18-661 Introduction to Machine Learning

Graphical Models - I

Fall 2020

ECE – Carnegie Mellon University

Midterm Information

Midterm will be on Tuesday, 10/20 in-class.

- Conducted as an online exam on Gradescope, with multiple-choice and short-answer questions
- Closed-book except for one double-sided letter-size handwritten page of notes that you can prepare as you wish.
- We will provide formulas for relevant probability distributions.
- You will not need a calculator. Only pen/pencil and scratch paper are allowed.

Will cover all topics up to and including Nearest Neighbors (10/15)

- (1) point estimation/MLE/MAP, (2) linear regression, (3) naive Bayes, (4) logistic regression, (5) SVMs, (6) Graphical Models, (7) Nearest Neighbors.
- Practice Midterm exam has been posted on Gradescope
- This friday's recitation will go over practice exam questions.

Outline

1. Review of Interdependence Random Variables
2. Probabilistic Graphical Models
3. Examples of Graphical Models
4. Bayes Ball Theorem (d-separation)
5. Related Concepts

Review of Interdependence Random Variables

Joint Distribution of Two Random Variables

- Consider two discrete random variables X and Y.
- X can take one of the values $\mathcal{X} = \{0, 1, \ldots, m\}$ and Y can take one of the values $\mathcal{Y} = \{0, 1, \ldots, n\}$
- The joint distribution $p(x, y)$ gives the probability of each pair of realizations (x, y)

$$\begin{aligned} p(x, y) &= \Pr(X = x, Y = y) \quad \text{for } x \in \mathcal{X}, y \in \mathcal{Y} \\ &= \Pr(X = x) \Pr(Y = y | X = x) \quad \text{by Bayes rule} \\ &= p(x)p(y|x) \quad \text{shorter way of writing it} \\ &= p(y)p(x|y) \end{aligned}$$

- What if X and Y are independent?

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y) = p(x)p(y)$$

because the conditionals $p(y|x) = p(y)$, and $p(x|y) = p(x)$.

Marginals and Conditionals from the Joint Distribution

- For two discrete random variables X and Y, the marginal distributions of X and Y are $p(x) = \Pr(X = x)$ and $p(y) = Pr(Y = y)$ respectively
- Given the joint distribution $p(x, y)$, how do you find the marginals?

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

$$p(y) = \sum_{x \in \mathcal{X}} p(x, y)$$

- And how do you express the conditional distributions in terms of the joint and marginals?

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\sum_{y \in \mathcal{Y}} p(x, y)}$$

Example 1: Find the Marginals and Conditionals

- X is whether a person is infected with COVID ($x = 1$) or not ($x = 0$)
- Y is whether they test positive ($y = 1$) or negative ($y = 0$)

p(y)		

p(x)

p(x,y)	y=0	y=1
x=0	0.5	0.1
x=1	0.1	0.3

p(y|x)	y=0	y=1
x=0		
x=1		

p(x|y)	y=0	y=1
x=0		
x=1		

Compute the following:

- Marginals $p(x)$ and $p(y)$
- Conditionals $p(x|y)$ and $p(y|x)$

Example 1: Find the Marginals and Conditionals

- X is whether a person is infected with COVID ($x = 1$) or not ($x = 0$)
- Y is whether they test positive ($y = 1$) or negative ($y = 0$)

p(y)	0.6	0.4

p(x)
0.6
0.4

p(x,y)	y=0	y=1
x=0	0.5	0.1
x=1	0.1	0.3

p(y|x)	y=0	y=1
x=0		
x=1		

p(x|y)	y=0	y=1
x=0		
x=1		

Compute the following:

- Marginals $p(x)$ and $p(y)$ are the row and column sums
- Conditionals $p(x|y)$ and $p(y|x)$

Example 1: Find the Marginals and Conditionals

- X is whether a person is infected with COVID ($x = 1$) or not ($x = 0$)
- Y is whether they test positive ($y = 1$) or negative ($y = 0$)

p(y)	0.6	0.4

p(x)
0.6
0.4

p(x,y)	y=0	y=1
x=0	0.5	0.1
x=1	0.1	0.3

p(y|x)	y=0	y=1
x=0	5/6	1/6
x=1	1/4	3/4

p(x|y)	y=0	y=1
x=0	5/6	1/4
x=1	1/6	3/4

Compute the following:

- Marginals $p(x)$ and $p(y)$ are the row and column sums
- Conditionals $p(x|y)$ and $p(y|x)$ are the normalized rows and columns

Example 1: Find the Marginals and Conditionals

- X is the first coin toss: heads ($x = 1$) or tails ($x = 0$)
- Y is the second coin toss: heads ($y = 1$) or tails ($y = 0$)
- X and Y are independent

p(y)	0.6	0.4

p(x)
0.6
0.4

p(x,y)	y=0	y=1
x=0	0.36	0.24
x=1	0.24	0.16

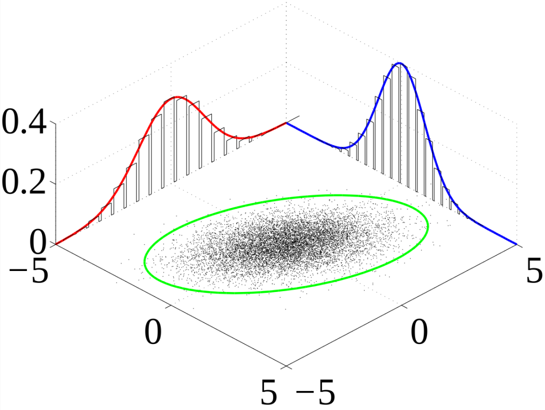
p(y|x)	y=0	y=1
x=0	0.6	0.4
x=1	0.6	0.4

p(x|y)	y=0	y=1
x=0	0.6	0.6
x=1	0.4	0.4

- Marginals $p(x)$ and $p(y)$ are the row and column sums
- Conditionals $p(x|y) = p(x)$ and $p(y|x) = p(y)$ due to independence

Example 2: Multi-variate Gaussian

- X, Y correlated Gaussian random variables

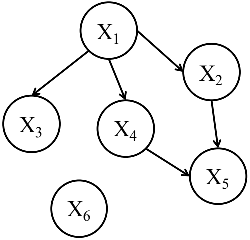


- Joint Distribution is shown in Green
- Marginals $p(x)$ and $p(y)$ are shown in red and blue

Probabilistic Graphical Models

Directed Graphical Models (also called Bayesian Networks)

- **Nodes** represent random variables
- **Edges** represent conditional dependencies
- Directed acyclic graph – no loops

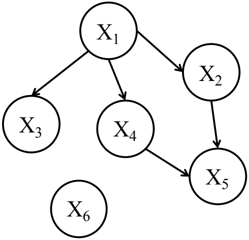


Advantages

1. Compact way of describing a family of joint dist. (this lecture)
2. Enable us to visualize conditional dependencies (this lecture)
3. Enable us to perform inference using observed data (next lecture)

Graphical Models (also called Bayesian Networks)

- **Nodes** represent random variables
- **Edges** represent conditional dependencies
- Directed acyclic graph – no loops



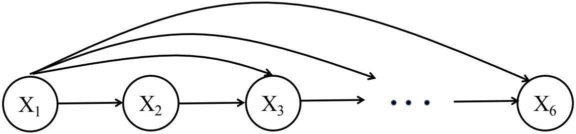
Advantages

1. **Compact way of describing a family of joint dist.** (this lecture)
2. Enable us to visualize conditional dependencies (this lecture)
3. Enable us to perform inference using observed data (next lecture)

Compact Way of Writing the Joint Distribution

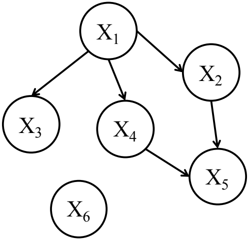
By Bayes Rule, the joint distribution of any set of random variables can be written as the product of conditionals

$$
\begin{aligned}
& p(x_1, x_2, \ldots, x_6) \\
& = p(x_1)p(x_2, x_3, x_4, x_5, x_6|x_1) \quad \text{Pull out } x_1 \\
& = p(x_1)p(x_2|x_1)p(x_3, x_4, x_5, x_6|x_1, x_2) \quad \text{Pull out } x_1 \text{ then } x_2 \\
& = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3) \ldots p(x_6|x_1, \ldots, x_5)
\end{aligned}
$$



- This is a fully connected graph – can represent any distribution
- Each node is conditioned on its parents
- Changing the order of the nodes will give a different fully connected graph, but correspond to the same joint distribution

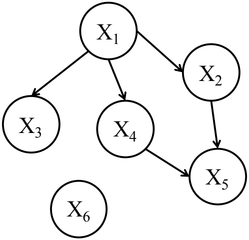
Compact Way of Writing the Joint Distribution



$$
\begin{aligned}
p(x_1, x_2, \ldots, x_6) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3)\ldots p(x_6|x_1, \ldots, x_5) \\
&= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_1)p(x_5|x_2, x_4)p(x_6)
\end{aligned}
$$

- Removing edges corresponds to conditional independence
- Given X_1, X_4 is independent of X_2 and X_3, that is, $p(x_4|x_1, x_2 x_3) = p(x_4)$
- X_6 is independent of all other variables: $p(x_6|x_1, \ldots, x_5) = p(x_6)$

Compact Way of Writing the Joint Distribution



Method to write the joint dist. described by any directed acyclic graph

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_{\pi_i})$$

where x_{π_i} is the set of parents of node i. For example,

- $x_{\pi_1} = \{\}$
- $x_{\pi_2} = \{x_1\}$, $x_{\pi_3} = \{x_1\}$, $x_{\pi_4} = \{x_1\}$
- $x_{\pi_5} = \{x_2, x_4\}$
- $x_{\pi_6} = \{\}$

Storage Complexity of the Joint Distribution

- An arbitrary joint distribution of 6 binary random variables will be a 6-dimensional table with 2^6 entries in total – this corresponds to the fully connected graph.

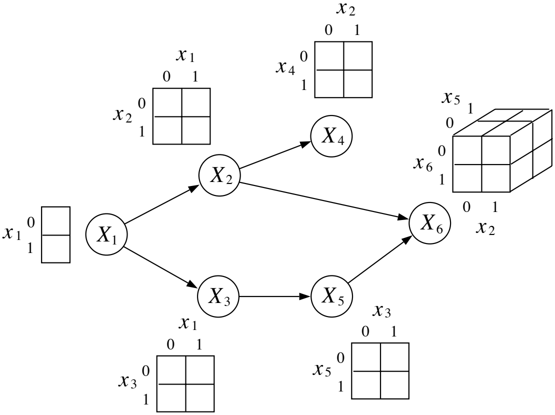


Image Source: Prof. Michael Jordan's lecture notes on Probabilistic Graphical Models

Storage Complexity of the Joint Distribution

- Due to conditional independencies, storage complexity is reduced
- Each node with d parents needs to store a $d + 1$ dimensional table

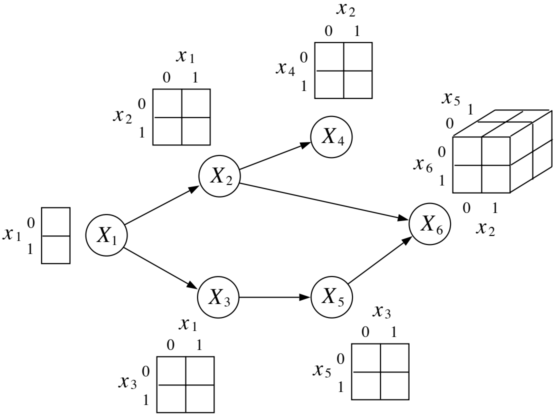
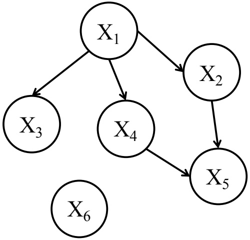


Image Source: Prof. Michael Jordan's lecture notes on Probabilistic Graphical Models

Graphical Models (also called Bayesian Networks)

- **Nodes** represent random variables
- **Edges** represent conditional dependencies
- Directed acyclic graph – no loops

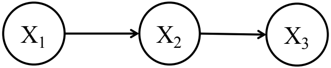


Advantages

1. Compact way of describing a family of joint dist. (this lecture)
2. **Enable us to visualize conditional dependencies** (this lecture)
3. Enable us to perform inference using observed data (next lecture)

Examples of Graphical Models

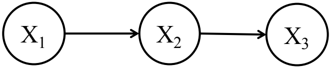
Three Canonical Graphs: (1) Markov Chain

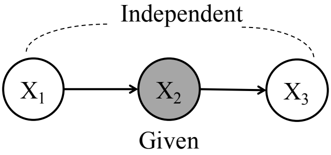


Each state X_i captures all the information about past states that is relevant for the future states

- **Example 1 (Drawing balls from an urn):** You have an urn with 10 balls. Each day, you draw 1 or 2 balls (without replacement). After day i, let X_i be the number of balls remaining in the urn. Observe that X_{i+1} only depends on X_i and the no. of balls pulled out on day i
- **Example 2 (Weather prediction):** Let X_i be the weather on the i-th day, which can be rainy or sunny. Suppose that tomorrow's weather depends on today. But if we know today's weather, then tomorrow's weather is independent of yesterday.

Three Canonical Graphs: (1) Markov Chain





- What is the joint distribution of X_1, X_2, X_3?

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$$

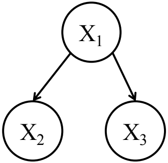
- X_1 and X_3 are conditionally independent given X_2

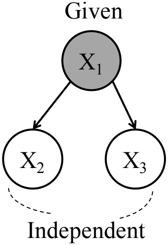
$$p(x_1, x_3|x_2) = p(x_1|x_2)p(x_3|x_2)$$

- Short-form notation of conditional independence $X_1 \perp\!\!\!\perp X_3|X_2$

Three Canonical Graphs: (2) Hidden Cause

- **Example (Shoe size and Gray Hair):** $X_2 =$ Shoe Size and $X_3 =$ Gray Hair, which are correlated in the general population because children have small feet and no gray hair. But, given the hidden cause $X_1 =$ Age, X_2 and X_3 are uncorrelated
- **Example (Icecream Sales and Homicides):** $X_2 =$ Icecream Sales and $X_3 =$ Homicide Rate are positively correlated. Does increased ice-cream consumption make people murderous? No, there is a hidden case $X_1 =$ summer.





Three Canonical Graphs: (2) Hidden Cause

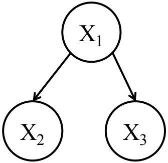
- What is the joint distribution of X_1, X_2, X_3?

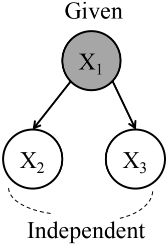
$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1)$$

- X_2 and X_3 are conditionally independent given X_1

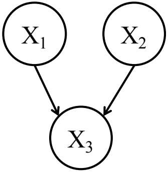
$$p(x_2, x_3|x_1) = p(x_2|x_1)p(x_3|x_1)$$

- Short-form notation of conditional independence $X_2 \perp\!\!\!\perp X_3 | X_1$

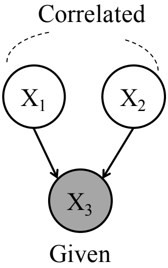




Three Canonical Graphs: (3) Explaining Away



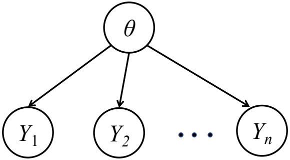
- Suppose we have observed an event which may result from one of two causes. If we then observe one of the causes, this makes the other one less likely i.e., it **explains it away**
- Observing X_3 induces a dependence between X_1 and X_3
- This graph is also called a V-structure.



Examples

- X_1 = Rain, X_2 = Sprinkler, X_3 = Ground wet
- X_1, X_2 Binary, $X_3 = X_1$ XOR X_2

Example: n tosses of a coin with bias θ



Recall from the MLE/MAP lecture

- n coin tosses are conditionally independent given the bias θ. That is why, in MLE we maximized the likelihood:

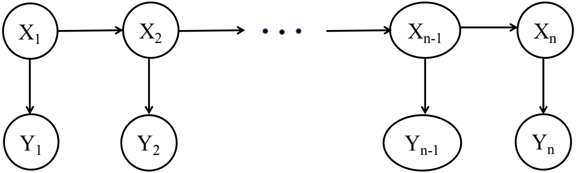
$$p(y_1, y_2, \ldots y_n|\theta) = p(y_1|\theta)p(y_2|\theta)\ldots p(y_n|\theta)$$

- What is the joint distribution of $\theta, y_1, \ldots, y_n$?

$$p(y_1, y_2, \ldots y_n, \theta) = p(\theta)p(y_1|\theta)p(y_2|\theta)\ldots p(y_n|\theta)$$

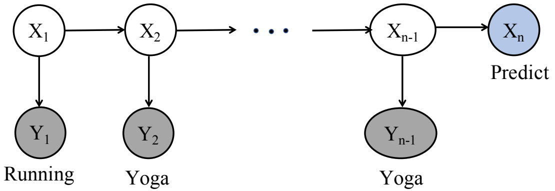
- In MAP we maximize $p(\theta|y_1, y_2, \ldots y_n) \propto p(y_1, y_2, \ldots y_n, \theta)$

Example: Hidden Markov Model (HMM)



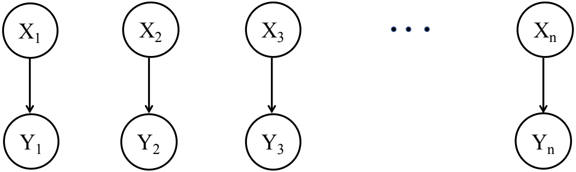
- Eg. Suppose there are two forms of exercise that I do: $Y =$ running (outdoor) or $Y =$ yoga (indoor) each day. My choice is governed by the weather, which can be $X =$ rainy or sunny.
- Given that the weather X_i on the i-th day is rainy, I am more likely to do yoga ($P(Y_i = \text{ yoga}|X_i)$ is larger)
- Tomorrow's weather depends on today's weather

Example: Hidden Markov Model (HMM)



- I tell you my exercise activity each day ($Y_1, Y_2, \ldots Y_{n-1}$ is observed), but don't share the weather ($X_1, X_2, \ldots X_{n-1}$ unknown)
- Can you predict tomorrow's weather X_n?
- More on this (inference on HMMs) in the next class

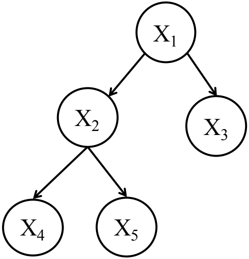
Example: Naive Bayes Model



- n is the number of words in the dictionary
- X_i indicates the probability with which word i occurs in a spam email
- Y_i is the number of times word i is observed in a given email

Example: Tree and Forest

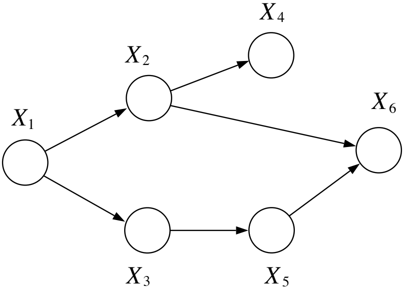
- Tree: Each node (except for the root node) has exactly one parent. An N-node graph will have $N-1$ edges
- Forest: A collection of disjoint trees is called a forest
- What are the size of the conditional distribution tables at each node?



$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_2)$$

Thus, each node corresponds to a 2-dimensional conditional probability distribution table.

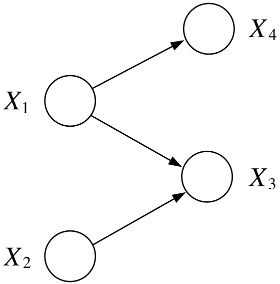
Q: Write the Joint Distribution of this Graphical Model



$$\begin{aligned} & p(x_1, x_2, x_3, x_4, x_5, x_6) \\ & = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5) \end{aligned}$$

Q: Draw the Graphical Model for this Joint Distribution

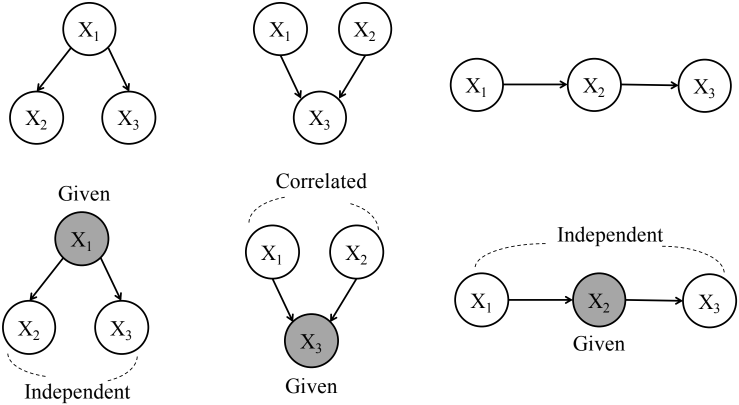
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_1)$$



Bayes Ball Theorem (d-separation)

Finding Variable Dependencies from a Graphical Model

- For the three canonical graphs, we inferred the conditional independences



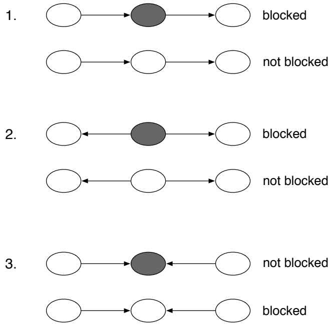
$$X_2 \perp\!\!\!\perp X_3 | X_1 \qquad X_2 \not\perp\!\!\!\perp X_3 | X_1 \qquad X_1 \perp\!\!\!\perp X_3 | X_2$$

- How do you identify these for a general graph?

Bayes Ball Theorem (also called d-separation)

Checking conditional dependencies between two nodes i and j given the observed values of nodes in set $\mathcal{S}$ (can also be an empty set).

- Shade the set of observed nodes $\mathcal{S}$ in grey
- Imagine a ball placed at node i. We want to move it to j
- The ball's movement along each edge is governed by the rules



- If the ball does not reach X_j, then $X_i \perp\!\!\!\perp X_j | X_{\mathcal{S}}$. Else $X_i \not\perp\!\!\!\perp X_j | X_{\mathcal{S}}$

Example of using the Bayes Ball algorithm

Graph 1

- $X_1 \perp\!\!\!\perp X_4 | \{X_2, X_3\}$
- $X_5 \not\perp\!\!\!\perp X_6$

Graph 2

- $X_2 \not\perp\!\!\!\perp X_3 | X_1, X_6$
- $X_2 \not\perp\!\!\!\perp X_4 | X_1, X_6$

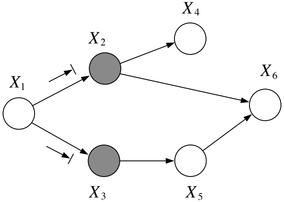


Figure 2.16: A ball cannot pass through X_2 to X_6 nor through X_3.

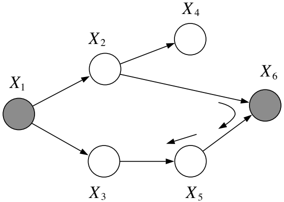
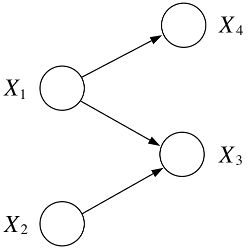


Figure 2.17: A ball can pass from X_2 through X_6 to X_5, and thence to X_3.

Example of using the Bayes Ball algorithm

List of all conditional independencies for this graph



$$X_1 \perp\!\!\!\perp X_2$$

$$X_2 \perp\!\!\!\perp X_4$$

$$X_2 \perp\!\!\!\perp X_4 \mid X_1$$

$$X_3 \perp\!\!\!\perp X_4 \mid X_1$$

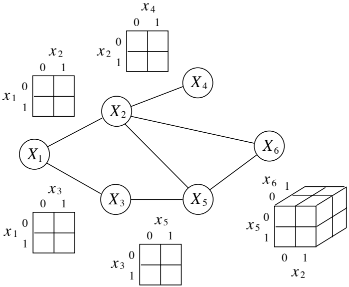
$$X_2 \perp\!\!\!\perp X_4 \mid \{X_1, X_3\}$$

$$\{X_2, X_3\} \perp\!\!\!\perp X_4 \mid X_1$$

Related Concepts

Undirected Graphical Models (also called Markov Random Fields)

- Edges do not have a direction
- Directed models can be converted to undirected models (out of the scope of this class)
- For tree graphs, the directed and undirected models represent the same joint distribution



Connection between Bayesian Networks and Neural Networks

- In Bayesian networks (directed graphs), each node is a random variable and has a meaning associated with it.
- In neural networks, each node (or neuron) is just a computational unit and does not have a specific meaning
- Graphical models help understand conditional independencies between nodes
- Neural networks do not give conditional independence information