CS 461: Machine Learning Principles

Class 19: Nov 11

Bayes Net: Representation of a Joint Probabilistic Density

Instructor: Diana Kim

Bayes Net: Probabilistic Generative Modeling

(1) querying posterior probabilities

& (2) learning a joint density of R.V

& (3) efficient marginalization based on conditional independence among R.Vs

Outline

0. Probabilistic reasoning based on posterior

1. Conditional independence

2. How conditional independence
   enhances computational efficiency in computing posterior

3. How Bayes Net represents a joint density and
   encodes conditional independence

The questions on uncertainty in our daily life?

(1)when a fire alarm rings,
what is the chance of the fire has actually happened?

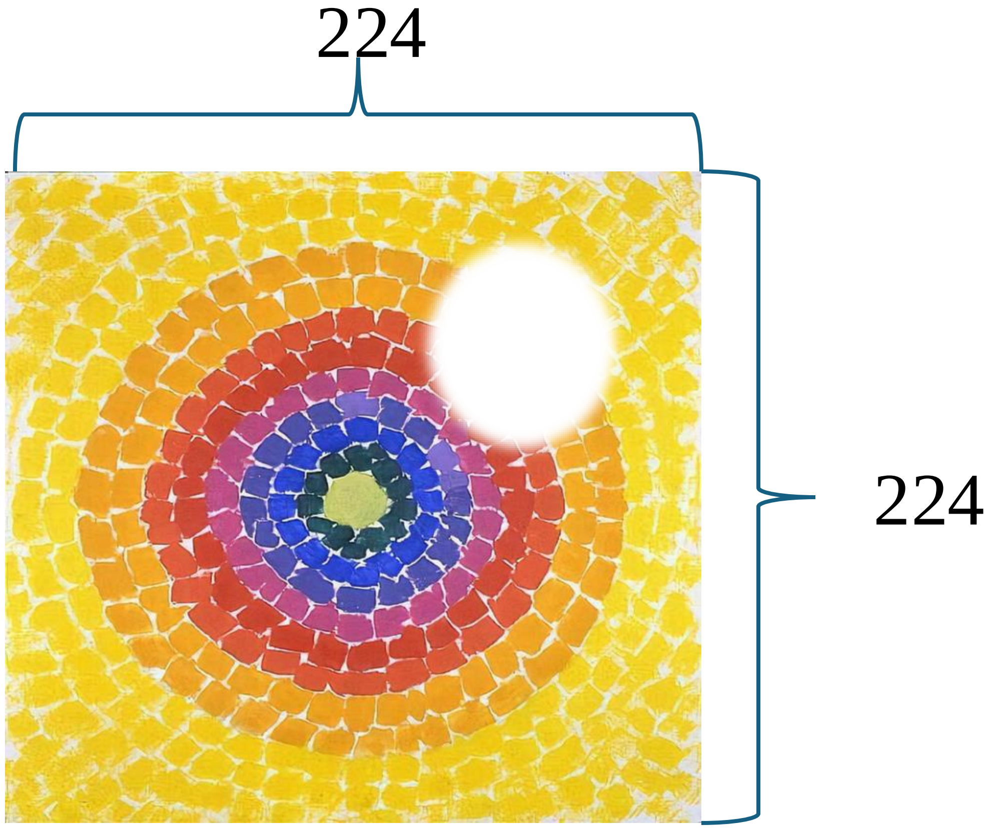
(2) what was the content of the original letter
when we only have a fragment of the paper?

(3) why do we feel better after seeing a doctor? What is the probability for other possible causes to our recovery?
(the medicine I took last night or the natural progression of time?)
The question can be represented by posterior probability.

[One example question that involves the higher dimensional variables]

The data can be encoded by multiple R.Vs.

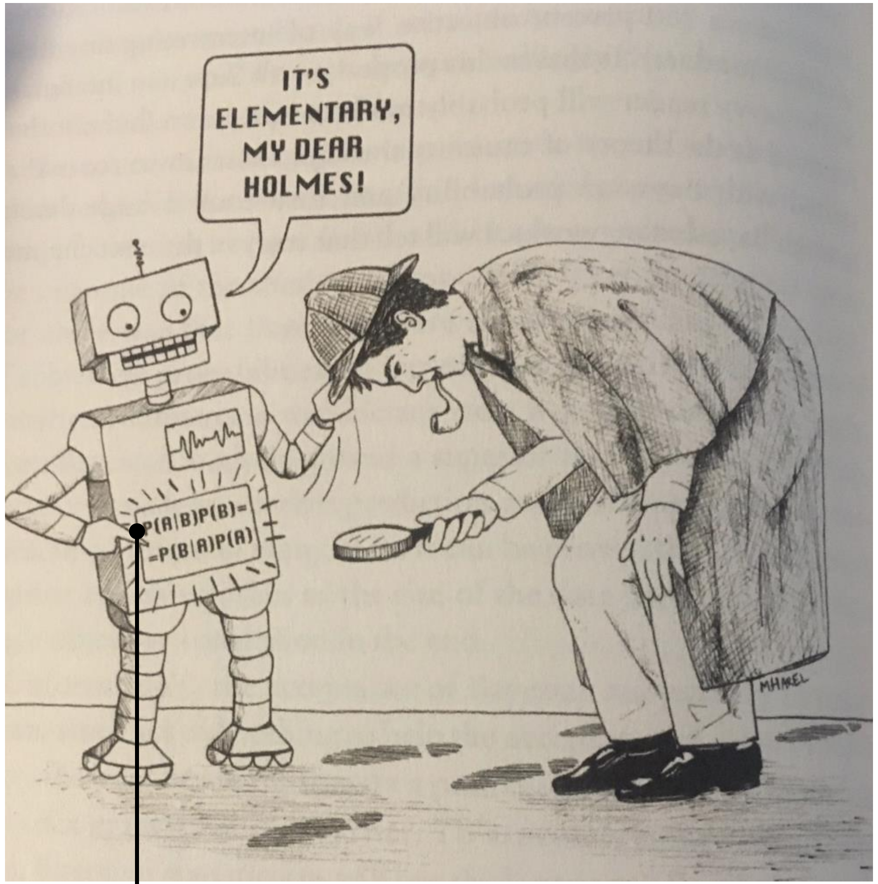
For example of the image, each intensity value is the realization of (224 x 224) random variables.



Q: suppose there are missing values defined by the white circle.
then how can we estimate the white circle? Based on what grounds?

Q: how about measuring $P(X_{circle} \mid X - X_{circle})$?

from "The book of Why" by Judea Pearl



$P[A|B]\ P[B] = P[B|A]\ P[A]$

- The questions can be answered systematically by computing posterior probability based on Bayes Rule.

Posterior computation based on Bayes rule

observations

Q: P(the circle | the intensities in other pixels)

$P(X_{circle} \mid X - X_{circle})$?

Q: $P(X_1|X_2)? = \frac{P(X_1, X_2)}{P(X_2)} = \frac{\sum_{X_3} P(X_1, X_2, X_3)}{\sum_{X_1 X_3} P(X_1, X_2, X_3)}?$

[Bayes Rule]

[**Joint density** and **Marginalization**]

All queries (posterior densities)
can be computed once we have a joint density. (the full of information)
However, the marginalization matters (in terms of computational efficiency);
the computation can be intractable as the number of variables grow.

Example) $a\ R.V\ sequence\ X_1^N$ but as we do not know the conditional independence relations among the R.Vs.

$$P(X_{36}) = \sum_{X_1}\sum_{X_2}\cdots\sum_{X_{35}}\sum_{X_{37}}\cdots\sum_{X_{1000}} P(X_1, X_2, ..., X_{35}, X_{36}, X_{37}, ..., X_{1000})$$

The marginalization involves the summation over K^{999} terms, where $X_i\ has\ K$ states. $O(K^N)$ operations.

Q: Suppose some conditional independence is known.
how we can use it to reduce the complexity ?

Conditional independence
enhances computational efficiency in computing posterior.

- Independence

Discrete Random Variables X and Y are independent $\leftrightarrow$
the joint PMF (probability Mass Function) is the product of the marginal PMFs.

$$\begin{aligned} P(X=x, Y=y) &= P(X=x) \cdot P(Y=y|X=x) \\ &= P(Y=y) \cdot P(X=x|Y=y) \\ &= P(Y=y) \cdot P(X=x) \quad \forall x, y \end{aligned}$$

- Conditional Independence

Discrete random variables X and Y are conditionally independent given Z

$\leftrightarrow$

the conditional join PMF is the product of the conditional marginal PMFs.

$$
\begin{aligned}
P(X=x,Y=y|Z=z) &= P(X=x|Z=z)\cdot P(Y=y|X=x,Z=z) \\
&= P(Y=y|Z=z)\cdot P(X=x|Y=y,Z=z) \\
&= P(Y=y|Z=z)\cdot P(X=x|Z=z) \quad \forall x,y,z
\end{aligned}
$$

Assume the random variables are <u>conditionally independent.</u>
Ex) $a\ R.V\ sequence\ X_1^N.\ Each\ X_i\ is\ independent\ to\ X_1^{i-2}$ given X_{i-1}

$$P(X_{36}) = \sum_{X_1}\sum_{X_2}...\sum_{X_{35}}\sum_{X_{37}}...\sum_{X_{1000}} P(X_1)P(X_2|X_1),...,P(X_{36}|X_{35})P(X_{37}|X_{36}),....P(X_{1000}|X_{999})$$

$$= \sum_{X_1} P(X_1)\sum_{X_2} P(X_2|X_1)...\sum_{X_{35}} P(X_{35}|X_{34})P(X_{36}|X_{35})\sum_{X_{37}} P(X_{37}|X_{36})...\sum_{X_{1000}} P(X_{1000}|X_{999})$$

How the conditional independence enables to solve marginalization efficiently?

EX) A Simple Case

X_i is a binary R.V and and $[X_i \perp X_1^{i-2} \mid X_{i-1}]$

$$P[X_0] = \sum_{X_1} \sum_{X_2} P(X_0, X_1, X_2)$$

$$= P(X_0) \sum_{X_1} P(X_1|X_0) \underbrace{\sum_{X_2} \underbrace{P(X_2|X_1)}_{f_1(X_1, X_2)}}_{f_2(X_1)}$$

[From factor $f_1(X_1, X_2)$ to $f_2(X_1)$]

X_1 \ X_2	$+$	$-$
$+$		
$-$		

$\longrightarrow$

X_1	
$+$	
$-$	

EX) A Simple Case

X_i is a binary R.V and $[X_i \perp X_1^{i-2} \mid X_{i-1}]$

$$f_3(X_0) = \sum_{X_1} P(X_1|X_0) f_2(X_1)$$
$$= P(X_1 = +|X_0) f_2(X_1 = +) + P(X_1 = -|X_0) f_2(X_1 = -)$$

$$P[X_0] = \sum_{X_1} \sum_{X_2} P(X_0, X_1, X_2)$$
$$= P(X_0) \underbrace{\sum_{X_1} P(X_1|X_0) \underbrace{\sum_{X_2} \underbrace{P(X_2|X_1)}_{f_1(X_1, X_2)}}_{f_2(X_1)}}_{f_3(X_0)}$$

$X_0 \backslash X_1$	$+$	$-$
$+$		
$-$		

$\longrightarrow$ [forming $f_3(X_0)$]

X_0	
$+$	
$-$	

EX) A Simple Case

X_i is a binary R.V and $[X_i \perp X_1^{i-2} \mid X_{i-1}]$

$$P[X_0] = \sum_{X_1}\sum_{X_2} P(X_0, X_1, X_2)$$

$$= \underbrace{P(X_0)\underbrace{\sum_{X_1} P(X_1|X_0)\underbrace{\sum_{X_2}\underbrace{P(X_2|X_1)}_{f_1(X_1, X_2)}}_{f_2(X_1)}}_{f_3(X_0)}}_{f_4(X_0)}$$

Every factor contains $2^2 = 4$ terms at most and the factors are processed <u>sequentially</u> from the right to left order. Hence, the total # of operation will be $N \times 4$. $O(2^2)$ operations.

$$P(X_{36}) = \sum_{X_1}\sum_{X_2}\cdots\sum_{X_{35}}\sum_{X_{37}}\cdots\sum_{X_{1000}} P(X_1, X_2, ..., X_{35}, X_{36}, X_{37}, ..., X_{1000})$$

The marginalization involves the summation over 2^{999} terms, where X_i has K states and N variables. $O(2^{999})$ operations.

$$P(X_{36}) = \sum_{X_1}\sum_{X_2}\cdots\sum_{X_{35}}\sum_{X_{37}}\cdots\sum_{X_{1000}} P(X_1)P(X_2|X_1), ..., P(X_{36}|X_{35})P(X_{37}|X_{36}), ...P(X_{1000}|X_{999})$$

$$= \sum_{X_1} P(X_1)\sum_{X_2} P(X_2|X_1)...\sum_{X_{35}} P(X_{35}|X_{34})P(X_{36}|X_{35})\sum_{X_{37}} P(X_{37}|X_{36})...\sum_{X_{1000}} P(X_{1000}|X_{999})$$

Every factor contains $2^2 = 4$ terms at most and the factors are processed <u>sequentially</u> in the right to left order. Hence, the total # of operation will be N $\times$ 4. $O(2^2)$ operations.

Bayes Net encodes a joint distribution with <u>its conditional independence.</u>
It provides a framework enabling to compute queries (posterior prob)
in a reasonable amount of time.

Bayes Net (Graphical Representation)

$\leftrightarrow$

a Joint Probabilistic Density (factorized)

Bayes Net directly represents a joint probabilistic density
by using DAG (**D**irected **A**cyclic **G**raph)
: nodes encode random variables and edges define conditions.

Suppose we three random variable X_1, X_2, X_3.
The same joint density $P(X_1, X_2, X_3)$ can be represented
by $3 \times 2 \times 1 = 6$ possibles ways but all are the same $P(X_1, X_2, X_3)$.

- $P\ (X_1)\ P(X_2 |\ X_1)\ P(X_3 |\ X_1, X_2)$
- $P\ (X_1)\ P(X_3 |\ X_1)\ P(X_2 |\ X_1, X_3)$
- ...

DAG representations

Example 1) a density can be factorized in different orders.

The factorizations in different orders result in different DAG.

- $P(X_1)\, P(X_2 | X_1)\, P(X_3 | X_1, X_2)$

- $P(X_1)\, P(X_3 | X_1)\, P(X_2 | X_1, X_3)$

They encode the same joint density.

a single joint density can have multiple DAG representations.

DAG representations

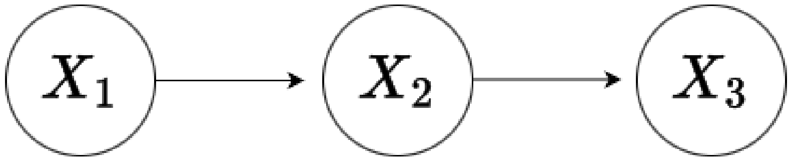
Example 2) $[X_3 \perp X_1 \ | X_2]$: X_3 and X_1 are conditionally independent given X_2.

- $P\ (X_1)\, P(X_2 | X_1)\, P(X_3 | X_2)$

$[X_3 \perp X_1 \mid X_2]$: X_3 and X_1 are conditionally independent given X_2.

The lack of edges indicates conditional independence.

- $P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_2)$



Baye Net Examples

Bayes Net defines a structure and parameters

- structure: a DAG encodes a joint density and conditional independence.
- parameters: the conditional densities (CPT: Conditional Probability Table)

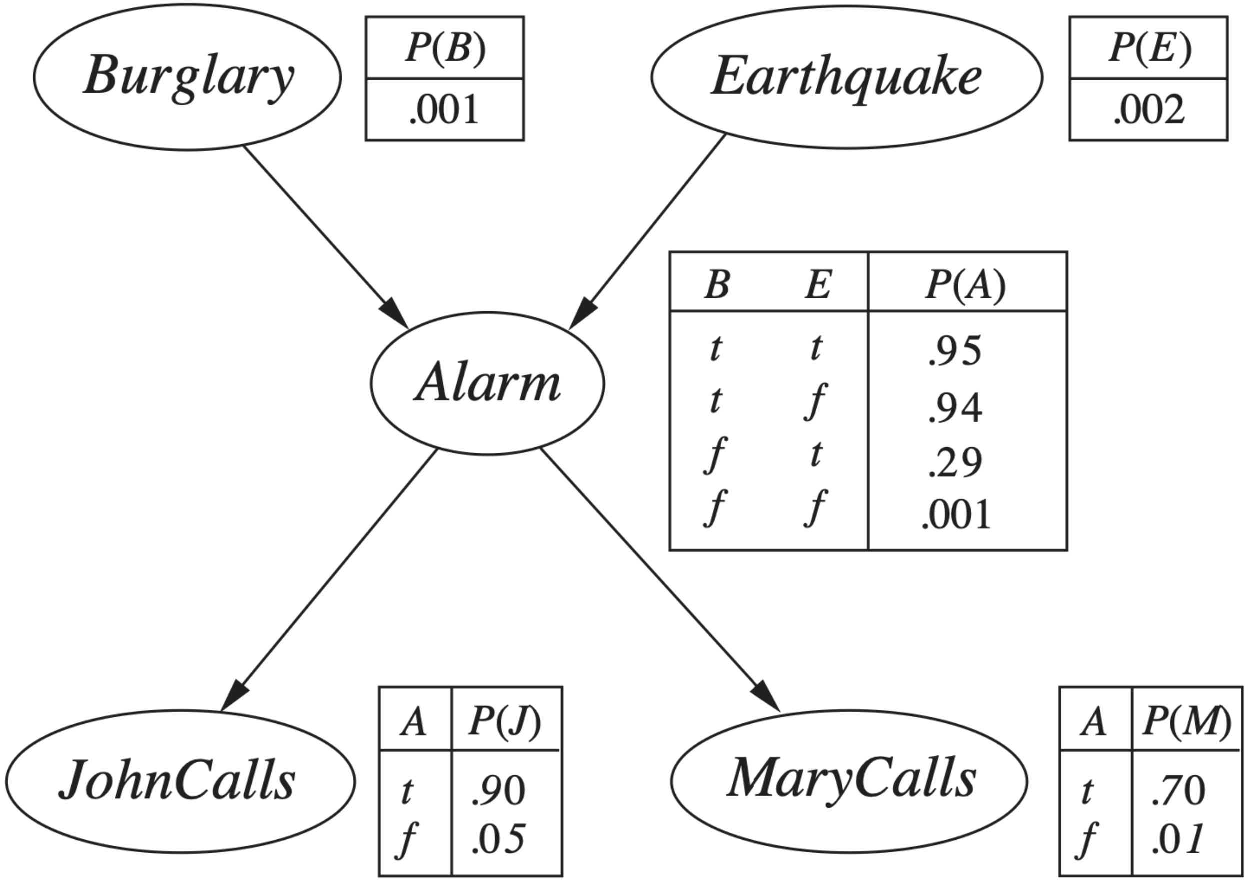


Figure 14.2 From the book “AI: A Modern Approach”

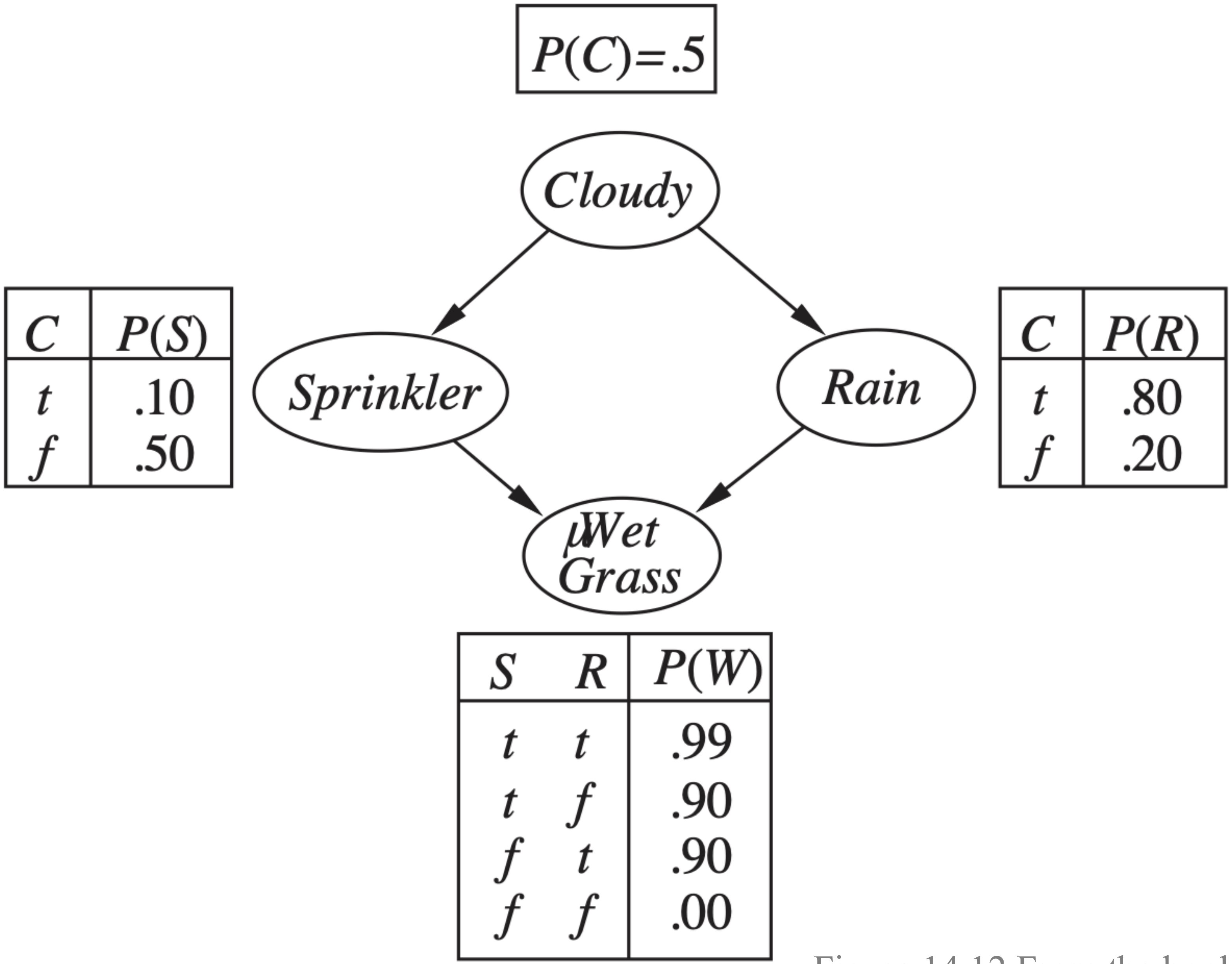


Figure 14.12 From the book "AI: A Modern Approach"

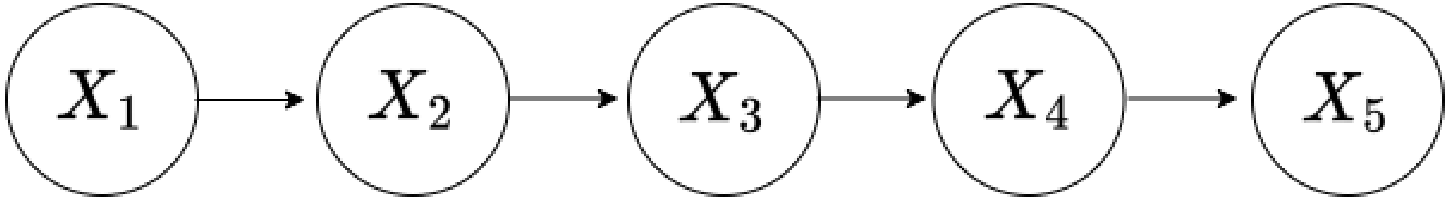
How can we learn the structure?

- We can start from a preset structure based on prior domain knowledge
- We can learn structure from data, too.

We can find a set of conditional independence that a Bayes Net encodes explicitly / implicitly.

This graph directly encodes the conditional densities.

- $[X_i \perp X_1^{i-2} \mid X_{i-1}]$



[Markov Chain]

Q: How about $[X_1 \perp X_5 \mid X_3]$?

$$\begin{aligned}
P(X_1, X_5|X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, ...X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, X_3) P(X_4, X_5|X_1, X_2, X_3)}{P(X_3)} \\
&= \sum_{X_2, X_4} P(X_1, X_2|X_3) P(X_4, X_5|X_3) \\
&= P(X_1|X_3) P(X_5|X_3)
\end{aligned}$$

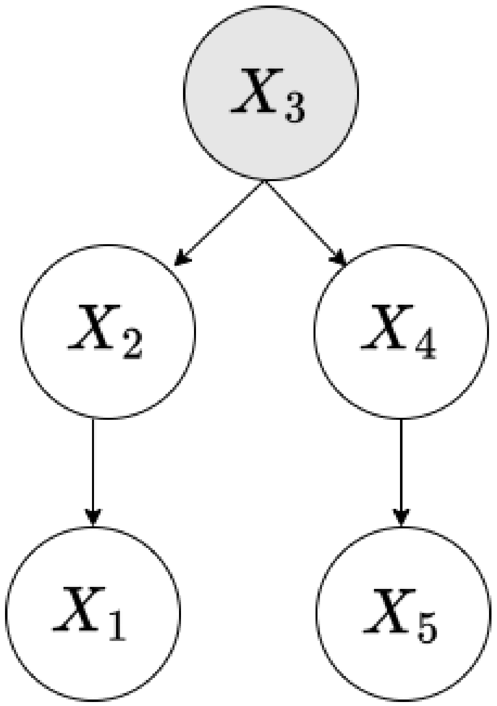
Some Rules to Identify the Conditional Independence in Bayes Net
[1] Chain rule
[2] Tent rule
[3] V-structure rule

[1] Chain Structure: the middle node of chain breaks into two.

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow X_5$$

$$\begin{aligned}
P(X_1, X_5|X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, ...X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, X_3)P(X_4, X_5|X_1, X_2, X_3)}{P(X_3)} \\
&= \sum_{X_2, X_4} P(X_1, X_2|X_3)P(X_4, X_5|X_3) \\
&= P(X_1|X_3)P(X_5|X_3)
\end{aligned}$$

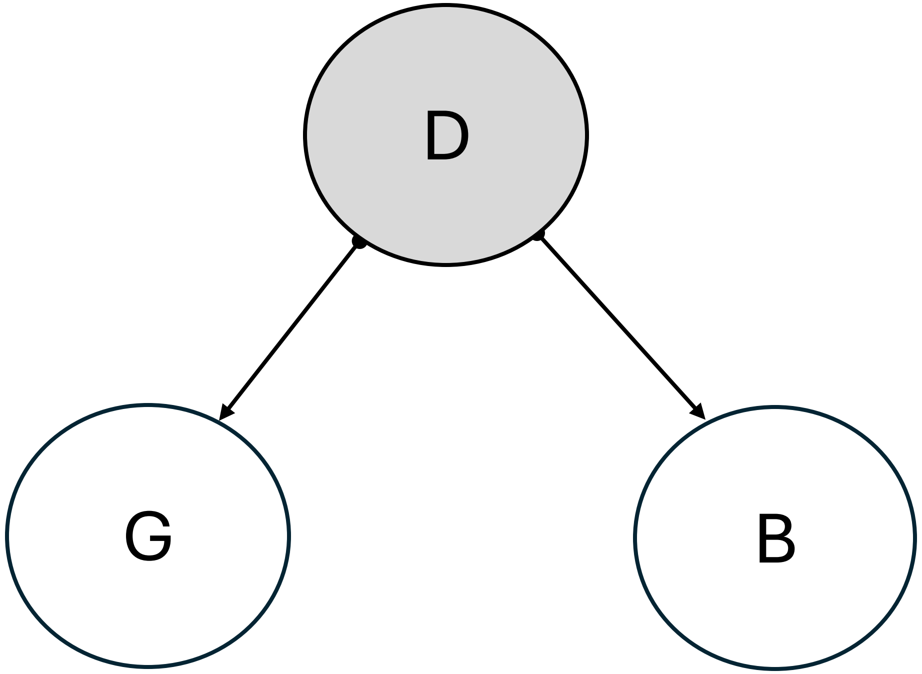
[2] Tent Structure: root note separates its children



$$
\begin{aligned}
P(X_1, X_5 | X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, ...X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, X_3) P(X_4, X_5 | X_1, X_2, X_3)}{P(X_3)} \\
&= \sum_{X_2, X_4} P(X_1, X_2 | X_3) P(X_4, X_5 | X_3) \\
&= P(X_1 | X_3) P(X_5 | X_3)
\end{aligned}
$$

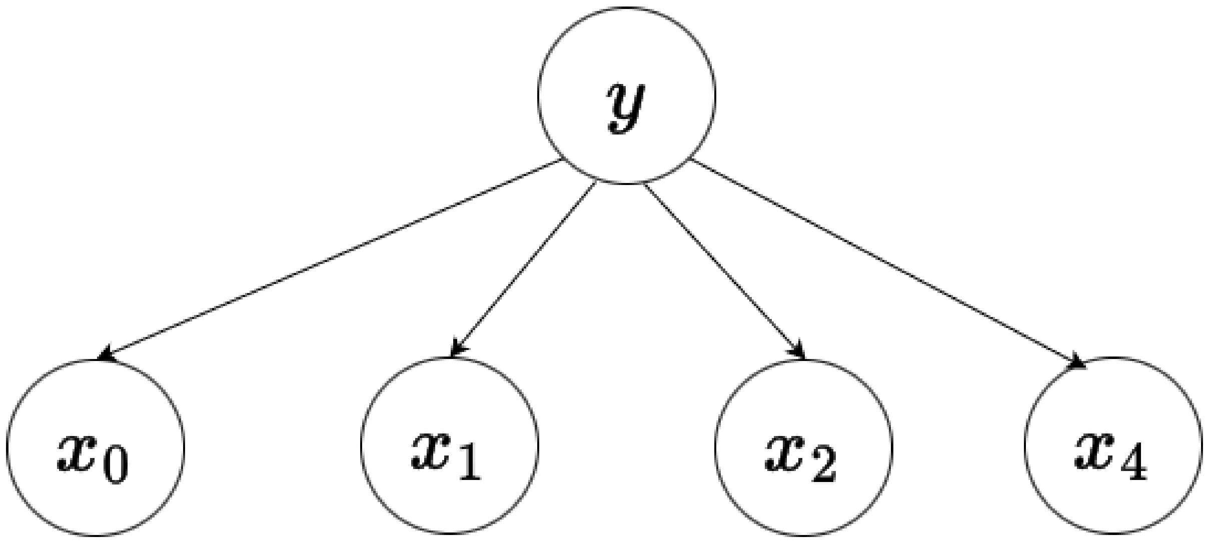
[Example of Tent Structure]

In the hw1, the feature of Glucose (G) and blood pressure (B) were conditionally independent given Diabetes (D). The conditional independence is based on the assumption of causal relationship between D and the feature of G and B.

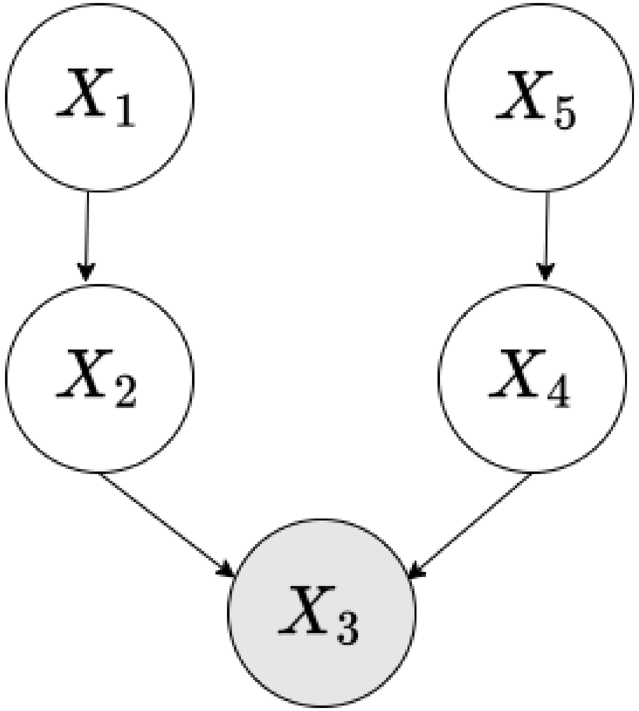


- When D information observed, G and B becomes independent.
  $P(G, B|D) = P(G|D)\,P(B|D)$

Conditional Independence Assumption for Naïve Bayes



[3] V-Structure: conditioning on a common child at the bottom of a v-structure makes its parents become dependent.



$$
\begin{aligned}
P(X_1, X_5 | X_3) &= \frac{P(X_1, X_3, X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2, \ldots X_5)}{P(X_3)} \\
&= \frac{\sum_{X_2, X_4} P(X_1, X_2) P(X_4, X_5) P(X_3 | X_2, X_4)}{P(X_3)} \\
&\neq P(X_1 | X_3) P(X_5 | X_3)
\end{aligned}
$$

[D separation for a path]

- An undirected path P is **d-separated** by a set of observation nodes E iff at least one of the conditions hold.

  (1) it contains a **chain structure.**
  (2) it contains a **tent structure.**
  (3) it contains a **V-structure.**

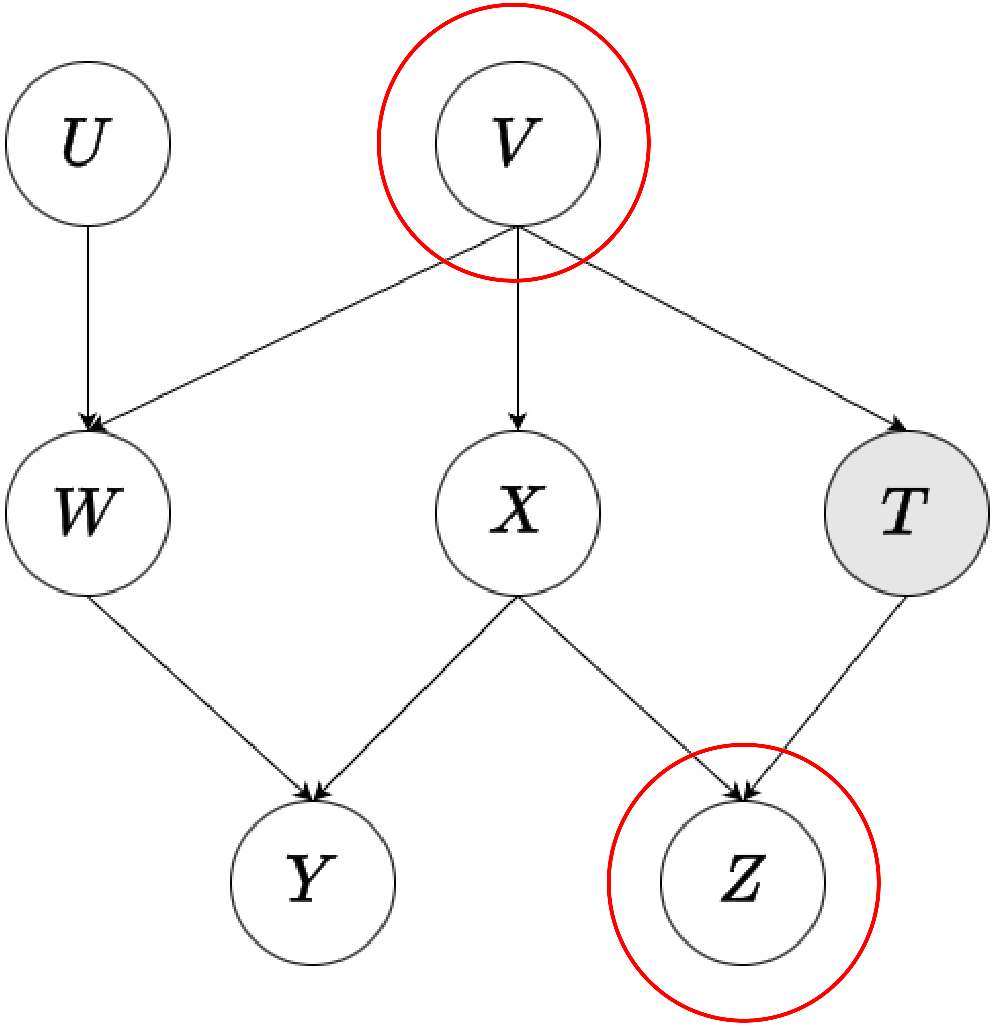
[D separation between two nodes]

The node A is **d-separate**d from B by a set of observation E
$\leftrightarrow$ each undirected path from A to B is d-separated.

$$\leftrightarrow [A \perp B \,|E]$$

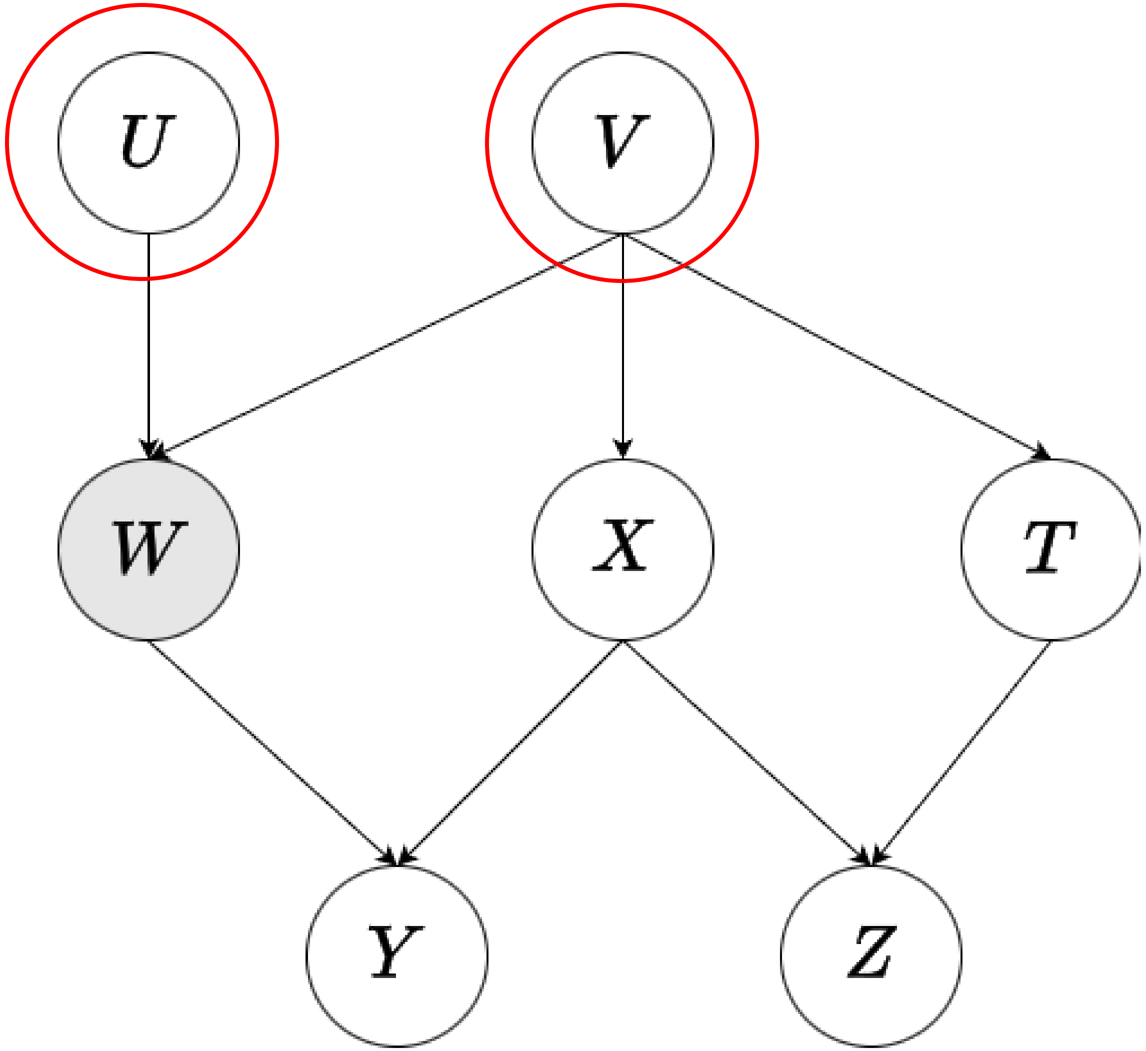
Q: What if we find a path between A and B that is not d-separated?

Ex1: D-Separation)



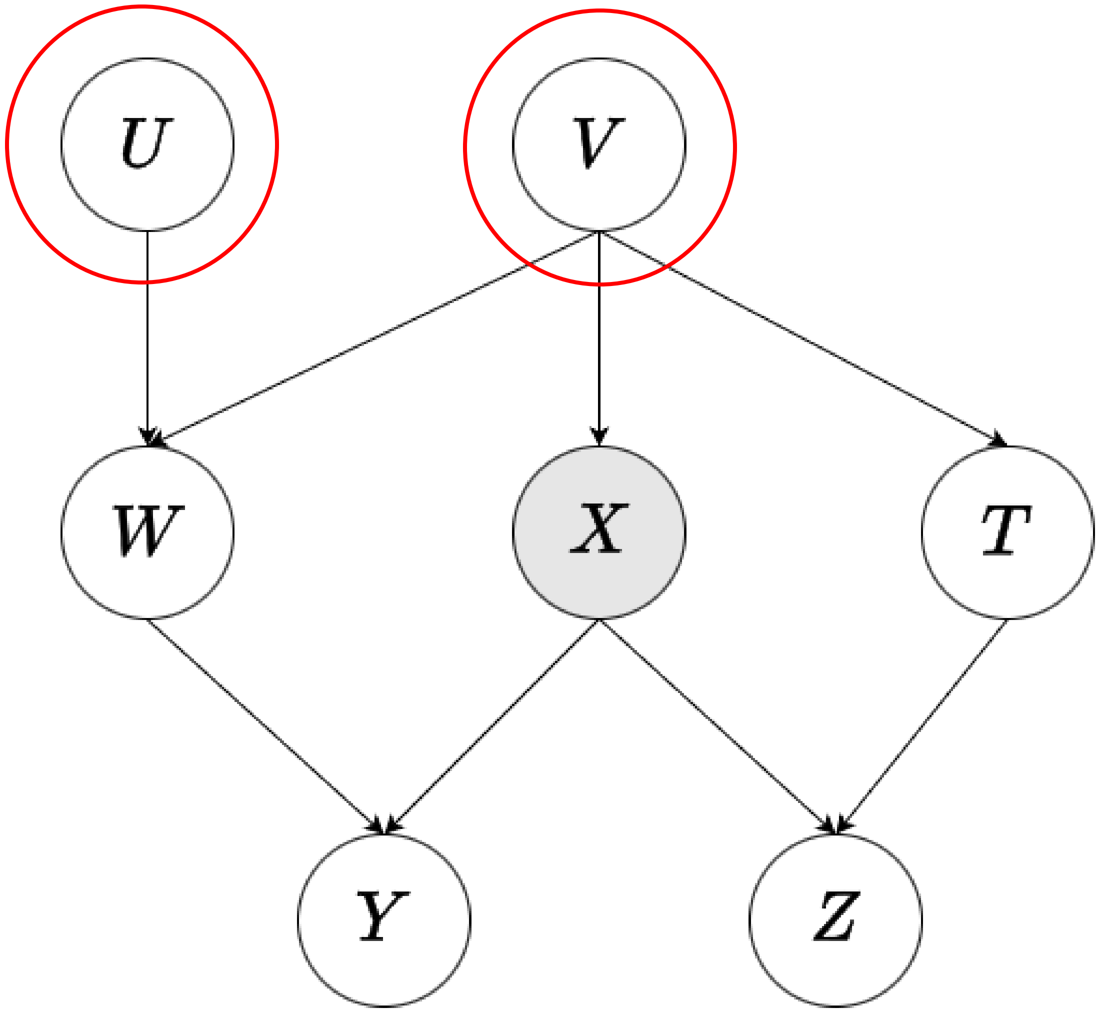
Q: $[V \perp Z | T]$?

Ex2: D-Separation)



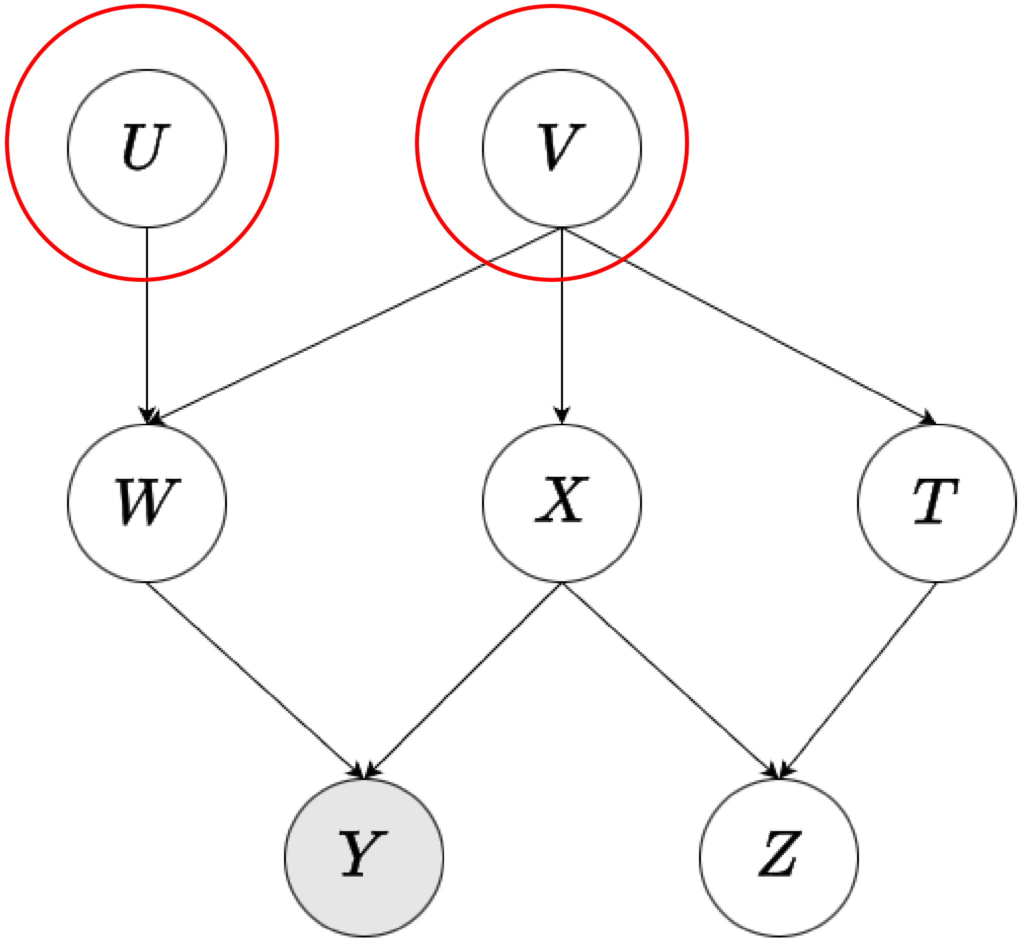
Q: $[U \perp V | X]$?

Ex3: D-Separation)



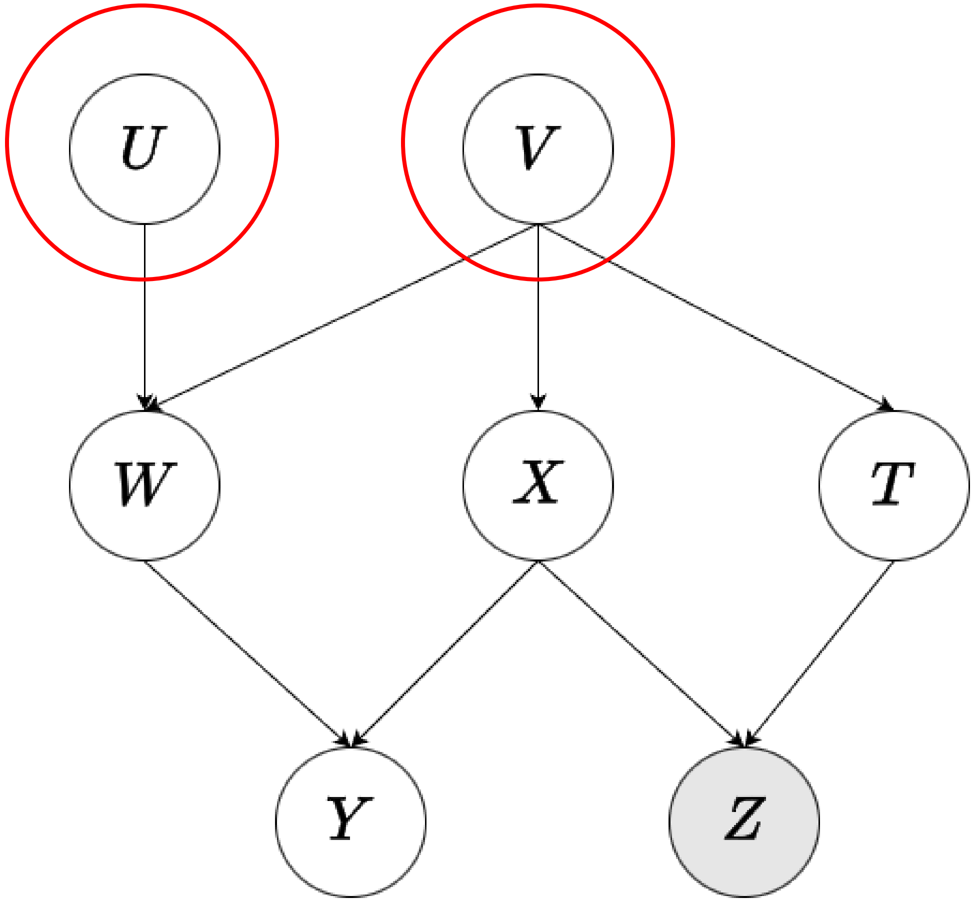
Q: $[U \perp V | X]$?

Ex4: D-Separation)



Q: $[U \perp V | Y]$?

Ex5: D-Separation)



Q: $[U \perp V | Z]$?