) + 0.01P(g^3|D, i^0) \]
\[ P(l^1 | i^0) = \frac{P(l^1, i^0)}{P(i^0)} \]

\[ P(l^1, i^0) = \sum_{D \in \{0,1\}} \sum_{S \in \{0,1\}} \sum_{G \in \{A,B,C\}} P(i^0, D, G, S, l^1) \]

\[ = \sum_{D \in \{0,1\}} P(D) \sum_{S \in \{0,1\}} P(S|i^0) \sum_{G \in \{A,B,C\}} P(G|D, i^0)P(l^1|G) \]

\[ = \sum_{D \in \{0,1\}} P(D) \sum_{S \in \{0,1\}} P(S|i^0) \sum_{G \in \{A,B,C\}} 0.9P(g^1|D, i^0) + 0.6P(g^2|D, i^0) + 0.01P(g^3|D, i^0) \]

\[ P(l^1 | i^0) = 0.389 \]
Causal Reasoning

- What if the course was easy?

![Diagram with nodes and edges representing Difficulty (D), Intelligence (I), Grade (G), SAT (S), and Letter (L).]

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d^0$ 0.6</td>
<td>$i^0$ 0.7</td>
</tr>
<tr>
<td>$d^1$ 0.4</td>
<td>$i^1$ 0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>SAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g^1$ 0.3</td>
<td>$s^1$ 0.2</td>
</tr>
<tr>
<td>$g^2$ 0.4</td>
<td>$s^2$ 0.08</td>
</tr>
<tr>
<td>$g^3$ 0.99</td>
<td>$s^3$ 0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^0$ 0.1</td>
</tr>
<tr>
<td>$l^1$ 0.9</td>
</tr>
</tbody>
</table>

```
P(l^1 | i^0, d^0) = \sum_{G \epsilon \{A, B, C\}} \sum_{S \epsilon \{0, 1\}} P(i^0, d^0, G, S, l^1)P(l^1 | i^0, d^1) = 0.513 \text{ (increases)}
```
Causal Reasoning

- What if the course was easy?
- A not so intelligent student may still be able to get a good grade and hence a good letter.
Causal Reasoning

- What if the course was easy?
- A not so intelligent student may still be able to get a good grade and hence a good letter

\[
P(l^1|i^0, d^0) = \sum_{G \in \{A, B, C\}} \sum_{S \in \{0, 1\}} P(i^0, d^0, G, S, l^1)
\]
Causal Reasoning

- What if the course was easy?
- A not so intelligent student may still be able to get a good grade and hence a good letter

\[
P(l^1|i^0, d^0) = \sum_{G \in \{A, B, C\}} \sum_{S \in \{0, 1\}} P(i^0, d^0, G, S, l^1)
\]

\[
P(l^1|i^0, d^1) = 0.513 \text{ (increases)}
\]
Evidential Reasoning

- Here, we reason about causes by looking at their effects

What is the probability of the student being intelligent?

What is the probability of the course being difficult?
Evidential Reasoning

- Here, we reason about causes by looking at their effects
- What is the probability of the student being intelligent?

\[ P(i^1) = ? \]
Evidential Reasoning

- Here, we reason about causes by looking at their effects
- What is the probability of the student being intelligent?

\[ P(i^1) = 0.3 \]
Evidential Reasoning

- Here, we reason about causes by looking at their effects
- What is the probability of the student being intelligent?
- What is the probability of the course being difficult?

\[ P(i^1) = 0.3 \]
\[ P(d^1) = ? \]
Evidential Reasoning

- Here, we reason about causes by looking at their effects
- What is the probability of the student being intelligent?
- What is the probability of the course being difficult?

\[
P(i^1) = 0.3 \\
P(d^1) = 0.4
\]
Evidential Reasoning

- Here, we reason about causes by looking at their effects
- What is the probability of the student being intelligent?
- What is the probability of the course being difficult?
- Now let us see what happens if we observe some effects

\[
P(i^1) = 0.3 \\
P(d^1) = 0.4
\]
\[ P(i^1) = 0.3 \]
\[ P(d^1) = 0.4 \]

**Evidential Reasoning**

- What if someone tells us that the student secured \( C \) grade?

---

Difficulty \( D \)

Intelligence \( I \)

Grade \( G \)

Letter \( L \)

SAT \( S \)
\[ P(i_1) = 0.3 \]
\[ P(d_1) = 0.4 \]
\[ P(i_1|g^3) = 0.079 (drops) \]
\[ P(d_1|g^3) = 0.629 (increases) \]

**Evidential Reasoning**
- What if someone tells us that the student secured \( C \) grade?

---

Difficulty  
Intelligence

\( D \)
\( I \)

Grade  
Letter

\( G \)
\( L \)

SAT

\( S \)
\[
P(i^1) = 0.3 \\
P(d^1) = 0.4 \\
P(i^1|g^3) = 0.079 (\text{drops}) \\
P(d^1|g^3) = 0.629 (\text{increases})
\]

**Evidential Reasoning**

- What if someone tells us that the student secured \( C \) grade?
- What if instead of getting to know the grade, we get to know that the student got a poor recommendation letter?
\[ P(i^1) = 0.3 \]
\[ P(d^1) = 0.4 \]
\[ P(i^1 | g^3) = 0.079 \text{(drops)} \]
\[ P(d^1 | g^3) = 0.629 \text{(increases)} \]
\[ P(i^1 | l^0) = 0.14 \text{(drops)} \]

**Evidential Reasoning**

- What if someone tells us that the student secured \( C \) grade?
- What if instead of getting to know the grade, we get to know that the student got a poor recommendation letter?
\[ P(i^1) = 0.3 \]
\[ P(d^1) = 0.4 \]
\[ P(i^1|g^3) = 0.079 (drops) \]
\[ P(d^1|g^3) = 0.629 (increases) \]
\[ P(i^1|l^0) = 0.14 (drops) \]

**Evidential Reasoning**

- What if someone tells us that the student secured C grade?
- What if instead of getting to know the grade, we get to know that the student got a poor recommendation letter?
- What if we know about the grade as well as the recommendation letter?
$P(i^1) = 0.3$
$P(d^1) = 0.4$
$P(i^1|g^3) = 0.079$ (drops)
$P(d^1|g^3) = 0.629$ (increases)
$P(i^1|l^0) = 0.14$ (drops)
$P(l^1|l^0, g^3) = 0.079$ (same as $P(i^1|g^3)$)

Evidential Reasoning

- What if someone tells us that the student secured $C$ grade?
- What if instead of getting to know the grade, we get to know that the student got a poor recommendation letter?
- What if we know about the grade as well as the recommendation letter?
\[ P(i^1) = 0.3 \]
\[ P(d^1) = 0.4 \]
\[ P(i^1|g^3) = 0.079 (\text{drops}) \]
\[ P(d^1|g^3) = 0.629 (\text{increases}) \]
\[ P(i^1|l^0) = 0.14 (\text{drops}) \]
\[ P(l^1|l^0,g^3) = 0.079 \]
(same as \( P(i^1|g^3) \))

**Evidential Reasoning**

- What if someone tells us that the student secured \( C \) grade?
- What if instead of getting to know the grade, we get to know that the student got a poor recommendation letter?
- What if we know about the grade as well as the recommendation letter?
- The last case is interesting! (We will return to it later)
Explaining Away

- Here, we see how different causes of the same effect can interact

\[ P(i^1) = 0.3 \]

\[ P(i^1 | g_3) = 0.079 \text{ (drops)} \]

\[ P(i^1 | g_3, d_1) = 0.11 \text{ (improves)} \]
$P(i^1) = 0.3$
$P(i^1|g^3) = 0.079(drops)$

Explaining Away

- Here, we see how different causes of the same effect can interact
- We already saw how knowing the grade influences our estimate of intelligence


\[ P(i^1) = 0.3 \]

\[ P(i^1|g^3) = 0.079 \text{(drops)} \]

**Explaining Away**

- Here, we see how different causes of the same effect can interact.
- We already saw how knowing the grade influences our estimate of intelligence.
- What if we were told the course was difficult?

Diagram:

- Difficulty \( D \)
- Intelligence \( I \)
- Grade \( G \)
- SAT \( S \)
- Letter \( L \)
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 (\text{drops}) \]
\[ P(i^1|g^3, d^1) = 0.11 (\text{improves}) \]

**Explaining Away**
- Here, we see how different causes of the same effect can interact.
- We already saw how knowing the grade influences our estimate of intelligence.
- What if we were told the course was difficult?
- Our belief in the student’s intelligence improves.
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 \text{(drops)} \]
\[ P(i^1|g^3, d^1) = 0.11 \text{(improves)} \]

### Explaining Away

- Here, we see how different causes of the same effect can interact.
- We already saw how knowing the grade influences our estimate of intelligence.
- What if we were told the course was difficult?
- Our belief in the student’s intelligence improves.
- Why? Let us see.
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 \]
\[ P(i^1|g^3, d^1) = 0.11 \]

**Explaining Away**

- Knowing that the course was difficult explains away the bad grade.
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 \]
\[ P(i^1|g^3, d^1) = 0.11 \]

**Explaining Away**

- Knowing that the course was difficult explains away the bad grade
- “Oh! Maybe the course was just too difficult and the student might have received a bad grade despite being intelligent!”
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 \]
\[ P(i^1|g^3,d^1) = 0.11 \]

**Explaining Away**

- Knowing that the course was difficult explains away the bad grade
- “Oh! Maybe the course was just too difficult and the student might have received a bad grade despite being intelligent!”
- The explaining away effect could be even more dramatic
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 \]
\[ P(i^1|g^3,d^1) = 0.11 \]

**Explaining Away**
- Knowing that the course was difficult explains away the bad grade
  - “Oh! Maybe the course was just too difficult and the student might have received a bad grade despite being intelligent!”
- The explaining away effect could be even more dramatic
- Let us consider the case when the grade was \( B \)
\[ P(i^1) = 0.3 \]
\[ P(i^1|g^3) = 0.079 \]
\[ P(i^1|g^3, d^1) = 0.11 \]
\[ P(i^1|g^2) = 0.175 \]
\[ P(i^1|g^2, d^1) = 0.34 \]

**Explaining Away**

- Knowing that the course was difficult explains away the bad grade
- “Oh! Maybe the course was just too difficult and the student might have received a bad grade despite being intelligent!”
- The explaining away effect could be even more dramatic
- Let us consider the case when the grade was \( B \)
\[ P(d^1) = 0.40 \]
\[ P(d^1|g^3) = 0.629 \]

**Explaining Away**

- Suppose we know that the student had a high SAT Score, what happens to our belief about the difficulty of the course?
\[ P(d^1) = 0.40 \]
\[ P(d^1|g^3) = 0.629 \]
\[ P(d^1|s^1,g^3) = 0.76 \]

**Explaining Away**
- Suppose we know that the student had a high SAT Score, what happens to our belief about the difficulty of the course?
\[ P(d^1) = 0.40 \]
\[ P(d^1 | g^3) = 0.629 \]
\[ P(d^1 | s^1, g^3) = 0.76 \]

**Explaining Away**

- Suppose we know that the student had a high SAT Score, what happens to our belief about the difficulty of the course?
- Knowing that the SAT score was high tells us that the student seems intelligent and perhaps the reason why he scored a poor grade is that the course was difficult.
Module 17.6: Independencies encoded by a Bayesian network (Case 1: Node and its parents)
Why do we care about independencies encoded in a Bayesian network?

We saw that if two variables are independent then the chain rule gets simplified, resulting in simpler factors which in turn reduces the number of parameters.

In the extreme case, we say that in the Bayesian network model, each factor was very simple (just $P(X_i|Y)$) and as a result each factor just added 3 parameters.

The more the number of independencies, the fewer the parameters and the lesser is the inference time.

For example, if we want to compute the marginal $P(S)$ then we just need to sum over the values of $I$ and not on any other variables.

Hence we are interested in finding the independencies encoded in a Bayesian network.
Why do we care about independencies encoded in a Bayesian network?

- We saw that if two variables are independent then the chain rule gets simplified, resulting in simpler factors which in turn reduces the number of parameters.
Why do we care about independencies encoded in a Bayesian network?

- We saw that if two variables are independent then the chain rule gets simplified, resulting in simpler factors which in turn reduces the number of parameters.

- In the extreme case, we say that in the Bayesian network model, each factor was very simple (just $P(X_i|Y)$) and as a result each factor just added 3 parameters.
Why do we care about independencies encoded in a Bayesian network?

- We saw that if two variables are independent then the chain rule gets simplified, resulting in simpler factors which in turn reduces the number of parameters.

- In the extreme case, we say that in the Bayesian network model, each factor was very simple (just $P(X_i|Y)$) and as a result each factor just added 3 parameters.

- The more the number of independencies, the fewer the parameters and the lesser is the inference time.
Why do we care about independencies encoded in a Bayesian network?

- We saw that if two variables are independent then the chain rule gets simplified, resulting in simpler factors which in turn reduces the number of parameters.

- In the extreme case, we say that in the Bayesian network model, each factor was very simple (just $P(X_i|Y)$) and as a result each factor just added 3 parameters.

- The more the number of independencies, the fewer the parameters and the lesser is the inference time.

- For example, if we want to compute the marginal $P(S)$ then we just need to sum over the values of $I$ and not on any other variables.
Why do we care about independencies encoded in a Bayesian network?

- We saw that if two variables are independent then the chain rule gets simplified, resulting in simpler factors which in turn reduces the number of parameters.
- In the extreme case, we say that in the Bayesian network model, each factor was very simple (just $P(X_i|Y)$) and as a result each factor just added 3 parameters.
- The more the number of independencies, the fewer the parameters and the lesser is the inference time.
- For example, if we want to the compute the marginal $P(S)$ then we just need to sum over the values of $I$ and not on any other variables.
- Hence we are interested in finding the independencies encoded in a Bayesian network.
In general, given \( n \) random variables, we are interested in knowing if

- \( X_i \perp X_j \)
In general, given $n$ random variables, we are interested in knowing if

- $X_i \perp X_j$
- $X_i \perp X_j | Z$, where $Z \subseteq X_1, X_2, ..., X_n/X_i, X_j$
In general, given $n$ random variables, we are interested in knowing if

- $X_i \perp X_j$
- $X_i \perp X_j | Z$, where $Z \subseteq X_1, X_2, ..., X_n/X_i, X_j$
- Let us answer some of the questions for our student Bayesian Network
To understand this let us return to our student example

- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)

Rule: A node is not independent of its parents.
To understand this let us return to our student example

First, let us see some independencies which clearly do not exist in the graph

- $D \perp G$? (No, by construction)
- $G \perp S$? (No, by construction)
- $G \perp I$? (No, by construction)
- $S \perp I$? (No, by construction)
To understand this let us return to our student example

First, let us see some independencies which clearly do not exist in the graph

- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)

Rule: A node is not independent of its parents.
To understand this let us return to our student example

First, let us see some independencies which clearly do not exist in the graph

- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)

Rule: A node is not independent of its parents
To understand this let us return to our student example

First, let us see some independencies which clearly do not exist in the graph

Is $L \perp G$? (No, by construction)
Is $G \perp D$? (No, by construction)
Is $G \perp I$? (No, by construction)
To understand this let us return to our student example.

First, let us see some independencies which clearly do not exist in the graph.

- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)
To understand this let us return to our student example

First, let us see some independencies which clearly do not exist in the graph

- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)

**Rule?**
To understand this let us return to our student example

First, let us see some independencies which clearly do not exist in the graph

- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)

**Rule:** A node is not independent of its parents
No, the instructor is not going to look at the SAT score but the grade. Rule: A node is not independent of its parents even when we are given the values of other variables. Let us focus on $G$ and $L$. We already know that $G \not \perp L$. What if we know the value of $I$? Does $G$ become independent of $L$? No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.) If we know the value of $D$, does $G$ become independent of $L$. No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course). What if we know the value of $S$? Does $G$ become independent of $L$?
No, the instructor is not going to look at the SAT score but the grade.

Rule: A node is not independent of its parents even when we are given the values of other variables.

Let us focus on $G$ and $L$.

We already know that $G \not\perp L$.

- What if we know the value of $I$? Does $G$ become independent of $L$?
  - No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)

- If we know the value of $D$, does $G$ become independent of $L$?
  - No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course)

- What if we know the value of $S$? Does $G$ become independent of $L$?

Mitesh M. Khapra
CS7015 (Deep Learning) : Lecture 17
Let us focus on $G$ and $L$.

We already know that $G \not\perp L$.

What if we know the value of $I$? Does $G$ become independent of $L$?
Let us focus on $G$ and $L$.

- We already know that $G \not\perp L$.
- What if we know the value of $I$? Does $G$ become independent of $L$?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
Let us focus on $G$ and $L$.

We already know that $G \not\perp L$.

What if we know the value of $I$? Does $G$ become independent of $L$?

No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)

If we know the value of $D$, does $G$ become independent of $L$.
Let us focus on $G$ and $L$.

- We already know that $G \not\perp L$.
- What if we know the value of $I$? Does $G$ become independent of $L$?
  - No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)

- If we know the value of $D$, does $G$ become independent of $L$.
  - No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course)
- Let us focus on $G$ and $L$.
- We already know that $G \not\perp L$.
- What if we know the value of $I$? Does $G$ become independent of $L$?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of $D$, does $G$ become independent of $L$.
- No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course)
- What if we know the value of $S$? Does $G$ become independent of $L$?
No, the instructor is not going to look at the SAT score but the grade.

- Let us focus on $G$ and $L$.
- We already know that $G \not\perp L$.
- What if we know the value of $I$? Does $G$ become independent of $L$?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of $D$, does $G$ become independent of $L$.
- No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course)
- What if we know the value of $S$? Does $G$ become independent of $L$?

\begin{itemize}
\item $D$
\item $I$
\item $G$
\item $S$
\item $L$
\end{itemize}
No, the instructor is not going to look at the SAT score but the grade.

**Rule?**

- Let us focus on $G$ and $L$.
- We already know that $G \not\perp L$.
- What if we know the value of $I$? Does $G$ become independent of $L$?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of $D$, does $G$ become independent of $L$.
- No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course)
- What if we know the value of $S$? Does $G$ become independent of $L$?
No, the instructor is not going to look at the SAT score but the grade.

**Rule:** A node is not independent of its parents even when we are given the values of other variables.

- Let us focus on $G$ and $L$.
- We already know that $G \not\perp L$.
- What if we know the value of $I$? Does $G$ become independent of $L$?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of $D$, does $G$ become independent of $L$?
- No (intuitively, the course may be easy or hard but the letter would depend on the performance in the course.)
- What if we know the value of $S$? Does $G$ become independent of $L$?
The same argument can be made about the following pairs:

- \( D \perp \perp G \) (even when other variables are given)
- \( I \perp \perp G \) (even when other variables are given)
- \( S \perp \perp I \) (even when other variables are given)
The same argument can be made about the following pairs:

- \(G \not\perp D\) (even when other variables are given)
- \(G \not\perp I\) (even when other variables are given)
- \(S \not\perp I\) (even when other variables are given)
Rule:

A node is not independent of its parents even when we are given the values of other variables.

The same argument can be made about the following pairs:

- \( G \not\perp D \) (even when other variables are given)
- \( G \not\perp I \) (even when other variables are given)
- \( S \not\perp I \) (even when other variables are given)
The same argument can be made about the following pairs:

- $G \not\perp D$ (even when other variables are given)
- $G \not\perp I$ (even when other variables are given)
- $S \not\perp I$ (even when other variables are given)
The same argument can be made about the following pairs

- \( G \not\perp D \) (even when other variables are given)
- \( G \not\perp I \) (even when other variables are given)
- \( S \not\perp I \) (even when other variables are given)

Rule?
Rule: A node is not independent of its parents even when we are given the values of other variables

- The same argument can be made about the following pairs
- \( G \not\perp D \) (even when other variables are given)
- \( G \not\perp I \) (even when other variables are given)
- \( S \not\perp I \) (even when other variables are given)
Module 17.7: Independencies encoded by a Bayesian network (Case 2: Node and its non-parents)
Now let’s look at the relation between a node and its non-parent nodes.

- Knowing the SAT score tells us about $I$, which in turn tells us something about $G$ and hence $L$.
- Hence we expect $P(l_1|s_1) > P(l_1|s_0)$.
- Similarly, we can argue $L \not\perp D$ and $L \not\perp I$.
Now let’s look at the relation between a node and its non-parent nodes

Is $L \perp S$?
Now let’s look at the relation between a node and its non-parent nodes

- Is $L \perp S$?
- No, knowing the SAT score tells us about $I$ which in turn tells us something about $G$ and hence $L$
Now let’s look at the relation between a node and its non-parent nodes

Is $L \perp S$?

No, knowing the SAT score tells us about $I$ which in turn tells us something about $G$ and hence $L$

Hence we expect $P(l^1|s^1) > P(l^1|s^0)$
Now let’s look at the relation between a node and its non-parent nodes

- Is \( L \perp S \)?
- No, knowing the SAT score tells us about \( I \) which in turn tells us something about \( G \) and hence \( L \)
- Hence we expect \( P(l^1 | s^1) > P(l^1 | s^0) \)
- Similarly we can argue \( L \not\perp D \) and \( L \not\perp I \)
But what if we know the value of $G$?

Yes, the grade completely determines the recommendation letter. Once we know the grade, other variables do not add any information. Hence $(L \perp S \mid G)$. Similarly, we can argue $(L \perp I \mid G)$ and $(L \perp D \mid G)$. 

Mitesh M. Khapra
CS7015 (Deep Learning) : Lecture 17
But what if we know the value of $G$?

Is $(L \perp S) | G$?

Yes, the grade completely determines the recommendation letter. Once we know the grade, other variables do not add any information. Hence $(L \perp S) | G$.

Similarly, we can argue $(L \perp I) | G$ and $(L \perp D) | G$. 

Mitesh M. Khapra
CS7015 (Deep Learning) : Lecture 17
But what if we know the value of $G$?
Is $(L \perp S)|G$?
Yes, the grade completely determines the recommendation letter.
But what if we know the value of $G$?

Is $(L \perp S)|G$?

Yes, the grade completely determines the recommendation letter.

Once we know the grade, other variables do not add any information.
But what if we know the value of $G$?
Is $(L \perp S)|G$?
Yes, the grade completely determines the recommendation letter.
Once we know the grade, other variables do not add any information.
Hence $(L \perp S)|G$.
But what if we know the value of $G$?

Is $(L \perp S)|G$?

Yes, the grade completely determines the recommendation letter

Once we know the grade, other variables do not add any information

Hence $(L \perp S)|G$

Similarly we can argue $(L \perp I)|G$ and $(L \perp D)|G$
But, wait a minute!

- The instructor may also want to look at the SAT score in addition to the grade.

Well, we "assumed" that the instructor only relies on the grade. That was our "belief" of how the world works and hence we drew the network accordingly.
• But, wait a minute!
• The instructor may also want to look at the SAT score in addition to the grade.
But, wait a minute!
The instructor may also want to look at the SAT score in addition to the grade.
Well, we “assumed” that the instructor only relies on the grade.
But, wait a minute!

The instructor may also want to look at the SAT score in addition to the grade.

Well, we “assumed” that the instructor only relies on the grade.

That was our “belief” of how the world works.
But, wait a minute!
The instructor may also want to look at the SAT score in addition to the grade.
Well, we “assumed” that the instructor only relies on the grade.
That was our “belief” of how the world works.
And hence we drew the network accordingly.
- Of course we are free to change our assumptions
- Of course we are free to change our assumptions

- We may want to assume that the instructor also looks at the SAT score
Of course we are free to change our assumptions

We may want to assume that the instructor also looks at the SAT score

But if that is the case we have to change the network to reflect this dependence
- Of course we are free to change our assumptions
- We may want to assume that the instructor also looks at the SAT score
- But if that is the case we have to change the network to reflect this dependence
- Why just SAT score? The instructor may even consult one of his colleagues and seek his/her opinion
Remember: The graph is a reflection of our assumptions about how the world works.
Remember: The graph is a reflection of our assumptions about how the world works

Our assumptions about dependencies are encoded in the graph

- **D**
- **I**
- **G**
- **S**
- **L**
Remember: The graph is a reflection of our assumptions about how the world works

Our assumptions about dependencies are encoded in the graph

Once we build the graph we freeze it and do all the reasoning and analysis (independence) on this graph
Remember: The graph is a reflection of our assumptions about how the world works

Our assumptions about dependencies are encoded in the graph

Once we build the graph we freeze it and do all the reasoning and analysis (independence) on this graph

It is not fair to ask “what if” questions involving other factors (For example, what if the professor was in a bad mood?)
If we believe Graph (a) is how the world works then \((L \perp S)|G\)

If we believe Graph (b) is how the world works then \((L \not\perp S)|G\)

We will stick to Graph (a) for the discussion.
If we believe Graph (a) is how the world works then $(L \perp S)|G$

If we believe Graph (b) is how the world works then $(L \not\perp S)|G$
If we believe Graph (a) is how the world works then $(L \perp S)|G$

If we believe Graph(b) is how the world works then $(L \not\perp S)|G$

We will stick to Graph(a) for the discussion
Let’s return back to our discussion of finding independence relations in the graph.
Let’s return back to our discussion of finding independence relations in the graph.

So far we have seen three cases as summarized in the next module.
Module 17.8: Independencies encoded by a Bayesian network (Case 3: Node and its descendants)
\[(G \not\perp D) \ (G \not\perp I) \ (S \not\perp I) \ (L \not\perp G)\]

A node is not independent of its parents.
A node is not independent of its parents

\[(G \not\perp D) \ (G \not\perp I) \ (S \not\perp I) \ (L \not\perp G)\]

A node is not independent of its parents even when other variables are given

\[(G \not\perp D, I)|S, L\]
\[(S \not\perp I)|D, G, L\]
\[(L \not\perp G)|D, I, S\]
A node is not independent of its parents

\[(G \not\perp D) \ (G \not\perp I) \ (S \not\perp I) \ (L \not\perp G)\]

A node is not independent of its parents even when other variables are given

\[(G \not\perp D, I) \mid S, L\]
\[(S \not\perp I) \mid D, G, L\]
\[(L \not\perp G) \mid D, I, S\]

A node seems to be independent of other variables given its parents

\[(S \perp G) \mid I?\]
\[(L \perp D, I, S) \mid G?\]
\[(G \perp L) \mid D, I?\]
Let us inspect this last rule

\[ \text{If you know that } d = 0 \text{ and } i = 1 \text{ then you would expect the student to get a good grade. But now if someone tells you that the student got a poor letter, your belief will change. So } (G \not\perp L) | D, I \]

The effect (letter) actually gives us information about the cause (grade).
Let us inspect this last rule

Is \((G \perp L) | D, I?\)
Let us inspect this last rule

Is \((G \perp L)\mid D, I?\)

If you know that \(d = 0\) and \(i = 1\) then you would expect the student to get a good grade

But now if someone tells you that the student got a poor letter, your belief will change

So \((G \not\perp L)\mid D, I\)

The effect (letter) actually gives us information about the cause (grade)
Let us inspect this last rule

Is \((G \perp L)|D, I?\)

If you know that \(d = 0\) and \(i = 1\) then you would expect the student to get a good grade

But now if someone tells you that the student got a poor letter, your belief will change
Let us inspect this last rule

Is \((G \perp L)\mid D, I?\)

If you know that \(d = 0\) and \(i = 1\) then you would expect the student to get a good grade

But now if someone tells you that the student got a poor letter, your belief will change

So \((G \not\perp L)\mid D, I\)
Let us inspect this last rule

Is \((G \perp L)|D, I?\)

If you know that \(d = 0\) and \(i = 1\) then you would expect the student to get a good grade.

But now if someone tells you that the student got a poor letter, your belief will change.

So \((G \not\perp L)|D, I\)

The effect (letter) actually gives us information about the cause (grade).
\[ (G \not\perp D) \ (G \not\perp I) \ (S \not\perp I) \ (L \not\perp G) \]

A node is not independent of its parents even when other variables are given.

Given its parents, a node is independent of all variables except its descendants.
A node is not independent of its parents

\[(G \not\perp D) \ (G \not\perp I) \ (S \not\perp I) \ (L \not\perp G)\]

A node is not independent of its parents even when other variables are given

\[(G \not\perp D, I)|S, L\]
\[(S \not\perp I)|D, G, L\]
\[(L \not\perp G)|D, I, S\]
A node is not independent of its parents even when other variables are given.

- \((G \perp D)\) \((G \perp I)\) \((S \perp I)\) \((L \perp G)\)
  - A node is not independent of its parents

- \((G \perp D, I)\mid S, L\)
- \((S \perp I)\mid D, G, L\)
- \((L \perp G)\mid D, I, S\)
  - A node is not independent of its parents even when other variables are given

- \((S \perp G)\mid I\)
- \((L \perp D, I, S)\mid G\)
- \((G \perp L)\mid D, I\)
  - Given its parents, a node is independent of all variables except its descendants
Module 17.9: Bayesian Networks: Formal Semantics
We are now ready to formally define the semantics of a Bayesian Network

**Bayesian Network Semantics:**

A Bayesian Network structure $G$ is a directed acyclic graph where nodes represent random variables $X_1, X_2, \ldots, X_n$. Let $P_G(i)$ denote the parents of $X_i$ in $G$ and $\text{NonDescendants}(X_i)$ denote the variables in the graph that are not descendants of $X_i$. Then $G$ encodes the following set of conditional independence assumptions called the local independencies and denoted by $I_i(G)$ for each variable $X_i$.

$$X_i \perp \text{NonDescendants}(X_i) | P_G(i)$$
We are now ready to formally define the semantics of a Bayesian Network

**Bayesian Network Semantics:**

A Bayesian Network structure G is a directed acyclic graph where nodes represent random variables $X_1, X_2, ..., X_n$. 

Let $P_{G}(X_i)$ denote the parents of $X_i$ in $G$ and $\text{NonDescendants}(X_i)$ denote the variables in the graph that are not descendants of $X_i$. Then $G$ encodes the following set of conditional independence assumptions called the local independencies and denoted by $I_i(G)$ for each variable $X_i$. 

$X_i \perp \text{NonDescendants}(X_i) | P_G(X_i)$
We are now ready to formally define the semantics of a Bayesian Network

**Bayesian Network Semantics:**
A Bayesian Network structure $G$ is a directed acyclic graph where nodes represent random variables $X_1, X_2, ..., X_n$. Let $P_{\text{a}_{X_i}}^G$ denote the parents of $X_i$ in $G$. Let $\text{NonDescendants}(X_i)$ denote the variables in the graph that are not descendants of $X_i$.
We are now ready to formally define the semantics of a Bayesian Network

**Bayesian Network Semantics:**

A Bayesian Network structure $G$ is a directed acyclic graph where nodes represent random variables $X_1, X_2, ..., X_n$. Let $P_{aX_i}^G$ denote the parents of $X_i$ in $G$ and $\text{NonDescendants}(X_i)$ denote the variables in the graph that are not descendants of $X_i$. 
We are now ready to formally define the semantics of a Bayesian Network

**Bayesian Network Semantics:**

A Bayesian Network structure $G$ is a directed acyclic graph where nodes represent random variables $X_1, X_2, \ldots, X_n$. Let $P^G_{a_{X_i}}$ denote the parents of $X_i$ in $G$ and $\text{NonDescendants}(X_i)$ denote the variables in the graph that are not descendants of $X_i$. Then $G$ encodes the following set of conditional independence assumptions called the local independencies and denoted by $I_i(G)$ for each variable $X_i$. 

\[
X_i \perp \text{NonDescendants}(X_i) | P^G_{a_{X_i}}
\]
We are now ready to formally define the semantics of a Bayesian Network

**Bayesian Network Semantics:**

A Bayesian Network structure $G$ is a directed acyclic graph where nodes represent random variables $X_1, X_2, \ldots, X_n$. Let $P_{a_{X_i}}^G$ denote the parents of $X_i$ in $G$ and NonDescendants($X_i$) denote the variables in the graph that are not descendants of $X_i$. Then $G$ encodes the following set of conditional independence assumptions called the local independencies and denoted by $I_i(G)$ for each variable $X_i$.

$$(X_i \perp \text{NonDescendants}(X_i) | P_{a_{X_i}}^G)$$
We will see some more formal definitions and then return to the question of independencies.
Module 17.10: I Maps
Let $P$ be a joint distribution over $X = X_1, X_2, \ldots, X_n$. We define $I(P)$ as the set of independence assumptions that hold in $P$.

For example, $I(P) = \{ (G \perp S | I, D), \ldots \}$.

Each element of this set is of the form $X_i \perp X_j | Z, Z \subseteq X | X_i, X_j$.

Let $I(G)$ be the set of independence assumptions associated with a graph $G$. 

![Diagram of nodes D, I, G, S, L with arrows connecting them]
Let $P$ be a joint distribution over $X = X_1, X_2, ..., X_n$. We define $I(P)$ as the set of independence assumptions that hold in $P$.

For example: $I(P) = \{ (G \perp S | I, D), ..., \}$

Each element of this set is of the form $X_i \perp X_j | Z, Z \subseteq X | X_i, X_j$.

Let $I(G)$ be the set of independence assumptions associated with a graph $G$.
Let $P$ be a joint distribution over $X = X_1, X_2, ..., X_n$

We define $I(P)$ as the set of independence assumptions that hold in $P$.

For Example:
$I(P) = \{(G \perp S|I, D), \ldots\}$
Let $P$ be a joint distribution over $X = X_1, X_2, ..., X_n$.

We define $I(P)$ as the set of independence assumptions that hold in $P$.

For Example:

$I(P) = \{(G \perp S| I, D), ...., \}$

Each element of this set is of the form $X_i \perp X_j|Z, Z \subseteq X|X_i, X_j$.
Let $P$ be a joint distribution over $X = X_1, X_2, \ldots, X_n$.

We define $I(P)$ as the set of independence assumptions that hold in $P$.

For Example:

$I(P) = \{ (G \perp S|I, D), \ldots \}$

Each element of this set is of the form $X_i \perp X_j|Z, Z \subseteq X|X_i, X_j$

Let $I(G)$ be the set of independence assumptions associated with a graph $G$. 

\begin{itemize}
  \item Let $P$ be a joint distribution over $X = X_1, X_2, \ldots, X_n$.
  \item We define $I(P)$ as the set of independence assumptions that hold in $P$.
  \item For Example:
    \[ I(P) = \{ (G \perp S|I, D), \ldots \} \]
  \item Each element of this set is of the form $X_i \perp X_j|Z, Z \subseteq X|X_i, X_j$.
  \item Let $I(G)$ be the set of independence assumptions associated with a graph $G$.
\end{itemize}
We say that $G$ is an I-map for $P$ if $I(G) \subseteq I(P)$.
We say that $G$ is an I-map for $P$ if $I(G) \subseteq I(P)$

$G$ does not mislead us about independencies in $P$
We say that $G$ is an I-map for $P$ if $I(G) \subseteq I(P)$

$G$ does not mislead us about independencies in $P$

Any independence that $G$ states must hold in $P$
We say that \( G \) is an I-map for \( P \) if
\[
I(G) \subseteq I(P)
\]

\( G \) does not mislead us about independencies in \( P \).

Any independence that \( G \) states must hold in \( P \).

But \( P \) can have additional independencies.
Consider this joint distribution over $X, Y$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$P(X,Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Consider this joint distribution over $X, Y$

We need to find a $G$ which is an I-map for this $P$
Consider this joint distribution over $X, Y$

- We need to find a $G$ which is an I-map for this $P$
- How do we find such a $G$?

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
<th>$P(X,Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Well since there are only 2 variables here the only possibilities are
$I(P) = \{(X \perp Y)\}$ or $I(P) = \emptyset$

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Well since there are only 2 variables here the only possibilities are 
$I(P) = \{(X \perp Y)\}$ or $I(P) = \emptyset$

From the table we can easily check $P(X, Y) = P(X).P(Y)$
Well since there are only 2 variables here the only possibilities are

$I(P) = \{(X \perp Y)\}$ or $I(P) = \emptyset$

From the table we can easily check

$P(X, Y) = P(X).P(Y)$

$I(P) = \{(X \perp Y)\}$
Well since there are only 2 variables here the only possibilities are $I(P) = \{(X \perp Y)\}$ or $I(P) = \emptyset$.

From the table we can easily check $P(X, Y) = P(X).P(Y)$.

$I(P) = \{(X \perp Y)\}$

Now can you come up with a $G$ which satisfies $I(G) \subseteq I(P)$?
Since we have only two variables, there are only 3 possibilities for $G$.

$I(G) = \Phi \quad I(G_2) = \Phi \quad I(G_3) = \{(X \perp Y)\}$
Since we have only two variables there are only 3 possibilities for $G$

Which of these is an I-Map for $P$?

$I(G) = \Phi \quad I(G_2) = \Phi \quad I(G_3) = \{(X \perp Y)\}$
Since we have only two variables there are only 3 possibilities for $G$

Which of these is an I-Map for $P$?

Well all three are I-Maps for $P$

$I(G) = \emptyset$ $I(G_2) = \emptyset$ $I(G_3) = \{(X \perp Y)\}$
Since we have only two variables there are only 3 possibilities for $G$

Which of these is an I-Map for $P$?

Well all three are I-Maps for $P$

They all satisfy the condition $I(G) \subseteq I(P)$

$I(G) = \Phi \quad I(G_2) = \Phi \quad I(G_3) = \{(X \perp Y)\}$
Of course, this was just a toy example
<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>P(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.32</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.12</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.48</td>
</tr>
</tbody>
</table>

- Of course, this was just a toy example
- In practice, we do not know $P$ and hence can’t compute $I(P)$
Of course, this was just a toy example.

In practice, we do not know $P$ and hence can’t compute $I(P)$.

We just make some assumptions about $I(P)$ and then construct a $G$ such that $I(G) \subseteq I(P)$.
So why do we care about I-Map?

If $G$ is an I-Map for a joint distribution $P$ then $P$ factorizes over $G$.

What does that mean?

Well, it just means that $P$ can be written as a product of factors where each factor is a c.p.d associated with the nodes of $G$. 

---

Mitesh M. Khapra
CS7015 (Deep Learning) : Lecture 17
So why do we care about I-Map?

If $G$ is an I-Map for a joint distribution $P$ then $P$ factorizes over $G$. 

Well, it just means that $P$ can be written as a product of factors where each factor is a c.p.d associated with the nodes of $G$. 

Diagram: 

- $D$
- $I$
- $G$
- $S$
- $L$
So why do we care about I-Map?

If $G$ is an I-Map for a joint distribution $P$ then $P$ factorizes over $G$.

What does that mean?
So why do we care about I-Map?

If $G$ is an I-Map for a joint distribution $P$ then $P$ factorizes over $G$

What does that mean?

Well, it just means that $P$ can be written as a product of factors where each factor is a c.p.d associated with the nodes of $G$
Theorem

Let G be a BN structure over a set of random variables X and let P be a joint distribution over these variables. If G is an I-Map for P, then P factorizes according to G

Proof: Exercise

Theorem

Let G be a BN structure over a set of random variables X and let P be a joint distribution over these variables. If P factorizes according to G, then G is an I-Map of P

Proof: Exercise
• Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$
• Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$

• There are many joint distributions possible
• Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$

• There are many joint distributions possible

• Each may entail different independence relations
Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$

There are many joint distributions possible

Each may entail different independence relations

For example, in some cases $L$ could be independent of $S$; in some not.
• Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$

• There are many joint distributions possible

• Each may entail different independence relations

• For example, in some cases $L$ could be independent of $S$; in some not.

• Can you think of a $G$ which will be an I-Map for any distribution over $P$?
Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$.

There are many joint distributions possible.

Each may entail different independence relations.

For example, in some cases $L$ could be independent of $S$; in some not.

Can you think of a $G$ which will be an I-Map for any distribution over $P$?

What is this graph called?
Consider a set of random variables $X_1, X_2, X_3, X_4, X_5$

There are many joint distributions possible

Each may entail different independence relations

For example, in some cases $L$ could be independent of $S$; in some not.

Can you think of a $G$ which will be an I-Map for any distribution over $P$?

- Answer: A complete graph
Answer: A complete graph

The factorization entailed by the above graph is

\[ P(X_3)P(X_5|X_3)P(X_1|X_3, X_5) \]
\[ P(X_2|X_1, X_3, X_5)P(X_4|X_1, X_2, X_3, X_5) \]

- Consider a set of random variables \( X_1, X_2, X_3, X_4, X_5 \)
- There are many joint distributions possible
- Each may entail different independence relations
- For example, in some cases \( L \) could be independent of \( S \); in some not.
- Can you think of a \( G \) which will be an I-Map for any distribution over \( P \)?
Answer: A complete graph
The factorization entailed by the above graph is
\[ P(X_3)P(X_5|X_3)P(X_1|X_3, X_5) \]
\[ P(X_2|X_1, X_3, X_5)P(X_4|X_1, X_2, X_3, X_5) \]
which is just chain rule of probability which holds for any distribution.

Consider a set of random variables \( X_1, X_2, X_3, X_4, X_5 \)

There are many joint distributions possible

Each may entail different independence relations

For example, in some cases \( L \) could be independent of \( S \); in some not.

Can you think of a \( G \) which will be an I-Map for any distribution over \( P \)?