CS7015 (Deep Learning) : Lecture 17

Recap of Probability Theory, Bayesian Networks, Conditional Independence in Bayesian Networks

Mitesh M. Khapra

Department of Computer Science and Engineering
Indian Institute of Technology Madras

Module 17.0: Recap of Probability Theory

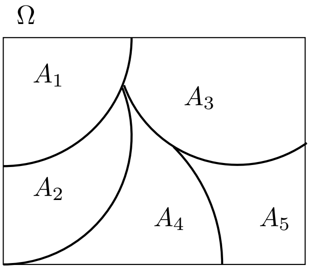
We will start with a quick recap of some basic concepts from probability

Axioms of Probability

Axioms of Probability

- For any event A,

$$P(A) \geq 0$$



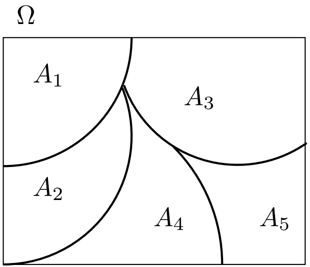
Axioms of Probability

- For any event A,

$$P(A) \geq 0$$

- If $A_1, A_2, A_3, ..., A_n$ are disjoint events (i.e., $A_i \cap A_j = \phi \quad \forall i \neq j$) then

$$P(\cup A_i) = \sum_i P(A_i)$$



Axioms of Probability

- For any event A,

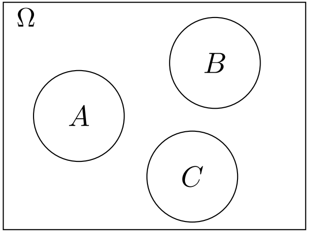
$$P(A) \geq 0$$

- If $A_1, A_2, A_3, ..., A_n$ are disjoint events (i.e., $A_i \cap A_j = \phi \quad \forall i \neq j$) then

$$P(\cup A_i) = \sum_i P(A_i)$$

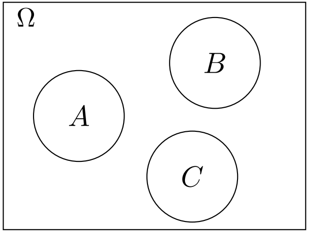
- If Ω is the universal set containing all events then

$$P(\Omega) = 1$$



Random Variable (intuition)

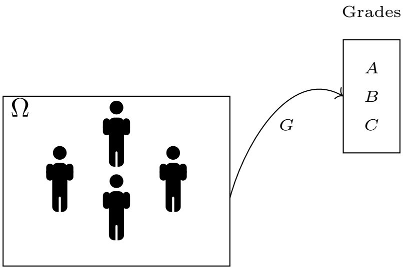
- Suppose a student can get one of 3 possible grades in a course: A, B, C



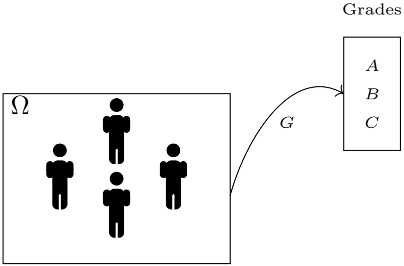
Random Variable (intuition)

- 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

Random Variable (intuition)

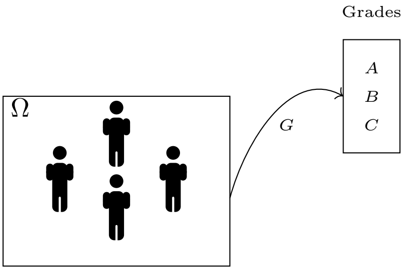


- Suppose a student can get one of 3 possible grades in a course: A, B, C
- One way of interpreting this is that there are 3 possible events here
- Another way of looking at this is there is a *random variable* G which each student to one of the 3 possible values



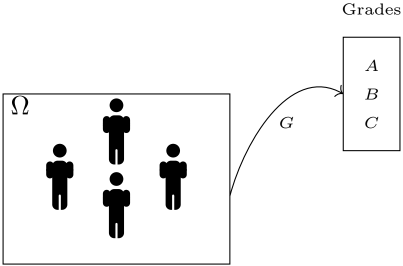
Random Variable (intuition)

- 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 $g \in \{A, B, C\}$



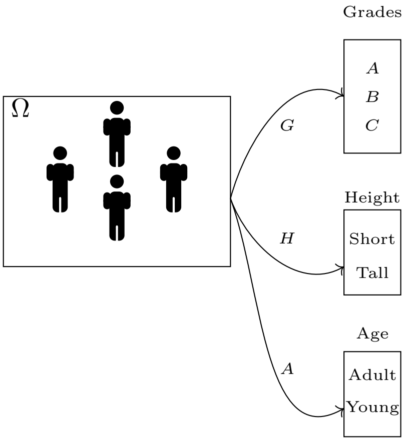
Random Variable (intuition)

- 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



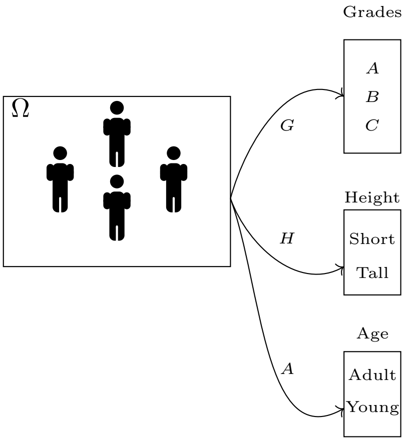
Random Variable (intuition)

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



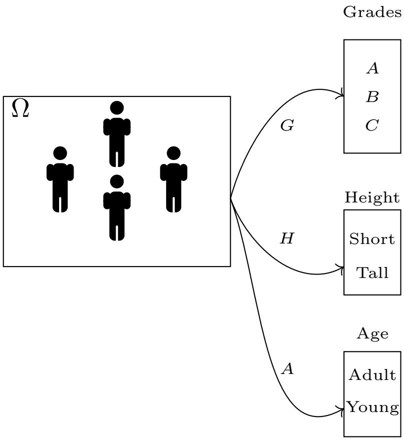
Random Variable (intuition)

- 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)

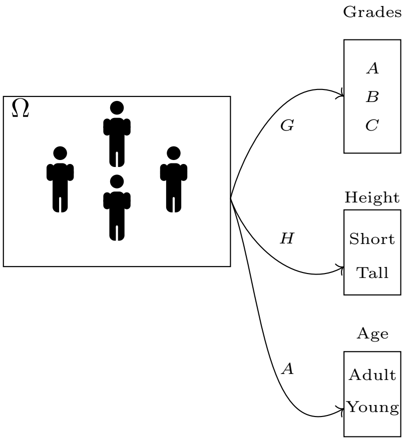
- But the second one (using random variables) is more compact
- Specially, when there are multiple attributes associated with a student (outcome) - *grade, height, age, etc.*
- We could have one random variable corresponding to each attribute

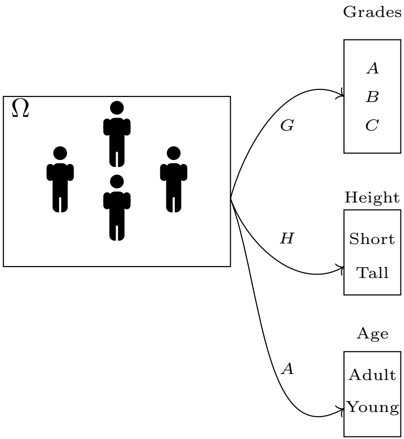


Random Variable (intuition)

- 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 $Grade = g$, $Height = h$, $Age = a$ and so on

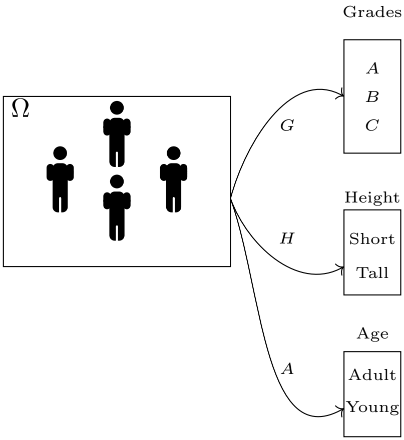
Random Variable (formal)





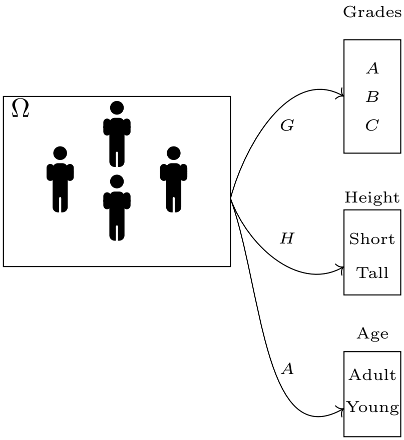
Random Variable (formal)

- A random variable is a ***function*** which maps each outcome in Ω to a value



Random Variable (formal)

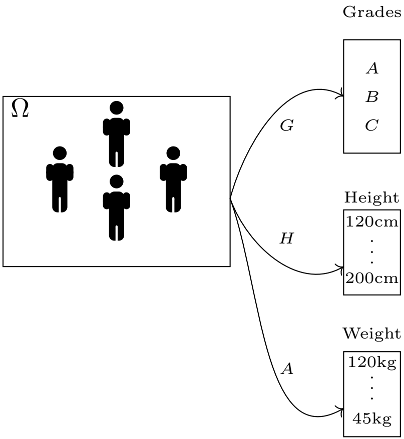
- A random variable is a ***function*** which maps each outcome in Ω to a value
- In the previous example, G (or f_{grade}) maps each student in Ω to a value: A, B or C



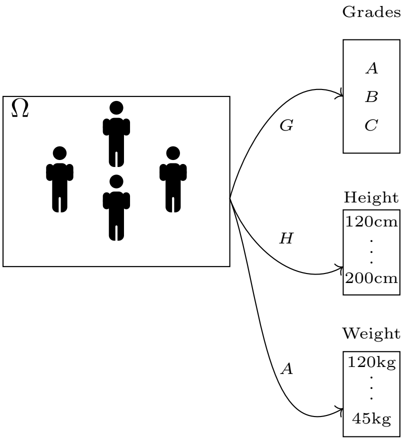
Random Variable (formal)

- A random variable is a ***function*** which maps each outcome in Ω to a value
- In the previous example, G (or f_{grade}) maps each student in Ω 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)

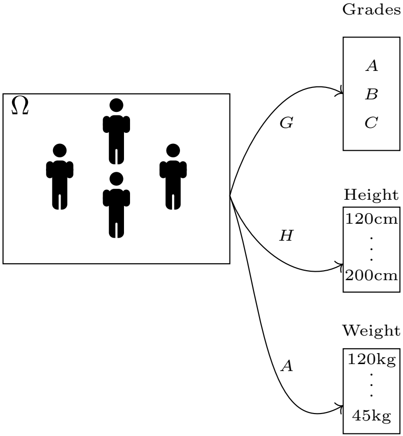


- A random variable can either take continuous values (for example, $weight, height$)



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 ?

Marginal Distribution

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

G	$P(G = g)$
A	0.1
B	0.2
C	0.7

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

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

G	$P(G = g)$
A	0.1
B	0.2
C	0.7

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 $P(G)$

Joint Distribution

- Consider two random variable G (grade) and I (intellegence $\in \{\mathbf{H}\text{igh}, \mathbf{L}\text{ow}\}$)

Joint Distribution

- Consider two random variable G (grade) and I (intellegence $\in$ {**H**igh, **L**ow})
- The joint distribution over these two random variables assigns probabilities to all events involving these two random variables

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

G	I	$P(G = g, I = i)$
A	High	0.3
A	Low	0.1
B	High	0.15
B	Low	0.15
C	High	0.1
C	Low	0.2

Joint Distribution

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

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

G	I	$P(G = g, I = i)$
A	High	0.3
A	Low	0.1
B	High	0.15
B	Low	0.15
C	High	0.1
C	Low	0.2

Joint Distribution

- Consider two random variable G (grade) and I (intellegence $\in$ {**H**igh, **L**ow})
- 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 $P(G, I)$

Conditional Distribution

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

- Consider two random variable G (grade) and I (intel-legence)

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

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

Conditional Distribution

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

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

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

- More compactly defined as

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

$$or \quad \underbrace{P(G, I)}_{joint} = \underbrace{P(G|I)}_{conditional} * \underbrace{P(I)}_{marginal}$$

Joint Distribution (n random variables)

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

$$\sum = 1$$

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

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

$$\sum = 1$$

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

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

$$\sum = 1$$

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)

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

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

  $$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Joint Distribution (n random variables)

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

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

$$P(X_1, X_2, ..., X_n)$$

Joint Distribution (n random variables)

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

- 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

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

X_1	$\dots$	X_n	$P(X_1, X_2, \dots, X_n)$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$
$\dots$	$\dots$	$\dots$	$\dots$

Joint Distribution (n random variables)

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

  $$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

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

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Joint Distribution (n random variables)

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

  $$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

  $$P(X_1, X_2, ..., X_n)$$
  $$= P(X_2, ..., X_n|X_1)P(X_1)$$
  $$= P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1)$$
  $$= P(X_4, ..., X_n|X_1, X_2, X_3)P(X_3|X_2, X_1)$$
  $$P(X_2|X_1)P(X_1)$$

X_1	$\ldots$	X_n	$P(X_1, X_2, \ldots, X_n)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Joint Distribution (n random variables)

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

$$P(X_1, X_2) = P(X_2|X_1)P(X_1) = P(X_1|X_2)P(X_2)$$

- Similarly for n random variables

$$\begin{aligned}
&P(X_1, X_2, ..., X_n) \\
&= P(X_2, ..., X_n|X_1)P(X_1) \\
&= P(X_3, ..., X_n|X_1, X_2)P(X_2|X_1)P(X_1) \\
&= P(X_4, ..., X_n|X_1, X_2, X_3)P(X_3|X_2, X_1) \\
&\quad P(X_2|X_1)P(X_1) \\
&= P(X_1)\prod_{i=2}^{n} P(X_i|X_1^{i-1}) \quad \textbf{\textit{(chain rule)}}
\end{aligned}$$

From Joint Distributions to Marginal Distributions

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A = a)$
High	0.55
Low	0.45

B	$P(B = a)$
High	0.65
Low	0.35

- Suppose we are given a joint distribtion over two random variables A, B

From Joint Distributions to Marginal Distributions

A	B	$P(A=a,B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

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

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

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

A	B	$P(A=a,B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

From Joint Distributions to Marginal Distributions

- Suppose we are given a joint distribtion 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

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

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

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A = a)$
High	0.55
Low	0.45

B	$P(B = a)$
High	0.65
Low	0.35

What if there are n random variables ?

- Suppose we are given a joint distribtion over n random variables X_1, X_2, ..., X_n

A	B	$P(A = a, B = b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A = a)$
High	0.55
Low	0.45

B	$P(B = a)$
High	0.65
Low	0.35

What if there are n random variables ?

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

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

What if there are n random variables ?

A	B	$P(A=a,B=b)$
High	High	0.3
High	Low	0.25
Low	High	0.35
Low	Low	0.1

A	$P(A=a)$
High	0.55
Low	0.45

B	$P(B=a)$
High	0.65
Low	0.35

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

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

- More compactly written as

$$P(X_1) = \sum_{X_2, X_3, ..., X_n} P(X_1, X_2, ..., X_n)$$

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 $\boldsymbol{G}rade$ to be dependent on $\boldsymbol{I}ntelligence$ but independent of $\boldsymbol{W}eight$

- Recall that by Chain Rule of Probability

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

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 $\boldsymbol{G}rade$ to be dependent on $\boldsymbol{I}ntelligence$ but independent of $\boldsymbol{W}eight$

Conditional Independence

- Recall that by Chain Rule of Probability

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

- However, if X and Y are independent, then

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

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

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

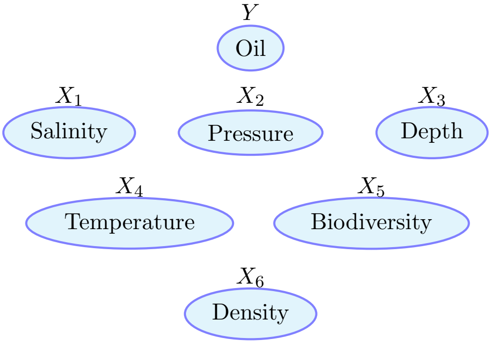
- We denote this as $X \perp\!\!\!\perp Y$
- In other words, knowing the value of Y does not change our belief about X
- We would expect $\boldsymbol{G}rade$ to be dependent on $\boldsymbol{I}ntelligence$ but independent of $\boldsymbol{W}eight$

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

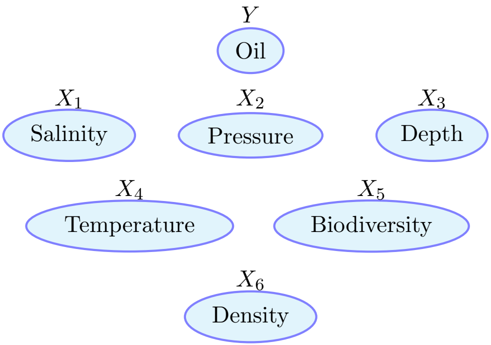
- In many real world applications, we have to deal with a large number of random variables

- 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



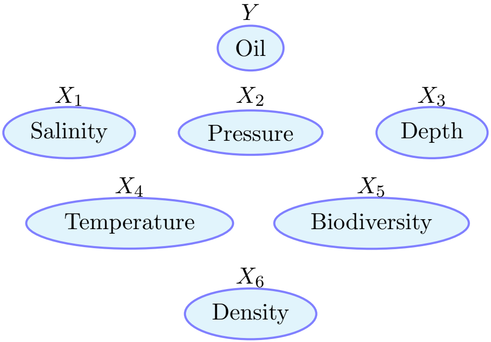
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- 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



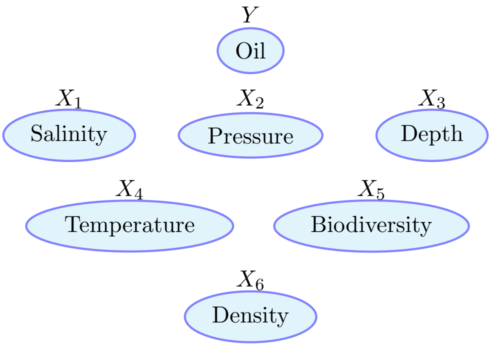
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- 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



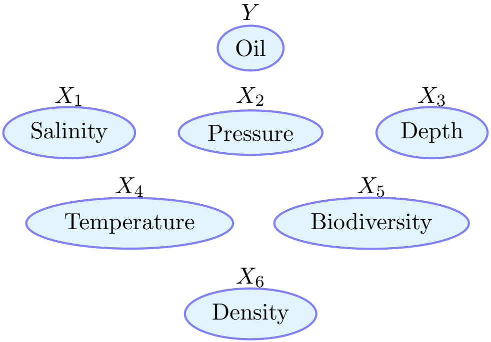
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- But why joint distribution?



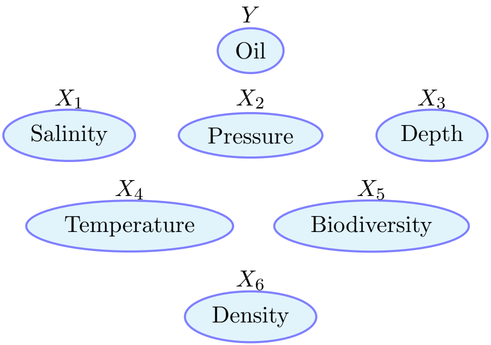
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- But why joint distribution?
- Aren't we just interested in $P(Y|X_1, X_2, ..., X_n)$?



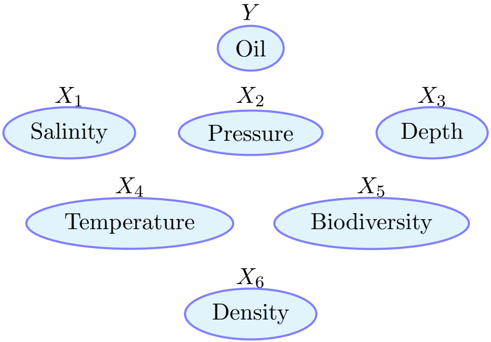
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- 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



$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

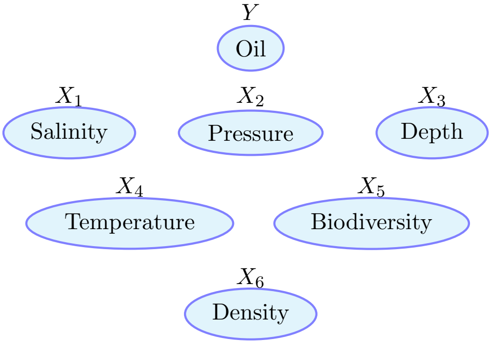
- But why joint distribution?
- Aren't we just interested in $P(Y|X_1, X_2, ..., X_n)$?
- Well, if we know the joint distribution, we can find answers to a bunch of interesting questions
- Let us see some such questions of interest



$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- We can find the conditional distribution

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



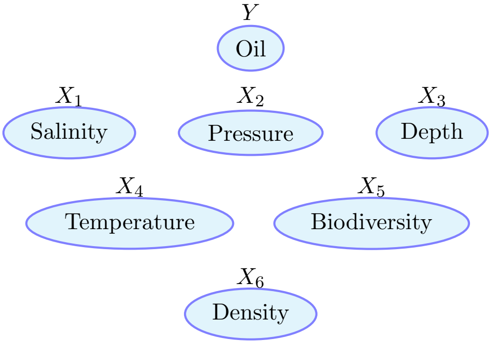
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- 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,

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



$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- We can find the conditional distribution

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

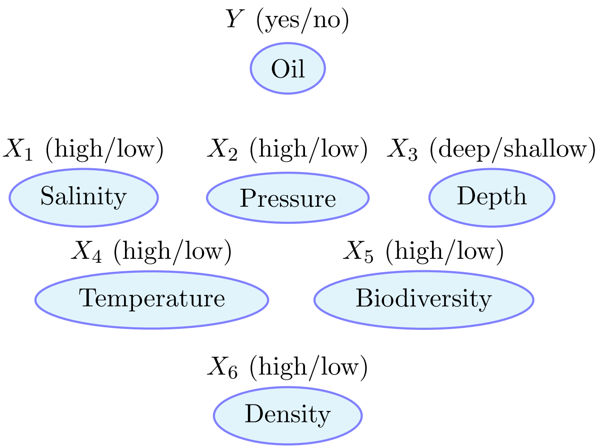
- We can find the marginal distribution,

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

- We can find the conditional independencies,

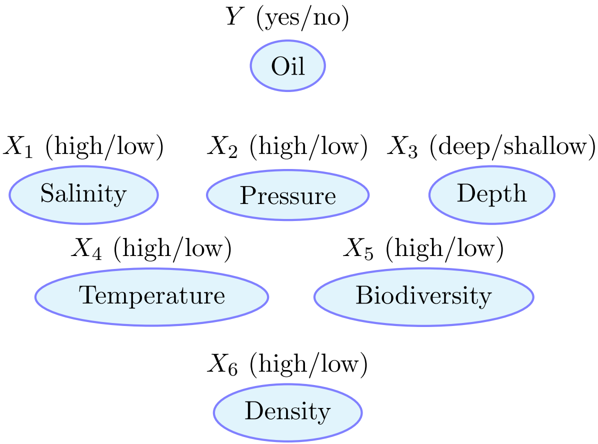
$$P(Y, X_1) = P(Y)P(X_1)$$

Module 17.2: How do we represent a joint distribution



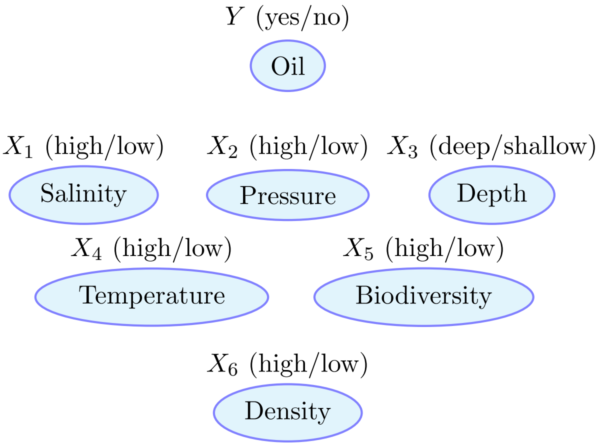
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- Let us return to the case of n random variables



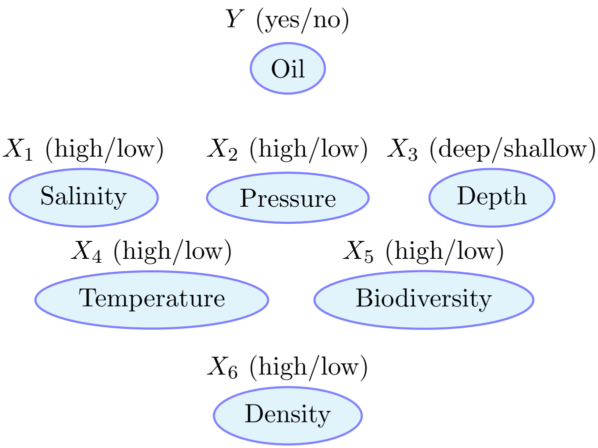
$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

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



$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

- Let us return to the case of n random variables
- For simplicity assume each of these variables can take binary values
- To specify the joint distribution, we need to specify $2^n - 1$ values. Why not (2^n)?



$$P(Y, X_1, X_2, X_3, X_4, X_5, X_6)$$

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

X_1	X_2	X_3	X_4	...	X_n	P
0	0	0	0	...	0	0.01
1	0	0	0	...	0	0.03
0	1	0	0	...	0	0.05
1	1	0	0	...	0	0.1
				...		
				...		
				...		
1	1	1	1	...	1	0.002

(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

X_1	X_2	X_3	X_4	...	X_n	P
0	0	0	0	...	0	0.01
1	0	0	0	...	0	0.03
0	1	0	0	...	0	0.05
1	1	0	0	...	0	0.1
				...		
				...		
				...		
1	1	1	1	...	1	0.002

(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 to store

X_1	X_2	X_3	X_4	...	X_n	P
0	0	0	0	...	0	0.01
1	0	0	0	...	0	0.03
0	1	0	0	...	0	0.05
1	1	0	0	...	0	0.1
				...		
				...		
				...		
1	1	1	1	...	1	0.002

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

X_1	X_2	X_3	X_4	...	X_n	P
0	0	0	0	...	0	0.01
1	0	0	0	...	0	0.03
0	1	0	0	...	0	0.05
1	1	0	0	...	0	0.1
				...		
				...		
				...		
1	1	1	1	...	1	0.002

(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 to store
- **Cognitive:** Impossible to acquire so many numbers from a human
- **Statistical:** Need huge amounts of data to learn the parameters

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)

I	S	$P(I, S)$
0	0	0.665
0	1	0.035
1	0	0.06
1	1	0.24

- 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

I	S	$P(I,S)$
0	0	0.665
0	1	0.035
1	0	0.06
1	1	0.24

- 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

I	S	$P(I,S)$
0	0	0.665
0	1	0.035
1	0	0.06
1	1	0.24

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

- 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

I	S	$P(I,S)$
0	0	0.665
0	1	0.035
1	0	0.06
1	1	0.24

- 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)

- 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

- 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

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

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

- 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 $P(S|I)$

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

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 $P(S|I)$

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

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)
- The number of parameters is still 3

- 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 $P(S|I)$

	i=0	i=1
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

- What have we achieved so far?

	i=0	i=1
$P(I)$	0.7	0.3

no.of parameters=1

	s=0	s=1
$P(S\|I=0)$	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

	i=0	i=1
$P(I)$	0.7	0.3

no.of parameters=1

	s=0	s=1
$P(S\|I=0)$	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
- But, we have a more natural way of representing the distribution

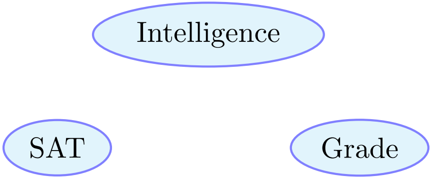
	i=0	i=1
$P(I)$	0.7	0.3

no.of parameters=1

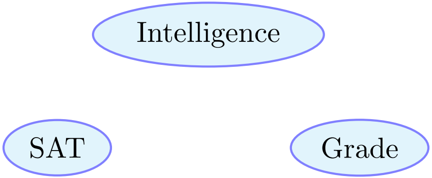
	s=0	s=1
$P(S\|I=0)$	0.95	0.05
$P(S\|I=1)$	0.2	0.8

no.of parameters=2

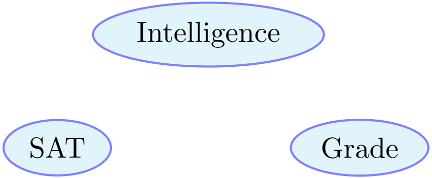
- What have we achieved so far?
- We were not able to reduce the number of parameters
- But, we have a more natural way of representing the distribution
- This is known as conditional parameterization



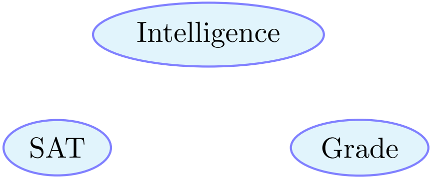
- Now consider a third random variable Grade (G)



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

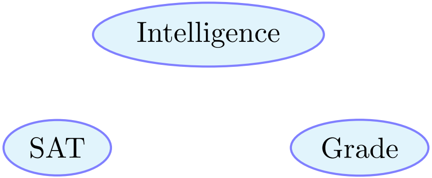


- Now consider a third random variable Grade (G)
- Notice that none of these 3 variables are independent of each other
- Grade and SAT Score are clearly correlated with Intelligence

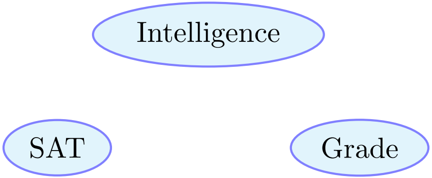


- Now consider a third random variable Grade (G)
- Notice that none of these 3 variables are independent of each other
- Grade and SAT Score are clearly correlated with Intelligence
- Grade and SAT Score are also correlated because we would expect

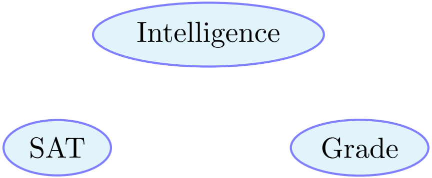
$$P(G = 1|S = 1) > P(G = 1|S = 0)$$



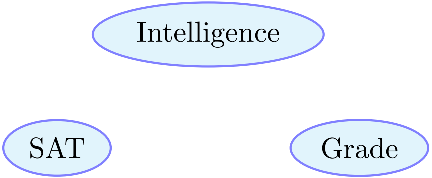
- However, it is possible that the distribution satisfies a conditional independence



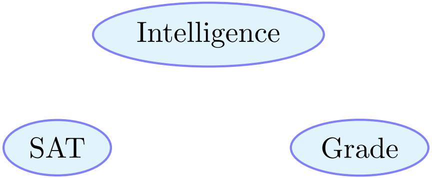
- However, it is possible that the distribution satisfies a conditional independence
- If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining G



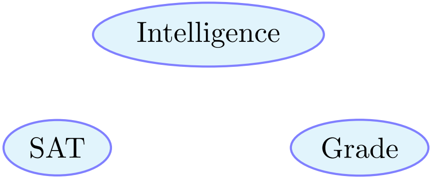
- However, it is possible that the distribution satisfies a conditional independence
- If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining G
- In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score



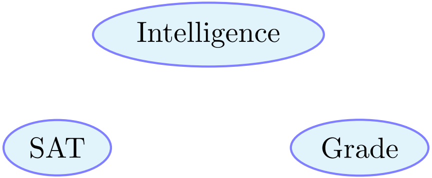
- However, it is possible that the distribution satisfies a conditional independence
- If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining G
- In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score
- Formally, we assume that $(S \perp G|I)$



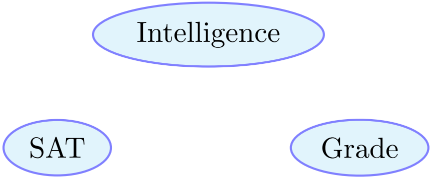
- However, it is possible that the distribution satisfies a conditional independence
- If we know that $I = H$, then it is possible that $S = H$ does not give any extra information for determining G
- In other words, if we know that the student is intelligent we can make inferences about his grade without even knowing the SAT score
- Formally, we assume that $(S \perp G | I)$
- Note that this is just an assumption



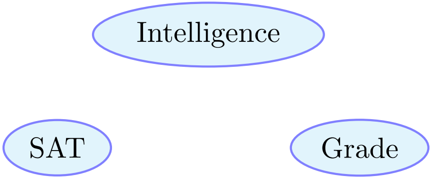
- We could argue that in many cases $S \not\perp G | I$



- 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 $P(I, G, S)$?

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

- 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?

- 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 $P(I, G, S)$?

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

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

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \end{aligned}$$

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i=0$	$i=1$
$P(I)$	0.7	0.3

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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i=0$	$i=1$
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

	g=A	g=B	g=C
P(G—I=0)	0.2	0.34	0.46
P(G—I=1)	0.74	0.17	0.09

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

	s=0	s=1
$P(S \mid I = 0)$	0.95	0.05
$P(S \mid I = 1)$	0.2	0.8

no.of parameters=2

	g=A	g=B	g=C
P(G—I=0)	0.2	0.34	0.46
P(G—I=1)	0.74	0.17	0.09

no.of parameters=4

- 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?

$$\begin{aligned} P(I, G, S) &= P(S, G \mid I)P(I) \\ &= P(S \mid G, I)P(G \mid I)P(I) \\ &= P(S \mid I)P(G \mid I)P(I) \end{aligned}$$

since $(S \perp G \mid I)$

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

	g=A	g=B	g=C
P(G—I=0)	0.2	0.34	0.46
P(G—I=1)	0.74	0.17	0.09

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?

$$\begin{aligned} P(I, G, S) &= P(S, G|I)P(I) \\ &= P(S|G, I)P(G|I)P(I) \\ &= P(S|I)P(G|I)P(I) \end{aligned}$$

since $(S \perp G|I)$

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

	g=A	g=B	g=C
P(G—I=0)	0.2	0.34	0.46
P(G—I=1)	0.74	0.17	0.09

no.of parameters=4

total no.of parameters=7

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

	g=A	g=B	g=C
P(G—I=0)	0.2	0.34	0.46
P(G—I=1)	0.74	0.17	0.09

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

	$i = 0$	$i = 1$
$P(I)$	0.7	0.3

no.of parameters=1

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

no.of parameters=2

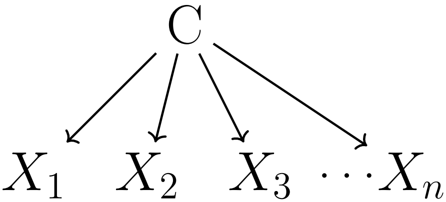
	g=A	g=B	g=C
P(G—I=0)	0.2	0.34	0.46
P(G—I=1)	0.74	0.17	0.09

no.of parameters=4

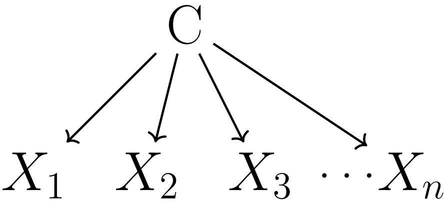
total no.of parameters=7

- 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?

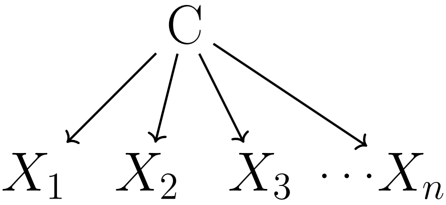


- Suppose we have n random variables, all of which are independent given another random variable C



- Suppose we have n random variables, all of which are independent given another random variable C
- The joint distribution factorizes as,

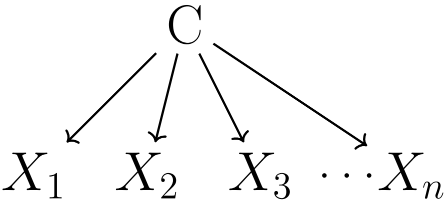
since $X_i \perp X_j | C$



- Suppose we have n random variables, all of which are independent given another random variable C
- The joint distribution factorizes as,

$$
\begin{aligned}
P(C, X_1, ..., X_n) = P(C)P(X_1|C) \\
P(X_2|X_1, C) \\
P(X_3|X_2, X_1, C)...
\end{aligned}
$$

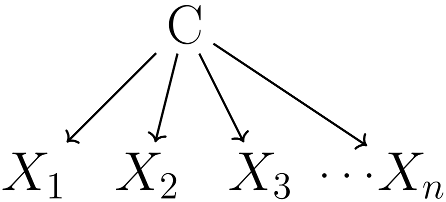
since $X_i \perp X_j | C$



- Suppose we have n random variables, all of which are independent given another random variable C
- The joint distribution factorizes as,

$$\begin{aligned} P(C, X_1, ..., X_n) = P(C)P(X_1|C) \\ P(X_2|X_1, C) \\ P(X_3|X_2, X_1, C)... \\ = P(C)\prod_{i=1}^{n} P(X_i|C) \end{aligned}$$

since $X_i \perp X_j | C$

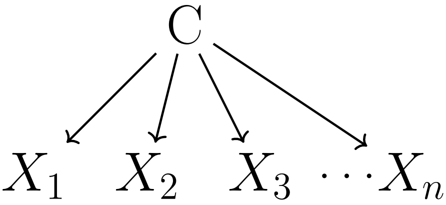


- This is called the Naive Bayes model

- Suppose we have n random variables, all of which are independent given another random variable C
- The joint distribution factorizes as,

$$\begin{aligned} P(C, X_1, ..., X_n) = P(C)P(X_1|C) \\ P(X_2|X_1, C) \\ P(X_3|X_2, X_1, C)... \\ = P(C)\prod_{i=1}^{n} P(X_i|C) \end{aligned}$$

since $X_i \perp X_j | C$



- This is called the Naive Bayes model
- It makes the Naive assumption that ${}^nC_2$ pairs are independent given C

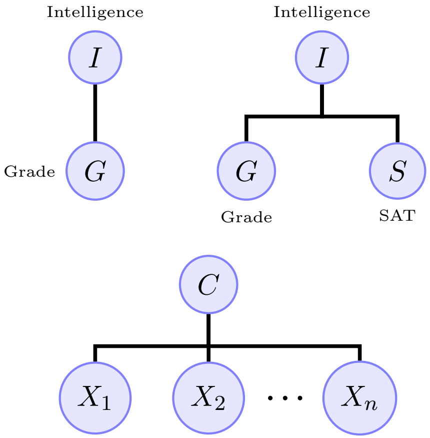
- Suppose we have n random variables, all of which are independent given another random variable C
- The joint distribution factorizes as,

$$
\begin{aligned}
P(C, X_1, ..., X_n) = P(C)P(X_1|C)\\
P(X_2|X_1, C)\\
P(X_3|X_2, X_1, C)...\\
= P(C)\prod_{i=1}^{n} P(X_i|C)
\end{aligned}
$$

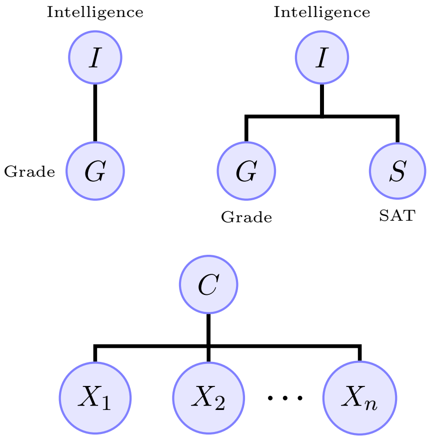
since $X_i \perp X_j | C$

- Bayesian networks build on the intuitions that we developed for the Naive Bayes model

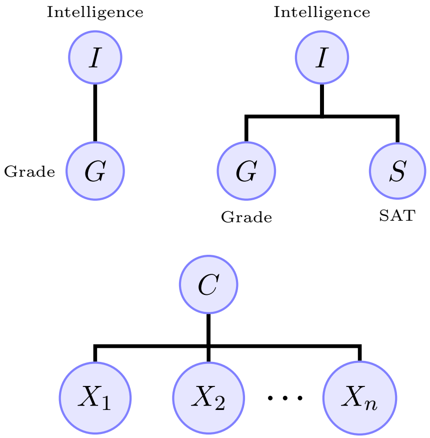
- Bayesian networks build on the intuitions that we developed for the Naive Bayes model
- But they are not restricted to strong (naive) independence assumptions



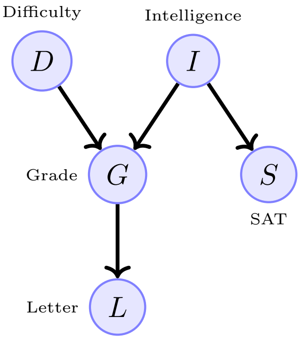
- Bayesian networks build on the intuitions that we developed for the Naive Bayes model
- But they are not restricted to strong (naive) independence assumptions
- We use graphs to represent the joint distribution



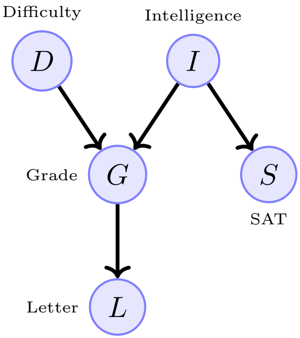
- Bayesian networks build on the intuitions that we developed for the Naive Bayes model
- But they are not restricted to strong (naive) independence assumptions
- We use graphs to represent the joint distribution
- **Nodes:** Random Variables



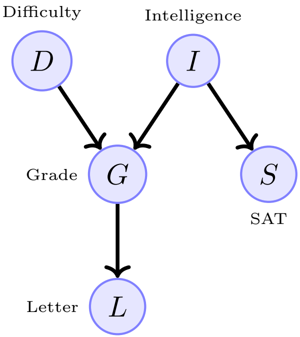
- Bayesian networks build on the intuitions that we developed for the Naive Bayes model
- But they are not restricted to strong (naive) independence assumptions
- We use graphs to represent the joint distribution
- **Nodes:** Random Variables
- **Edges:** Indicate dependence



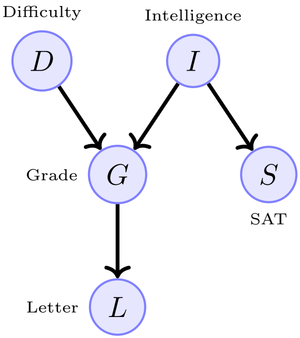
- Let's revisit the student example



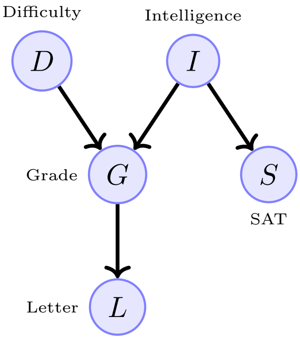
- Let's revisit the student example
- We will introduce a few more random variables and independence assumptions



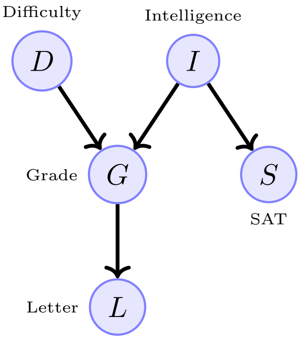
- Let's revisit the student example
- We will introduce a few more random variables and independence assumptions
- The grade now depends on student's Intelligence & exam's Difficulty level



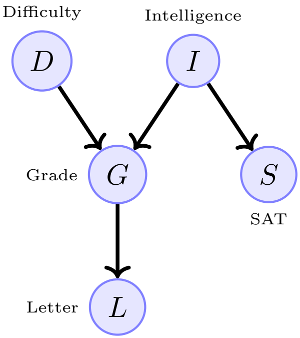
- Let's revisit the student example
- We will introduce a few more random variables and independence assumptions
- The grade now depends on student's Intelligence & exam's Difficulty level
- The SAT score depends on Intelligence



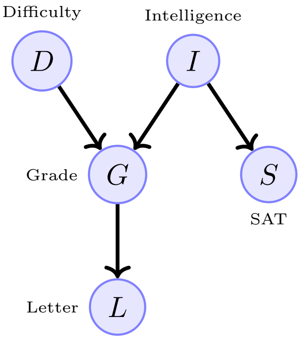
- Let's revisit the student example
- We will introduce a few more random variables and independence assumptions
- The grade now depends on student's Intelligence & exam's Difficulty level
- The SAT score depends on Intelligence
- The recommendation Letter from the course instructor depends on the Grade



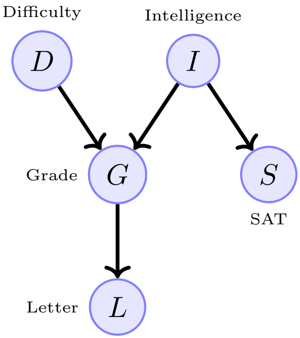
- The Bayesian network contains a node for each random variable



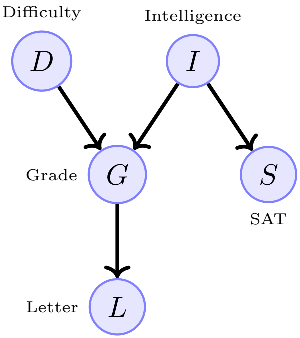
- The Bayesian network contains a node for each random variable
- The edges denote the dependencies between the random variables



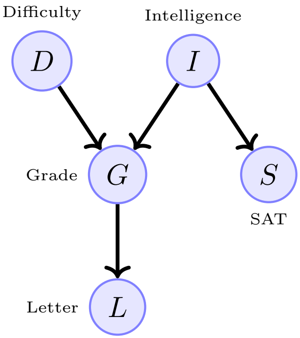
- The Bayesian network contains a node for each random variable
- The edges denote the dependencies between the random variables
- Each variable depends directly on its parents in the network



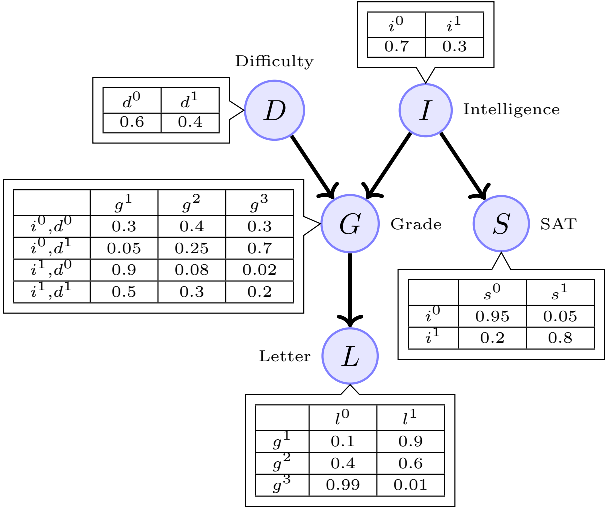
- The Bayesian network can be viewed as a data structure



- 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



i^0	i^1
0.7	0.3

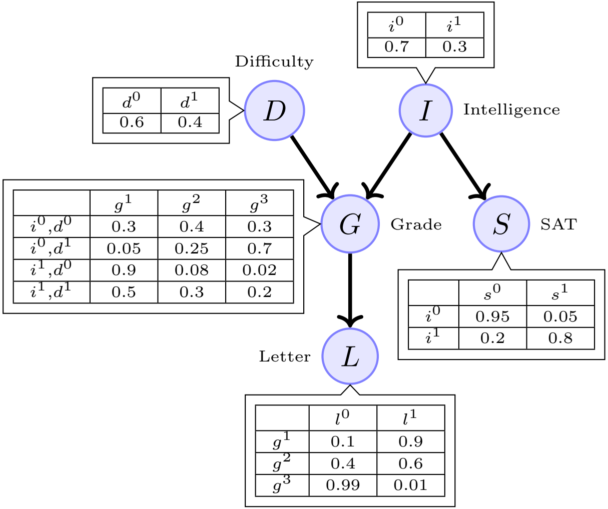
d^0	d^1
0.6	0.4

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- Each node is associated with a local probability model



d^0	d^1
0.6	0.4

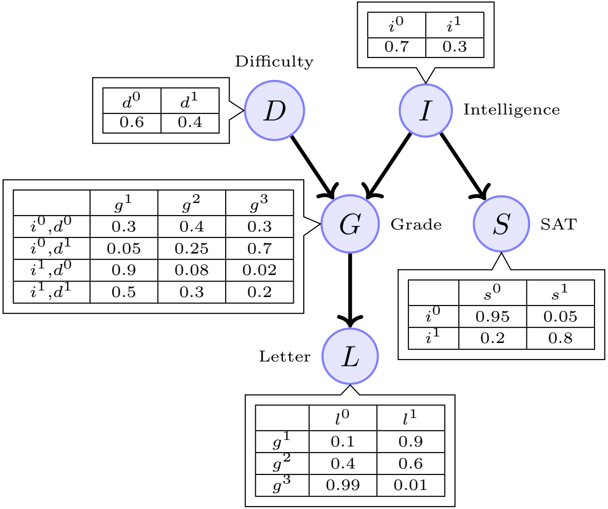
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

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



d^0	d^1
0.6	0.4

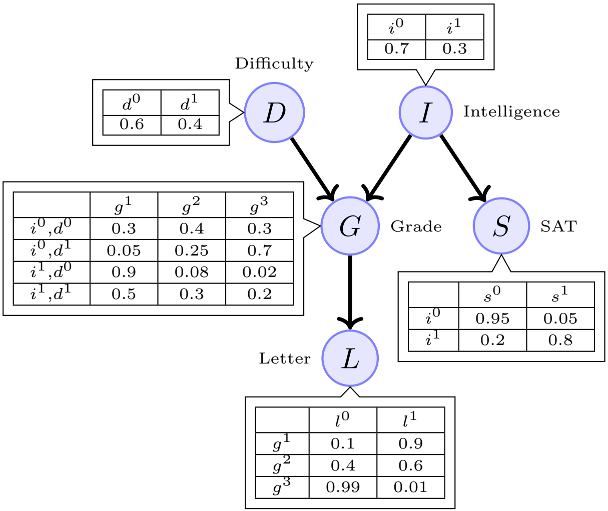
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

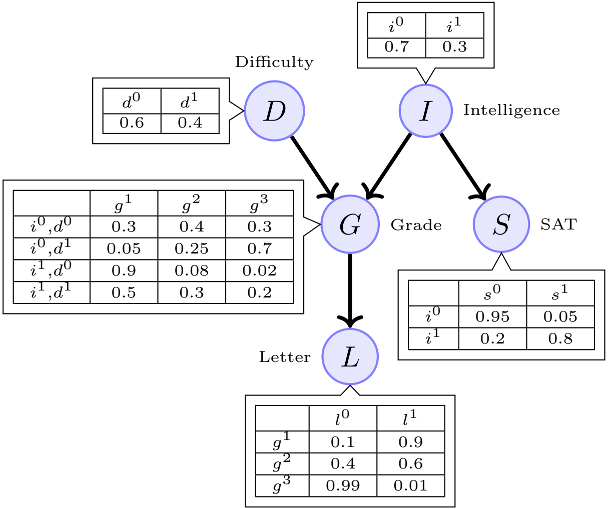
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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



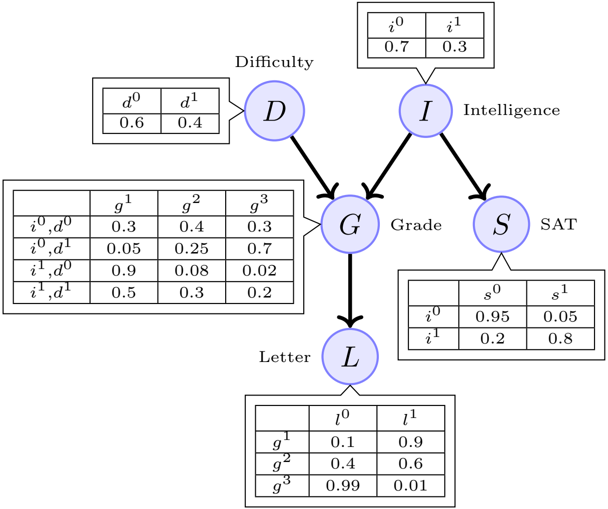
i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

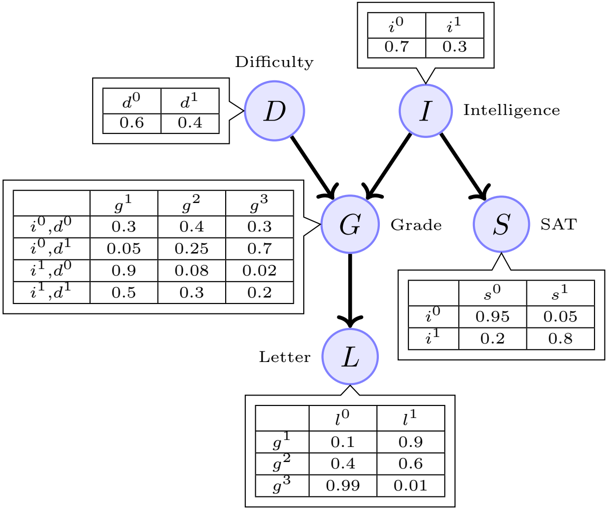
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- The graph gives us a natural factorization for the joint distribution
- In this case,

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



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

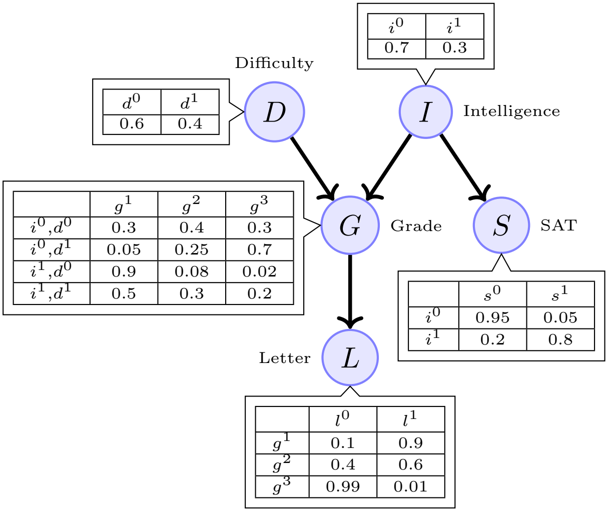
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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,

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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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

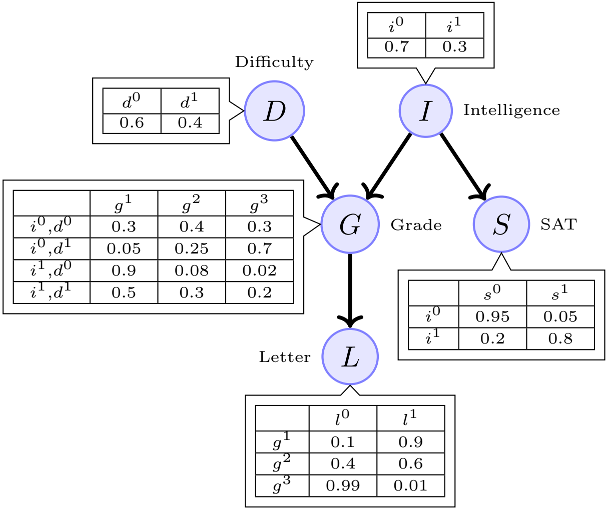
- We will denote $P(I = 0)$ by $P(i^0)$

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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

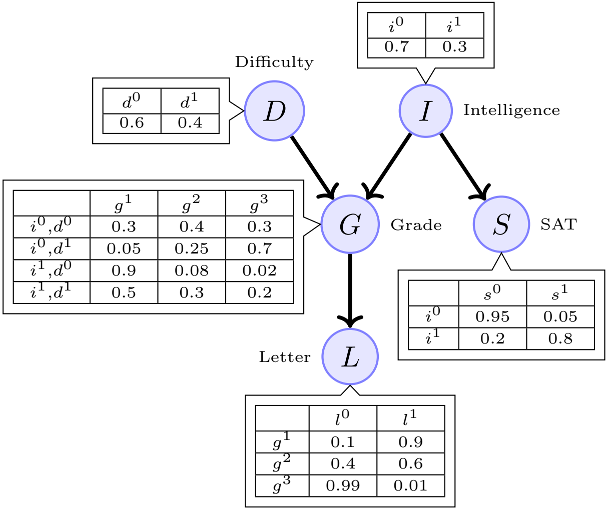
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Causal Reasoning

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



d^0	d^1
0.6	0.4

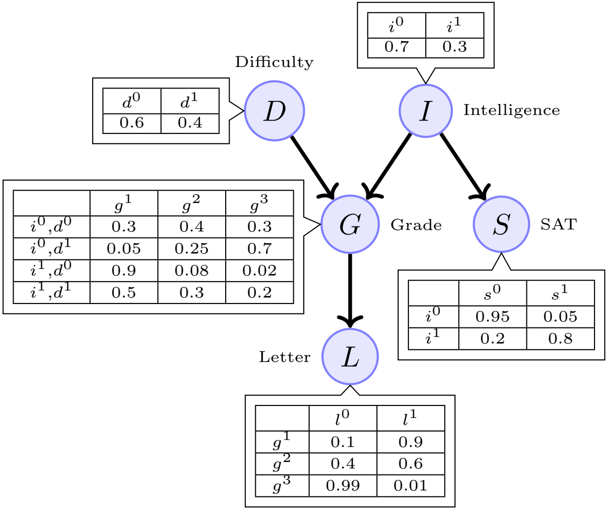
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Causal Reasoning



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

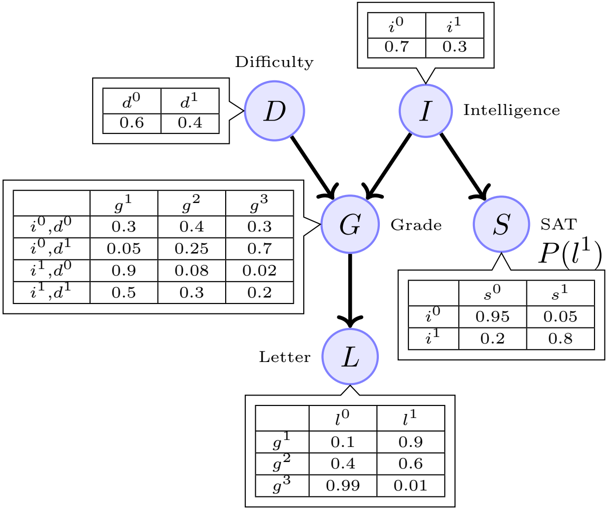
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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, $P(l^1)$?

Causal Reasoning



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

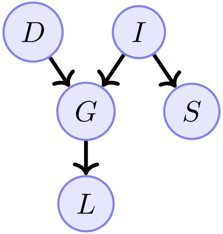
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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, $P(l^1)$?

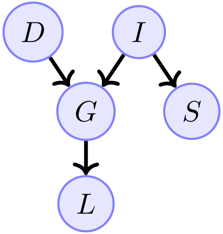
$$P(l^1) = \sum_{I\epsilon(0,1)} \sum_{D\epsilon(0,1)} \sum_{S\epsilon(0,1)} \sum_{G\epsilon(A,B,C)} P(I,D,G,S,l^1)$$

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



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

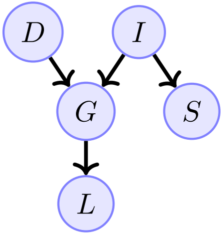
$$= \sum_{I\epsilon(0,1)} P(I) \sum_{D\epsilon(0,1)} P(D|I) \sum_{S\epsilon(0,1)} P(S|I, D) \sum_{G\epsilon(A,B,C)} P(G|I, D, S).P(l^1|G, I, D, S)$$



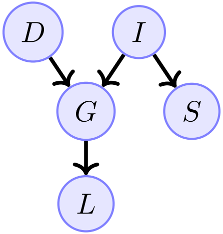
$$P(l^1) = \sum_{I\epsilon(0,1)} \sum_{D\epsilon(0,1)} \sum_{S\epsilon(0,1)} \sum_{G\epsilon(A,B,C)} P(I, D, G, S, l^1)$$

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

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



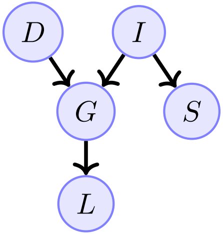
$$P(l^1) = \sum_{I\epsilon(0,1)} P(I) \sum_{D\epsilon(0,1)} P(D) \sum_{S\epsilon(0,1)} P(S|I) \sum_{G\epsilon(A,B,C)} P(G|I,D)P(l^1|G)$$



	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(l^1) = \sum_{I\epsilon(0,1)} P(I) \sum_{D\epsilon(0,1)} P(D) \sum_{S\epsilon(0,1)} P(S|I) \sum_{G\epsilon(A,B,C)} P(G|I,D)P(l^1|G)$$

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

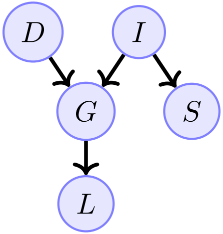


	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

$$P(l^1) = \sum_{I\epsilon(0,1)} P(I) \sum_{D\epsilon(0,1)} P(D) \sum_{S\epsilon(0,1)} P(S|I) \sum_{G\epsilon(A,B,C)} P(G|I,D)P(l^1|G)$$

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

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



	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

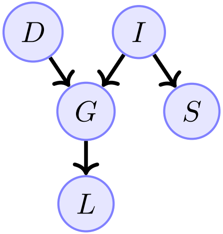
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

$$P(l^1) = \sum_{I\epsilon(0,1)} P(I) \sum_{D\epsilon(0,1)} P(D) \sum_{S\epsilon(0,1)} P(S|I) \sum_{G\epsilon(A,B,C)} P(G|I,D)P(l^1|G)$$

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

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

$$P(l^1) = 0.502$$

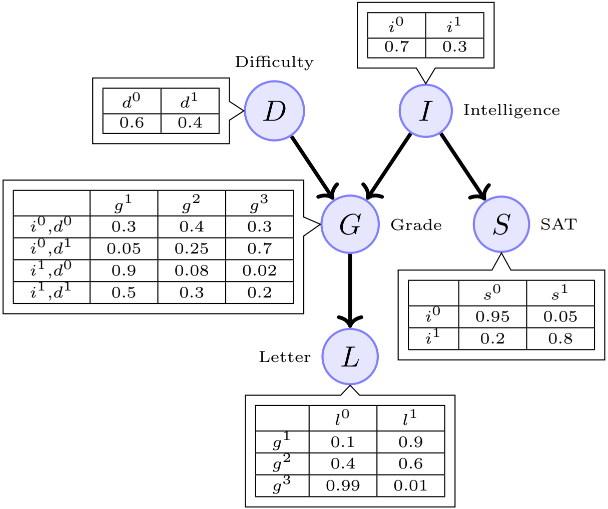


	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

Causal Reasoning

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



i^0	i^1
0.7	0.3

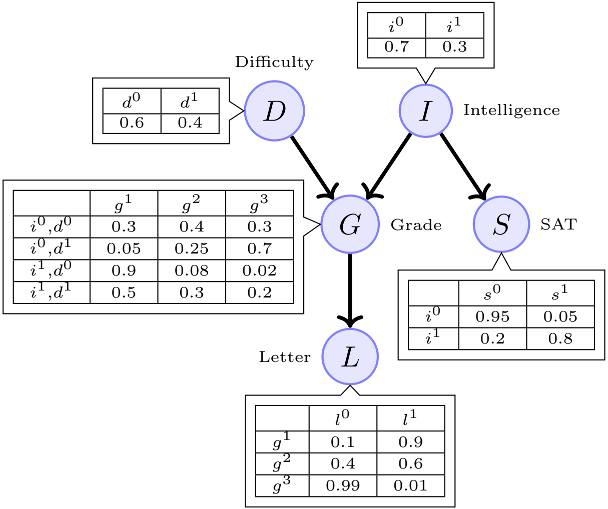
d^0	d^1
0.6	0.4

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Causal Reasoning



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

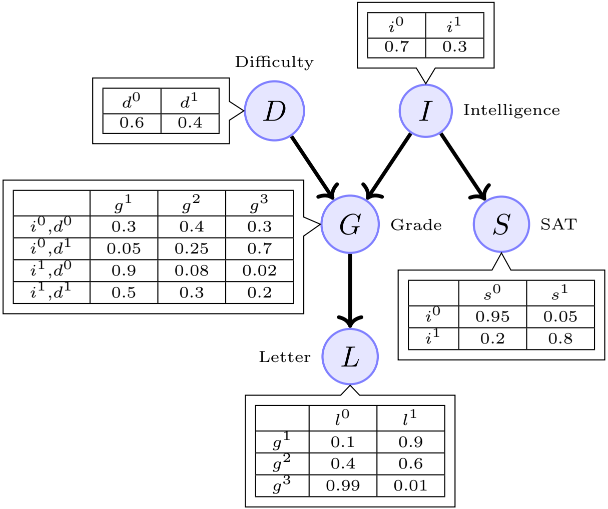
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

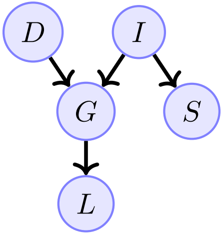
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

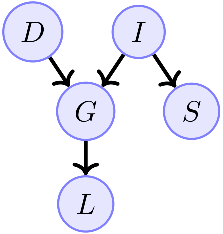
- Now what if we start adding information about the factors that could influence l^1
- What if someone reveals that the student is not intelligent?
- Intelligence will affect the score and hence the grade

$$P(l^1|i^0) = \frac{P(l^1, i^0)}{P(i^0)}$$



$$P(l^1|i^0) = \frac{P(l^1, i^0)}{P(i^0)}$$

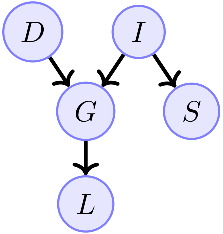
$$P(l^1, i^0) = \sum_{D\in\{0,1\}} \sum_{S\in\{0,1\}} \sum_{G\in\{A,B,C\}} P(i^0, D, G, S, l^1)$$



$$P(l^1|i^0) = \frac{P(l^1, i^0)}{P(i^0)}$$

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

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

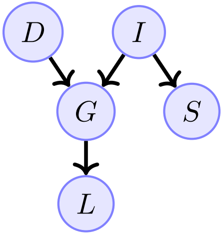


$$P(l^1|i^0) = \frac{P(l^1, i^0)}{P(i^0)}$$

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

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

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



	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

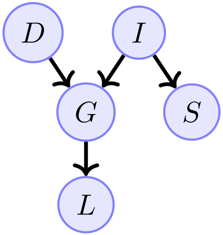
$$P(l^1|i^0) = \frac{P(l^1, i^0)}{P(i^0)}$$

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

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

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

$$P(l^1|i^0) = 0.389$$

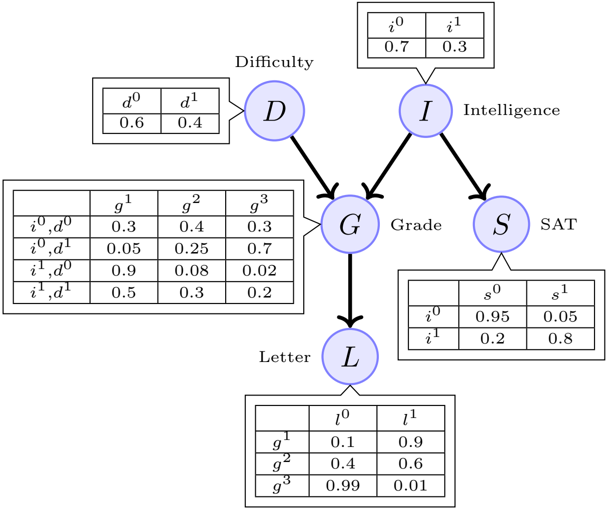


	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

Causal Reasoning

- What if the course was easy?



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

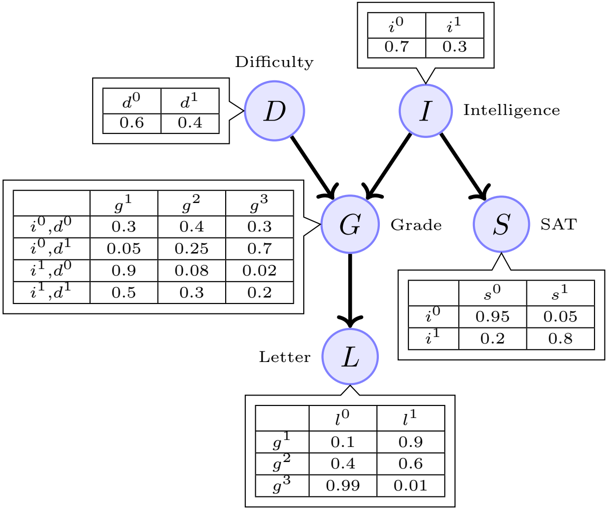
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

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



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

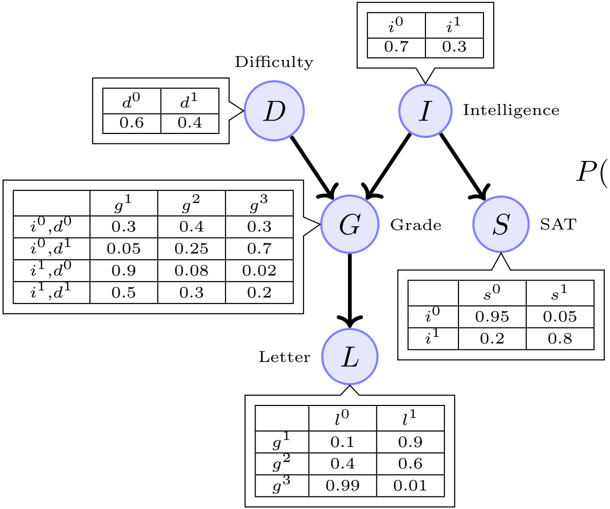
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

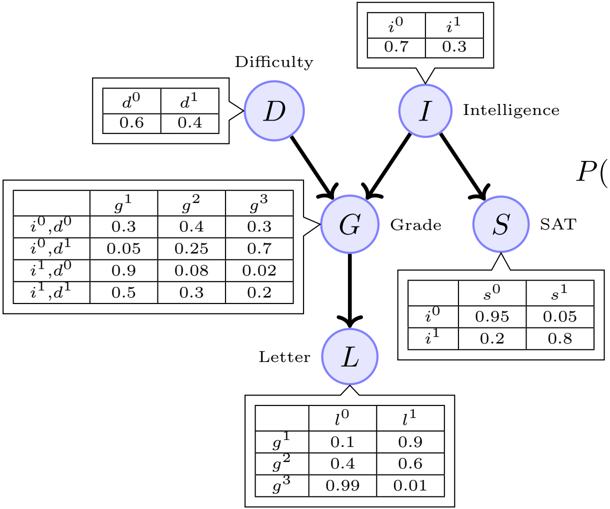
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

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

Causal Reasoning



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

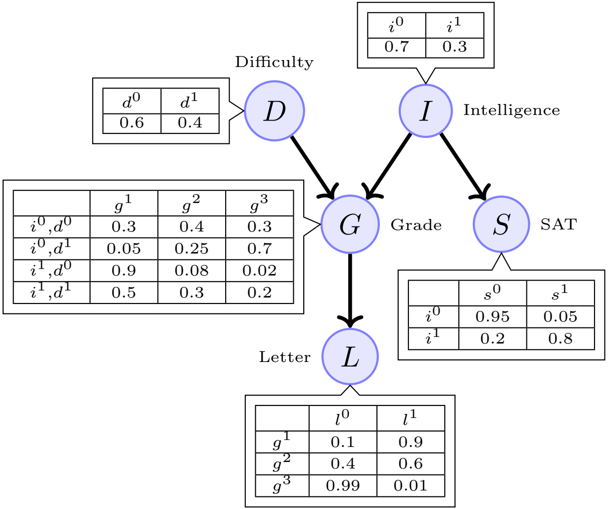
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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

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

$$P(l^1|i^0, d^1) = 0.513 \text{ (increases)}$$



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

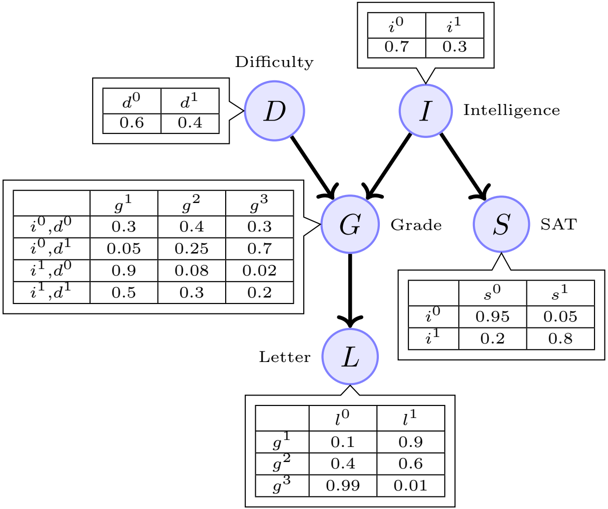
	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Evidential Reasoning

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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

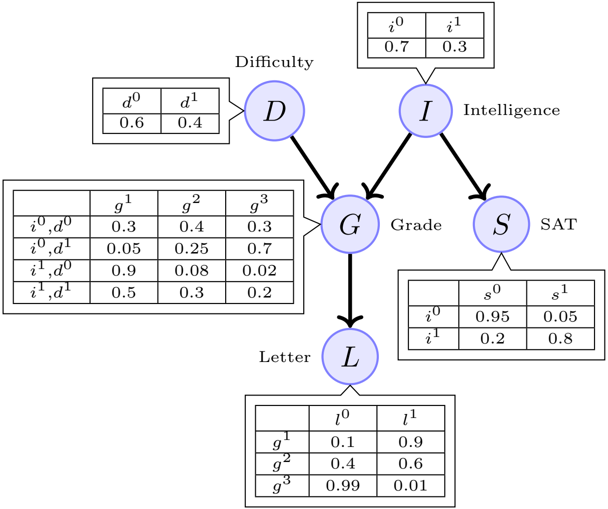
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Evidential Reasoning

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

$$P(i^1) = ?$$



i^0	i^1
0.7	0.3

d^0	d^1
0.6	0.4

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

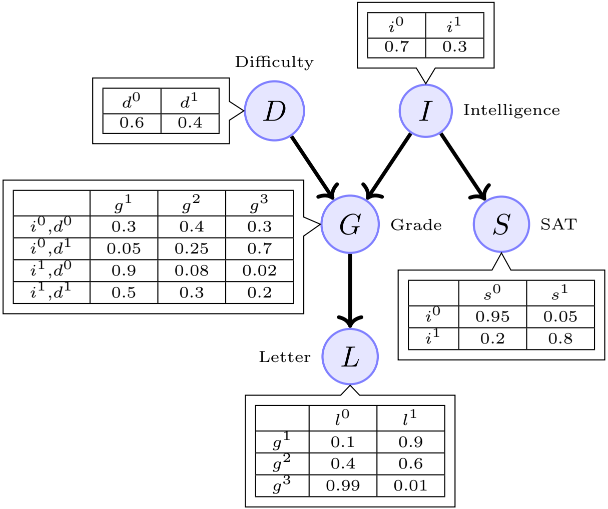
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

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



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

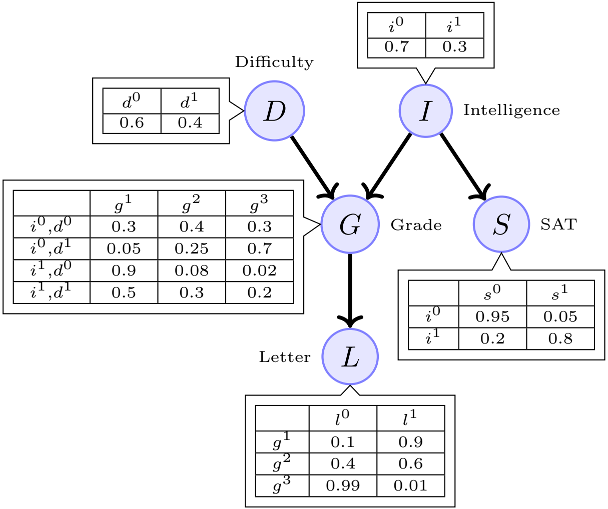
	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- 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?

$$P(i^1) = 0.3$$
$$P(d^1) = ?$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

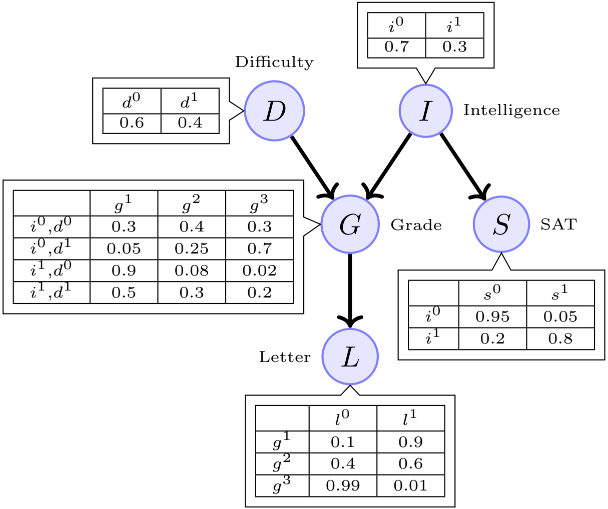
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

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?

$$P(i^1) = 0.3$$

$$P(d^1) = 0.4$$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

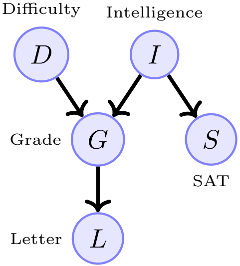
Evidential Reasoning

- 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

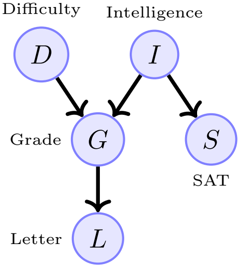
- What if someone tells us that the student secured C grade?

$P(i^1) = 0.3$

$P(d^1) = 0.4$

$P(i^1|g^3) = 0.079(drops)$

$P(d^1|g^3) = 0.629(increases)$



Evidential Reasoning

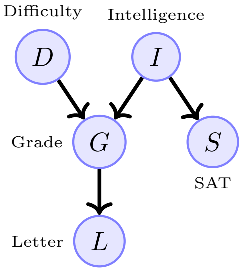
- What if someone tells us that the student secured C grade?

$P(i^1) = 0.3$

$P(d^1) = 0.4$

$P(i^1|g^3) = 0.079(drops)$

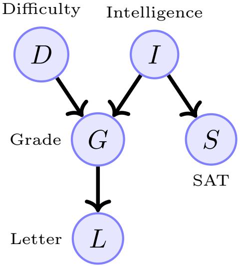
$P(d^1|g^3) = 0.629(increases)$



Evidential Reasoning

- What if someone tells us that the student secured C grade?
- What if instead of getting to know the grade, we get to know that the student got a poor recommendation letter?

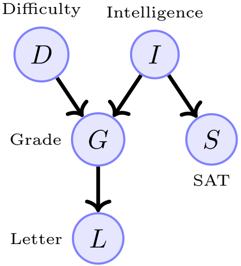
$$P(i^1) = 0.3$$
$$P(d^1) = 0.4$$
$$P(i^1|g^3) = 0.079(drops)$$
$$P(d^1|g^3) = 0.629(increases)$$
$$P(i^1|l^0) = 0.14(drops)$$



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?

$$P(i^1) = 0.3$$
$$P(d^1) = 0.4$$
$$P(i^1|g^3) = 0.079(drops)$$
$$P(d^1|g^3) = 0.629(increases)$$
$$P(i^1|l^0) = 0.14(drops)$$



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?

$P(i^1) = 0.3$

$P(d^1) = 0.4$

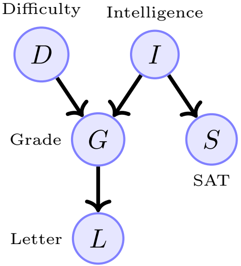
$P(i^1|g^3) = 0.079(drops)$

$P(d^1|g^3) = 0.629(increases)$

$P(i^1|l^0) = 0.14(drops)$

$P(l^1|l^0, g^3) = 0.079$

(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?

$P(i^1) = 0.3$

$P(d^1) = 0.4$

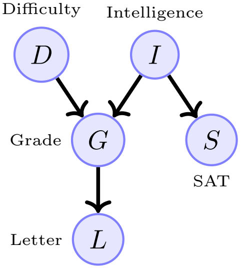
$P(i^1|g^3) = 0.079(drops)$

$P(d^1|g^3) = 0.629(increases)$

$P(i^1|l^0) = 0.14(drops)$

$P(l^1|l^0, g^3) = 0.079$

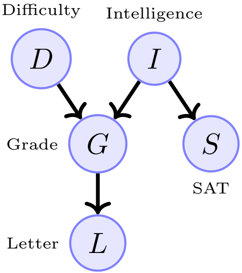
(same as $P(i^1|g^3)$)



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?
- The last case is interesting! (We will return to it later)

$$P(i^1) = 0.3$$

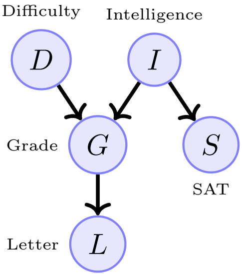


Explaining Away

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

$$P(i^1) = 0.3$$

$$P(i^1|g^3) = 0.079(drops)$$

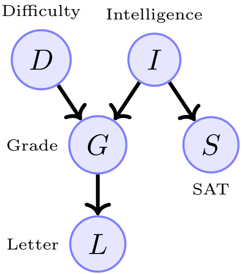


Explaining Away

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

$P(i^1) = 0.3$

$P(i^1|g^3) = 0.079(drops)$



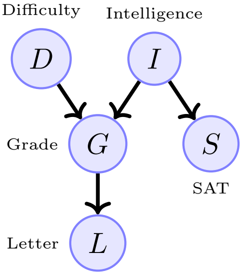
Explaining Away

- 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?

$P(i^1) = 0.3$

$P(i^1|g^3) = 0.079(drops)$

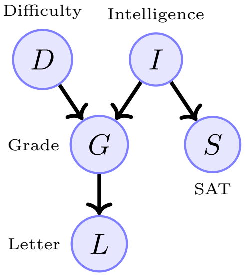
$P(i^1|g^3, d^1) = 0.11(improves)$



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

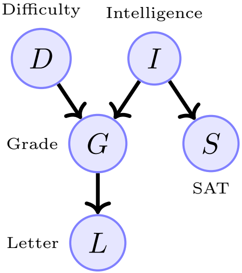
$$P(i^1) = 0.3$$
$$P(i^1|g^3) = 0.079(drops)$$
$$P(i^1|g^3, d^1) = 0.11(improves)$$



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

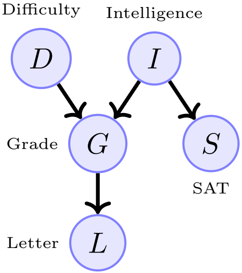
$$P(i^1) = 0.3$$
$$P(i^1|g^3) = 0.079$$
$$P(i^1|g^3, d^1) = 0.11$$



Explaining Away

- Knowing that the course was difficult explains away the bad grade

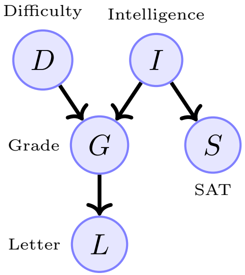
$$P(i^1) = 0.3$$
$$P(i^1|g^3) = 0.079$$
$$P(i^1|g^3, d^1) = 0.11$$



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!"

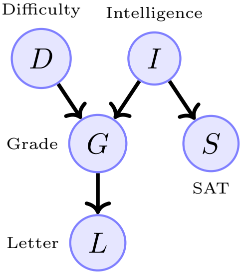
$$P(i^1) = 0.3$$
$$P(i^1|g^3) = 0.079$$
$$P(i^1|g^3, d^1) = 0.11$$



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

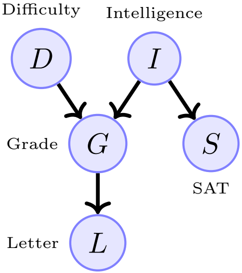
$$P(i^1) = 0.3$$
$$P(i^1|g^3) = 0.079$$
$$P(i^1|g^3, d^1) = 0.11$$



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

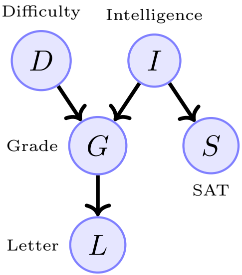
$$P(i^1) = 0.3$$
$$P(i^1|g^3) = 0.079$$
$$P(i^1|g^3, d^1) = 0.11$$
$$P(i^1|g^2) = 0.175$$
$$P(i^1|g^2, d^1) = 0.34$$



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

$$P(d^1) = 0.40$$
$$P(d^1|g^3) = 0.629$$



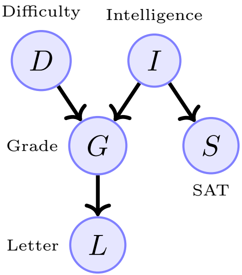
Explaining Away

- Suppose we know that the student had a high SAT Score, what happens to our belief about the difficulty of the course?

$$P(d^1) = 0.40$$

$$P(d^1|g^3) = 0.629$$

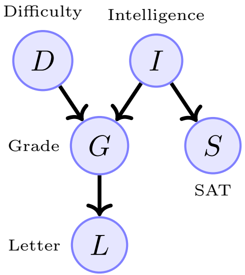
$$P(d^1|s^1, g^3) = 0.76$$



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?

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 the 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

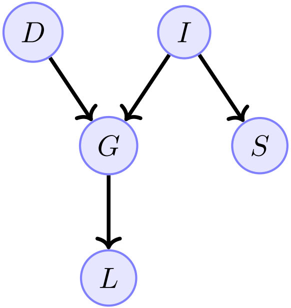
- $X_i \perp X_j$

In general, given n random variables, we are interested in knowing if

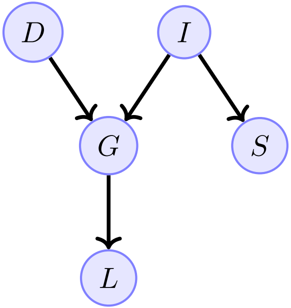
- $X_i \perp X_j$
- $X_i \perp X_j | Z$, where $Z \subseteq X_1, X_2, ..., X_n / X_i, X_j$

In general, given n random variables, we are interested in knowing if

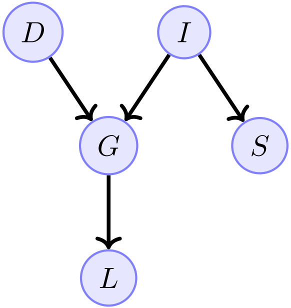
- $X_i \perp X_j$
- $X_i \perp X_j | Z$, where $Z \subseteq X_1, X_2, ..., X_n / X_i, X_j$
- Let us answer some of the questions for our student Bayesian Network



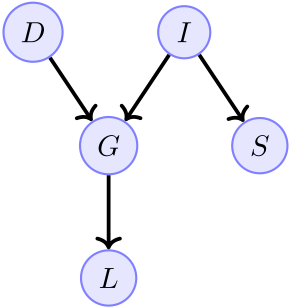
- To understand this let us return to our student example



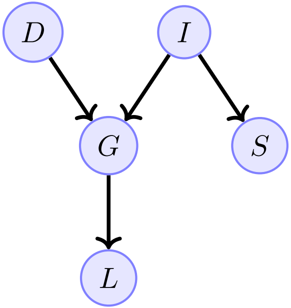
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph



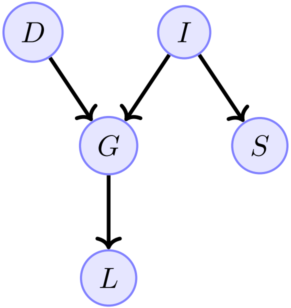
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph
- Is $L \perp G$? (No, by construction)



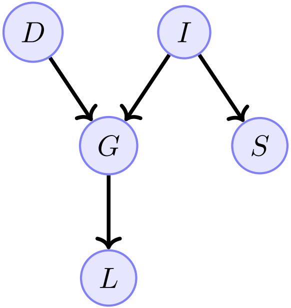
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph
- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)



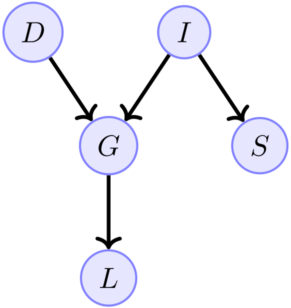
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph
- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)



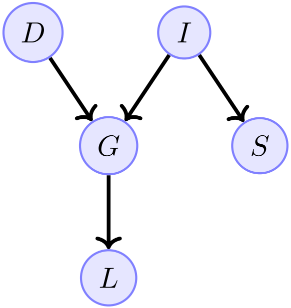
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph
- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)



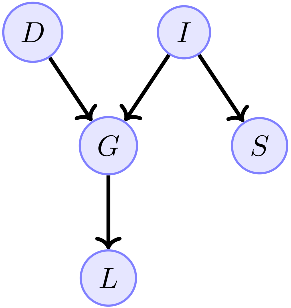
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph
- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)
- **Rule?**



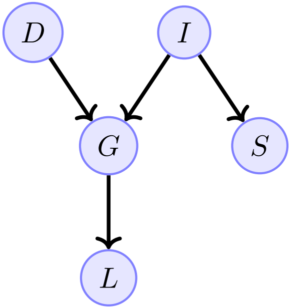
- To understand this let us return to our student example
- First, let us see some independencies which clearly do not exist in the graph
- Is $L \perp G$? (No, by construction)
- Is $G \perp D$? (No, by construction)
- Is $G \perp I$? (No, by construction)
- Is $S \perp I$? (No, by construction)
- **Rule:** A node is not independent of its parents



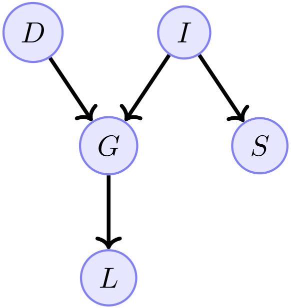
- Let us focus on G and L.



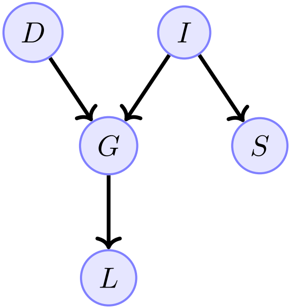
- Let us focus on G and L.
- We already know that $G \not\perp L$.



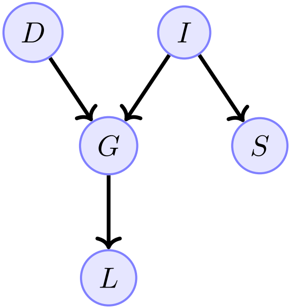
- Let us focus on G and L.
- We already know that $G \not\perp L$.
- What if we know the value of I? Does G become independent of L?



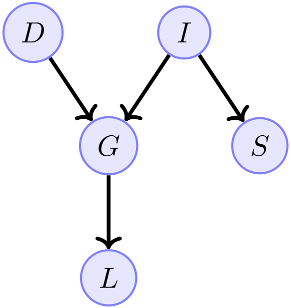
- Let us focus on G and L.
- We already know that $G \not\perp L$.
- What if we know the value of I? Does G become independent of L?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)



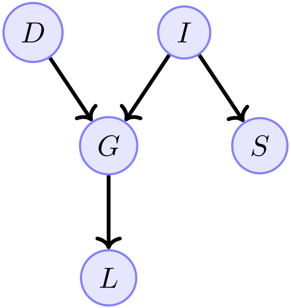
- Let us focus on G and L.
- We already know that $G \not\perp L$.
- What if we know the value of I? Does G become independent of L?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of D, does G become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

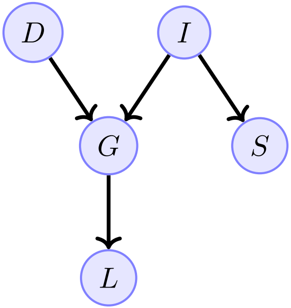


- Let us focus on G and L.
- We already know that $G \not\perp L$.
- What if we know the value of I? Does G become independent of L?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of D, does G become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



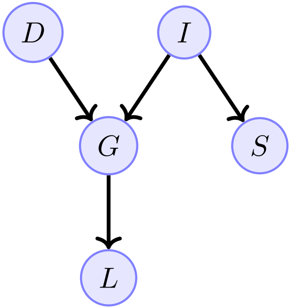
- No, the instructor is not going to look at the SAT score but the grade

- 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?



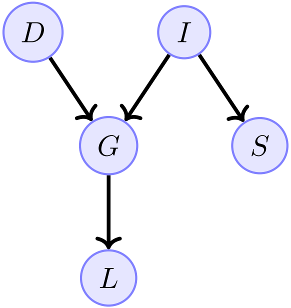
- No, the instructor is not going to look at the SAT score but the grade
- **Rule?**

- 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?

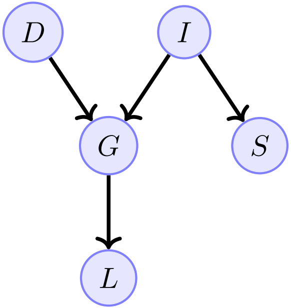


- No, the instructor is not going to look at the SAT score but the grade
- **Rule**: A node is not independent of its parents even when we are given the values of other variables

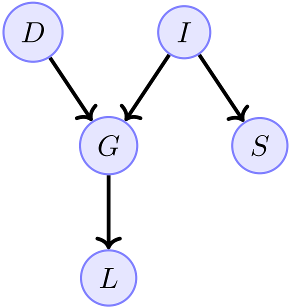
- Let us focus on G and L.
- We already know that $G \not\perp L$.
- What if we know the value of I? Does G become independent of L?
- No (intuitively, the student may be intelligent or not but ultimately, the letter depends on the performance in the course.)
- If we know the value of D, does G become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



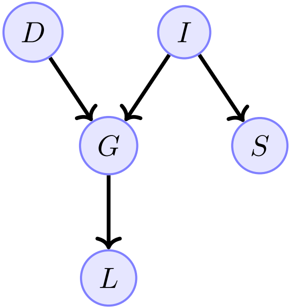
- The same argument can be made about the following pairs



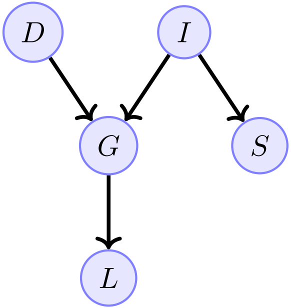
- The same argument can be made about the following pairs
- $G \not\perp D$ (even when other variables are given)



- 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)

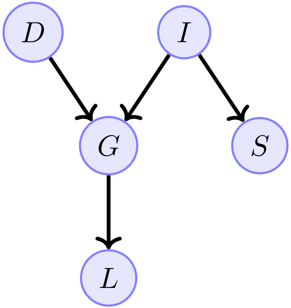


- The same argument can be made about the following pairs
- $G \not\perp D$ (even when other variables are given)
- $G \not\perp I$ (even when other variables are given)
- $S \not\perp I$ (even when other variables are given)



- **Rule?**

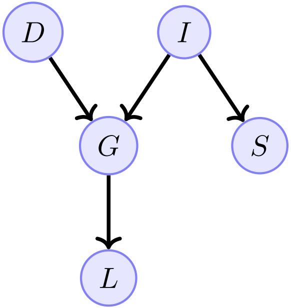
- The same argument can be made about the following pairs
- $G \not\perp D$ (even when other variables are given)
- $G \not\perp I$ (even when other variables are given)
- $S \not\perp I$ (even when other variables are given)



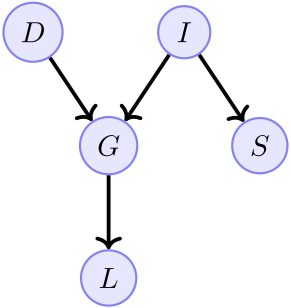
- **Rule:** A node is not independent of its parents even when we are given the values of other variables

- 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)

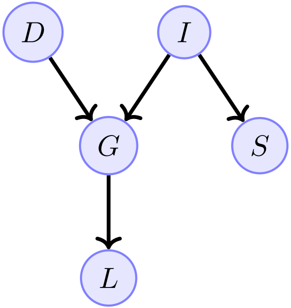
Module 17.7: Independencies encoded by a Bayesian network (Case 2: Node and its non-parents)



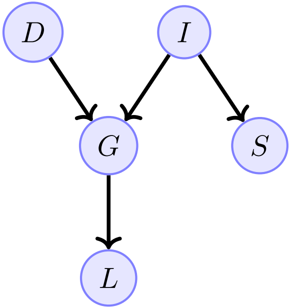
- Now let's look at the relation between a node and its non-parent nodes



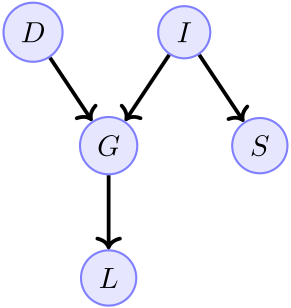
- Now let's look at the relation between a node and its non-parent nodes
- Is $L \perp S$?



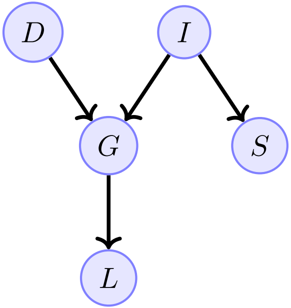
- Now let's look at the relation between a node and its non-parent nodes
- Is $L \perp S$?
- No, knowing the SAT score tells us about I which in turn tells us something about G and hence L



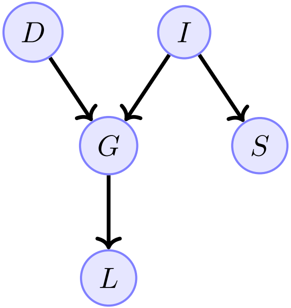
- Now let's look at the relation between a node and its non-parent nodes
- Is $L \perp S$?
- No, knowing the SAT score tells us about I which in turn tells us something about G and hence L
- Hence we expect $P(l^1|s^1) > P(l^1|s^0)$



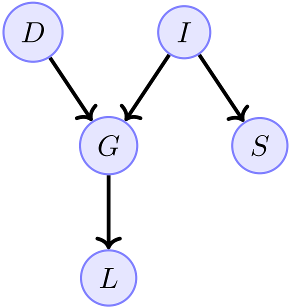
- Now let's look at the relation between a node and its non-parent nodes
- Is $L \perp S$?
- No, knowing the SAT score tells us about I which in turn tells us something about G and hence L
- Hence we expect $P(l^1|s^1) > P(l^1|s^0)$
- Similarly we can argue $L \not\perp D$ and $L \not\perp I$



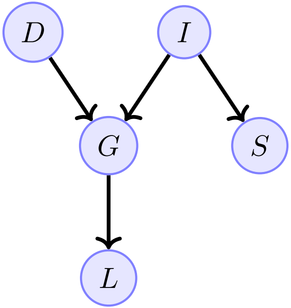
- But what if we know the value of G?



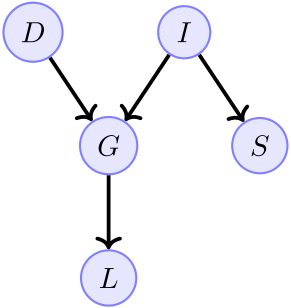
- But what if we know the value of G?
- Is $(L \perp S)|G$?



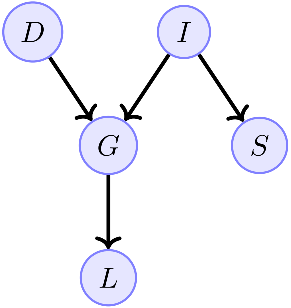
- But what if we know the value of G?
- Is $(L \perp S)|G$?
- Yes, the grade completely determines the recommendation letter



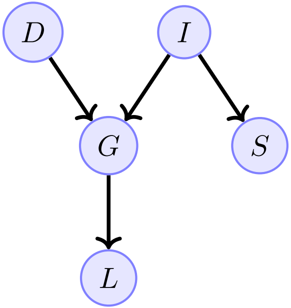
- But what if we know the value of G?
- Is $(L \perp S)|G$?
- Yes, the grade completely determines the recommendation letter
- Once we know the grade, other variables do not add any information



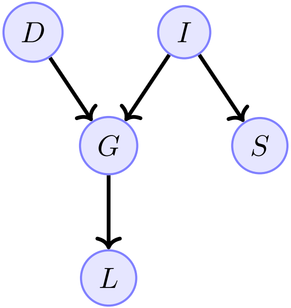
- But what if we know the value of G?
- Is $(L \perp S)|G$?
- Yes, the grade completely determines the recommendation letter
- Once we know the grade, other variables do not add any information
- Hence $(L \perp S)|G$



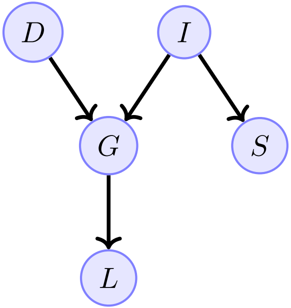
- But what if we know the value of G?
- Is $(L \perp S)|G$?
- Yes, the grade completely determines the recommendation letter
- Once we know the grade, other variables do not add any information
- Hence $(L \perp S)|G$
- Similarly we can argue $(L \perp I)|G$ and $(L \perp D)|G$



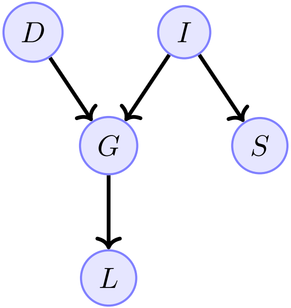
- But, wait a minute!



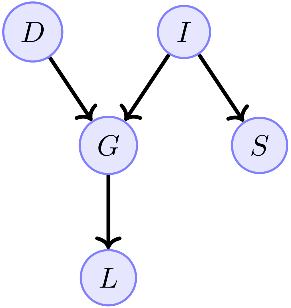
- But, wait a minute!
- The instructor may also want to look at the SAT score in addition to the grade



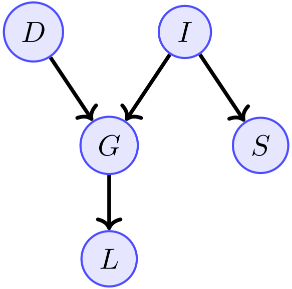
- But, wait a minute!
- The instructor may also want to look at the SAT score in addition to the grade
- Well, we "assumed" that the instructor only relies on the grade.



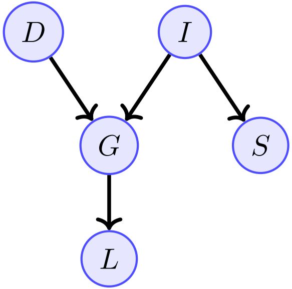
- But, wait a minute!
- The instructor may also want to look at the SAT score in addition to the grade
- Well, we "assumed" that the instructor only relies on the grade.
- That was our "belief" of how the world works



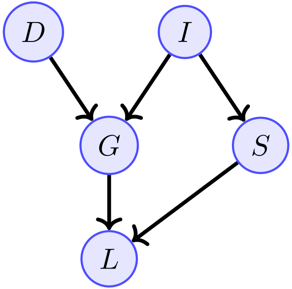
- But, wait a minute!
- The instructor may also want to look at the SAT score in addition to the grade
- Well, we "assumed" that the instructor only relies on the grade.
- That was our "belief" of how the world works
- And hence we drew the network accordingly



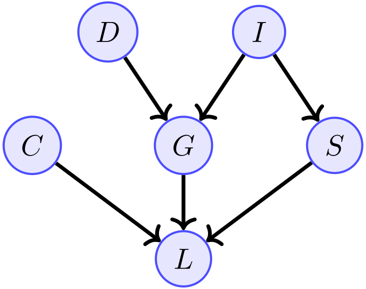
- Of course we are free to change our assumptions



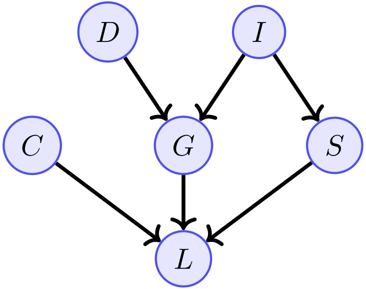
- Of course we are free to change our assumptions
- We may want to assume that the instructor also looks at the SAT score



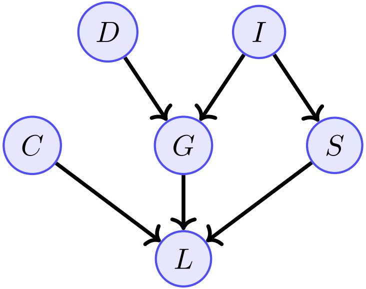
- Of course we are free to change our assumptions
- We may want to assume that the instructor also looks at the SAT score
- But if that is the case we have to change the network to reflect this dependence



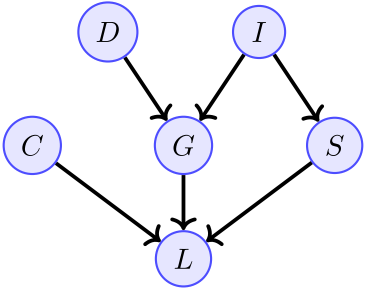
- Of course we are free to change our assumptions
- We may want to assume that the instructor also looks at the SAT score
- But if that is the case we have to change the network to reflect this dependence
- Why just SAT score? The instructor may even consult one of his colleagues and seek his/her opinion



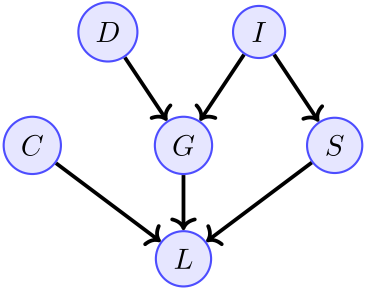
- Remember: The graph is a reflection of our assumptions about how the world works



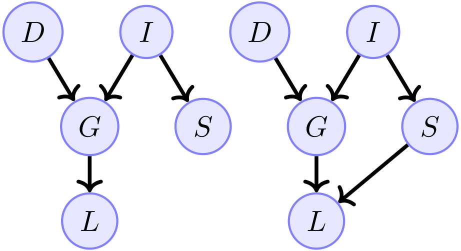
- Remember: The graph is a reflection of our assumptions about how the world works
- Our assumptions about dependencies are encoded in the graph



- 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

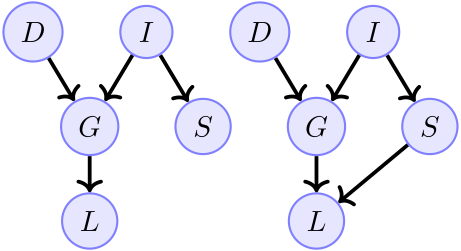


- Remember: The graph is a reflection of our assumptions about how the world works
- Our assumptions about dependencies are encoded in the graph
- Once we build the graph we freeze it and do all the reasoning and analysis (independence) on this graph
- It is not fair to ask "what if" questions involving other factors (For example, what if the professor was in a bad mood?)



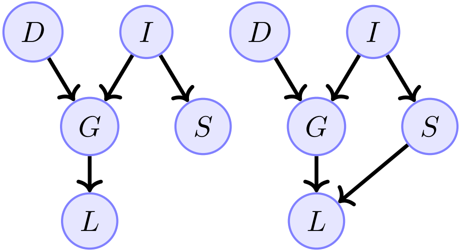
(a) **(b)**

- If we believe Graph (a) is how the world works then $(L \perp S)|G$



(a) (b)

- 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 $(L \not\perp S)|G$



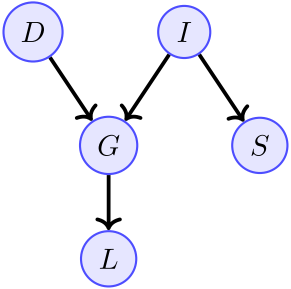
(a) **(b)**

- 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

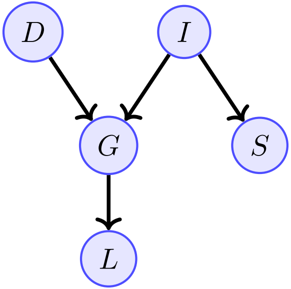
- Let's return back to our discussion of finding independence relations in the graph

- 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

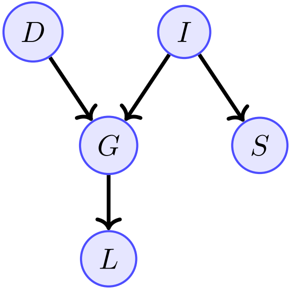
Module 17.8: Independencies encoded by a Bayesian network (Case 3: Node and its descendants)



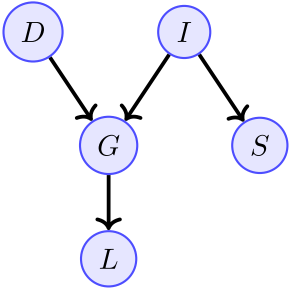
- $(G \not\perp D)$ $(G \not\perp I)$ $(S \not\perp I)$ $(L \not\perp G)$
  A node is not independent of its parents



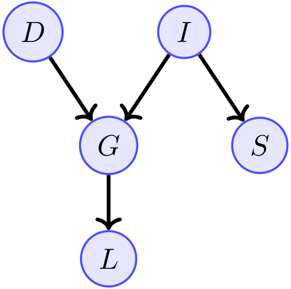
- $(G \not\perp D)$ $(G \not\perp I)$ $(S \not\perp I)$ $(L \not\perp G)$
  A node is not independent of its parents
- $(G \not\perp D, I)|S, L$
  $(S \not\perp I)|D, G, L$
  $(L \not\perp G)|D, I, S$
  A node is not independent of its parents even when other variables are given



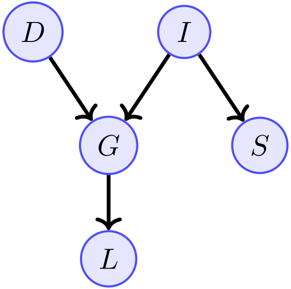
- $(G \not\perp D)$ $(G \not\perp I)$ $(S \not\perp I)$ $(L \not\perp G)$
  A node is not independent of its parents
- $(G \not\perp D, I)|S, L$
  $(S \not\perp I)|D, G, L$
  $(L \not\perp G)|D, I, S$
  A node is not independent of its parents even when other variables are given
- $(S \perp G)|I?$
  $(L \perp D, I, S)|G?$
  $(G \perp L)|D, I?$
  A node **seems to be** independent of other variables given its parents



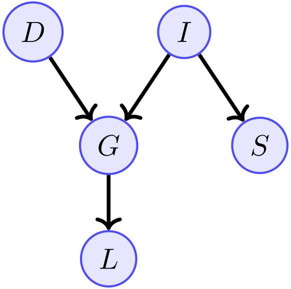
- Let us inspect this last rule



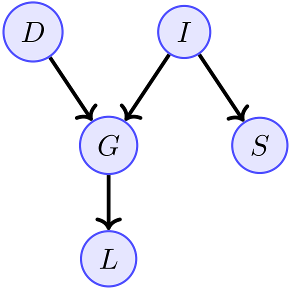
- Let us inspect this last rule
- Is $(G \perp L)|D, I$?



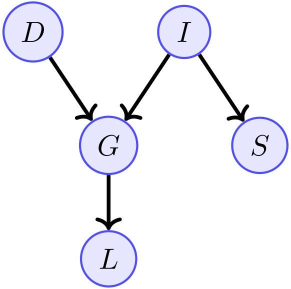
- Let us inspect this last rule
- Is $(G \perp L)|D, I$?
- If you know that $d = 0$ and $i = 1$ then you would expect the student to get a good grade



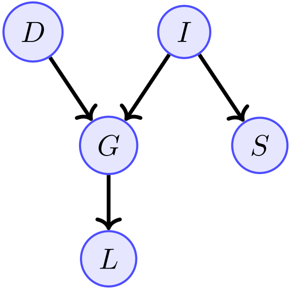
- Let us inspect this last rule
- Is $(G \perp L)|D, I$?
- If you know that $d = 0$ and $i = 1$ then you would expect the student to get a good grade
- But now if someone tells you that the student got a poor letter, your belief will change



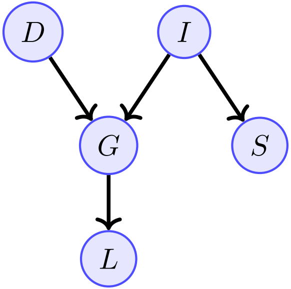
- Let us inspect this last rule
- Is $(G \perp L)|D, I$?
- If you know that $d = 0$ and $i = 1$ then you would expect the student to get a good grade
- But now if someone tells you that the student got a poor letter, your belief will change
- So $(G \not\perp L)|D, I$



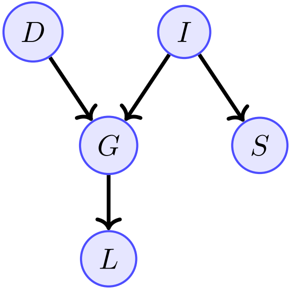
- Let us inspect this last rule
- Is $(G \perp L)|D, I$?
- If you know that $d = 0$ and $i = 1$ then you would expect the student to get a good grade
- But now if someone tells you that the student got a poor letter, your belief will change
- So $(G \not\perp L)|D, I$
- The effect (letter) actually gives 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



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



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

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 $X_1, X_2, ..., X_n$.

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 $P_{a_{X_i}}^{G}$ denote the parents of X_i in G

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 $P_{a_{X_i}}^G$ 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.

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 $P_{a_{X_i}}^G$ 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.

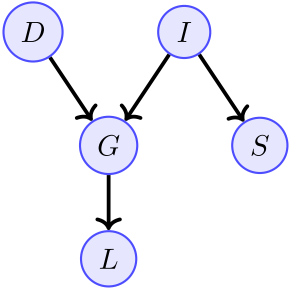
We are now ready to formally define the semantics of a Bayesian Network

Bayesian Network Semantics:

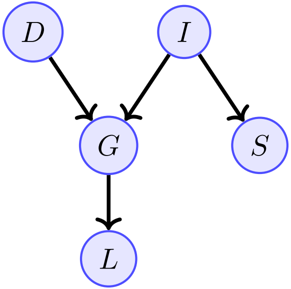
A Bayesian Network structure G is a directed acyclic graph where nodes represent random variables $X_1, X_2, ..., X_n$. Let $P^G_{a_{X_i}}$ denote the parents of X_i in G and NonDescendants(X_i) denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of conditional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

- We will see some more formal definitions and then return to the question of independencies.

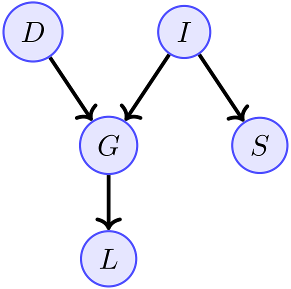
Module 17.10: I Maps



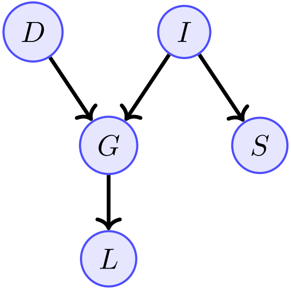
- Let P be a joint distribution over $X = X_1, X_2, ..., X_n$



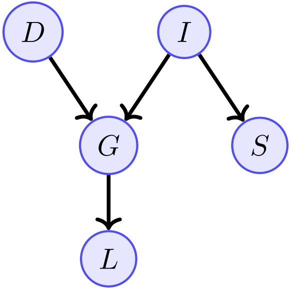
- Let P be a joint distribution over $X = X_1, X_2, ..., X_n$
- We define $I(P)$ as the set of independence assumptions that hold in P.



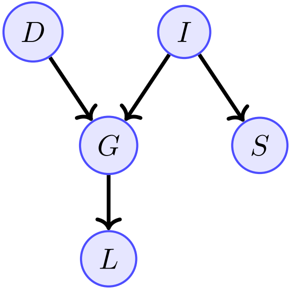
- Let P be a joint distribution over $X = X_1, X_2, ..., X_n$
- We define $I(P)$ as the set of independence assumptions that hold in P.
- For Example: $I(P) = \{(G \perp S | I, D), ...\}$



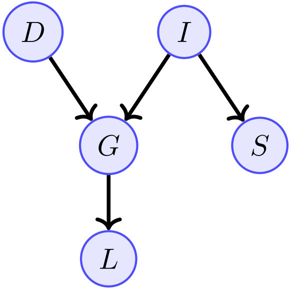
- Let P be a joint distribution over $X = X_1, X_2, ..., X_n$
- We define $I(P)$ as the set of independence assumptions that hold in P.
- For Example: $I(P) = \{(G \perp S | I, D), ...\}$
- Each element of this set is of the form $X_i \perp X_j | Z, Z \subseteq X | X_i, X_j$



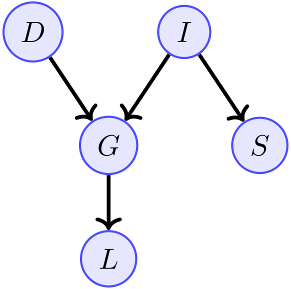
- Let P be a joint distribution over $X = X_1, X_2, ..., X_n$
- We define $I(P)$ as the set of independence assumptions that hold in P.
- For Example: $I(P) = \{(G \perp S|I, D), ...\}$
- Each element of this set is of the form $X_i \perp X_j|Z, Z \subseteq X|X_i, X_j$
- Let $I(G)$ be the set of independence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



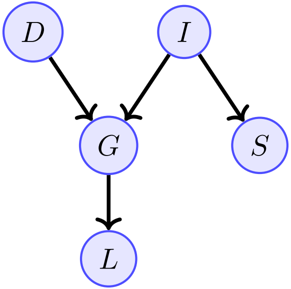
- We say that G is an I-map for P if $I(G) \subseteq I(P)$



- 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.

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

- Consider this joint distribution over X, Y

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

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

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

- 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?

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

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

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

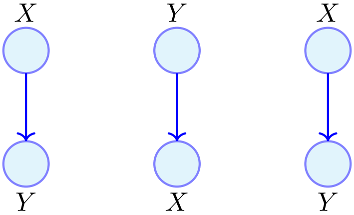
- Well since there are only 2 variables here the only possibilities are $I(P) = \{(X \perp Y)\}$ or $I(P) = \Phi$
- From the table we can easily check $P(X,Y) = P(X).P(Y)$

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

- Well since there are only 2 variables here the only possibilities are $I(P) = \{(X \perp Y)\}$ or $I(P) = \Phi$
- From the table we can easily check $P(X, Y) = P(X).P(Y)$
- $I(P) = \{(X \perp Y)\}$

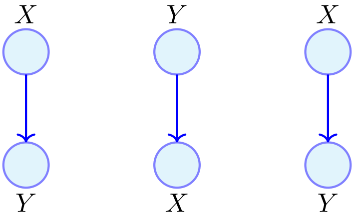
X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

- Well since there are only 2 variables here the only possibilities are $I(P) = \{(X \perp Y)\}$ or $I(P) = \Phi$
- 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 $I(G) \subseteq I(P)$?



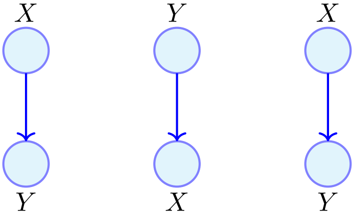
$I(G) = \Phi \quad I(G_2) = \Phi \quad I(G_3) = \{(X \perp Y)\}$

- Since we have only two variables there are only 3 possibilities for G



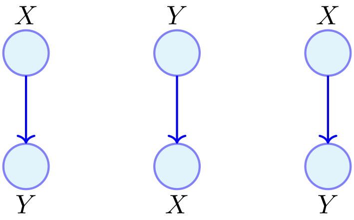
$I(G) = \Phi \quad I(G_2) = \Phi \quad I(G_3) = \{(X \perp Y)\}$

- 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



$I(G) = \Phi \quad I(G_2) = \Phi \quad I(G_3) = \{(X \perp Y)\}$

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

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

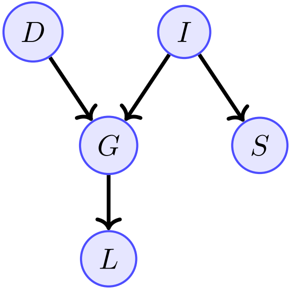
- Of course, this was just a toy example

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

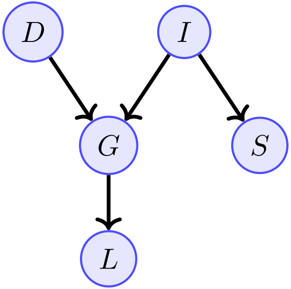
- Of course, this was just a toy example
- In practice, we do not know P and hence can't compute $I(P)$

X	Y	P(X,Y)
0	0	0.08
0	1	0.32
1	0	0.12
1	1	0.48

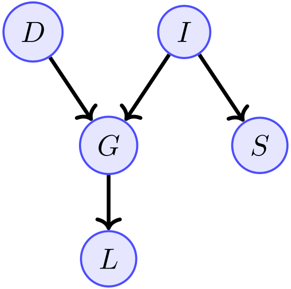
- Of course, this was just a toy example
- In practice, we do not know P and hence can't compute $I(P)$
- We just make some assumptions about $I(P)$ and then construct a G such that $I(G) \subseteq I(P)$



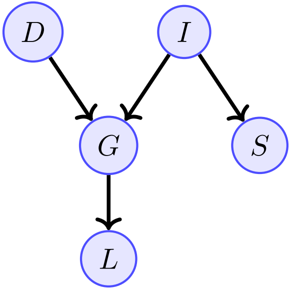
- So why do we care about I-Map?



- So why do we care about I-Map?
- If G is an I-Map for a joint distribution P then P factorizes over G



- 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

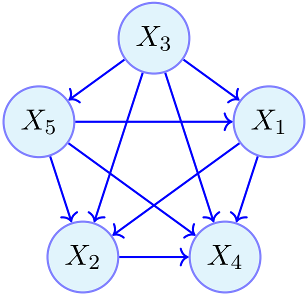
- Consider a set of random variables X_1, X_2, X_3, X_4, X_5

- 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

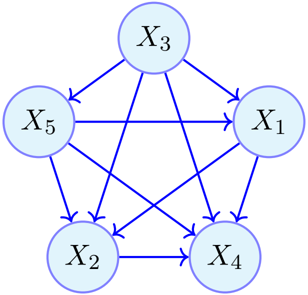
- Consider a set of random variables X_1, X_2, X_3, X_4, X_5
- There are many joint distributions possible
- Each may entail different independence relations
- For example, in some cases L could be independent of S; in some not.

- 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?



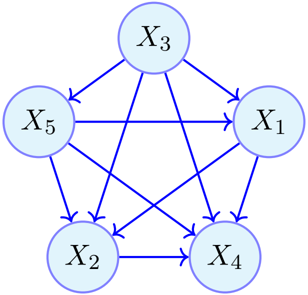
- What is this graph called?

- 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?



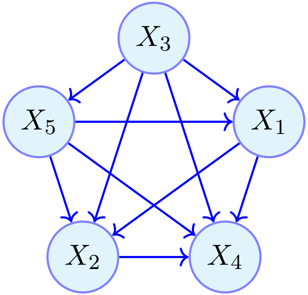
- Answer: A complete graph

- 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 $P(X_3)P(X_5|X_3)P(X_1|X_3,X_5)$ $P(X_2|X_1,X_3,X_5)P(X_4|X_1,X_2,X_3,X_5)$

- 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?