Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 4: Conditional Independence and Directed Graphical Models

January 30, 2025

Reading: See course homepage

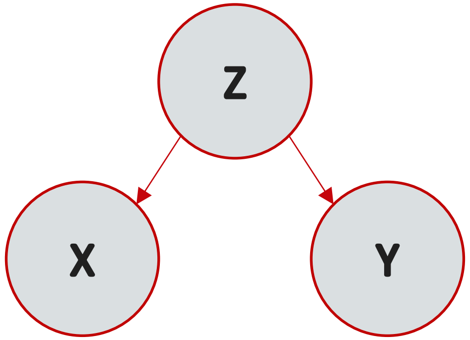
Today

- Conditional Independence
- Directed Graphical Models
  - Markov Chains
  - Hidden Markov Models
  - Bayesian Networks

Conditional Independence

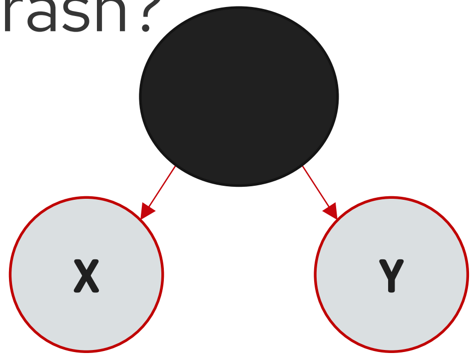
Introduction to Conditional Independence

- Variables X and Y are **independent** if:
$$P(X,Y) = P(X)P(Y)$$
  - Notation: $X \perp Y$
- Variables X and Y are **conditionally independent given Z** if:
$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$
  - Equivalently: $P(X|Y,Z) = P(X,Z)$
  - Notation: $X \perp Y \mid Z$

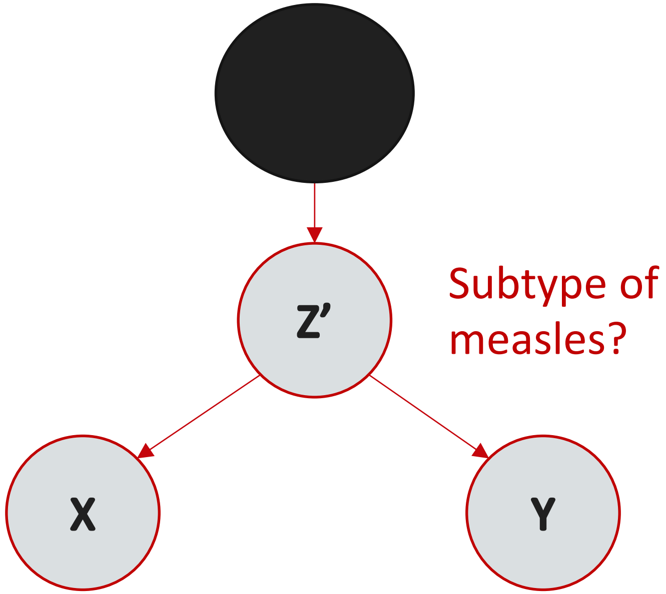


Example of Conditional Independence

- Let X = Fever, Y = Rash, Z = Measles
- Given that a patient has measles, does knowing if they have a fever give us any additional information about whether they have a rash?



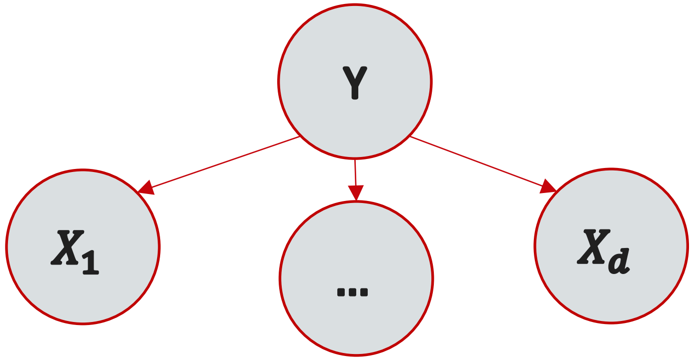
$$P(X,Y|Z) = \frac{P(X,Y,Z)}{P(Z)}$$
$$= \frac{P(X|Z)P(Y|Z)P(Z)}{P(Z)}$$
$$= P(X \mid Z)P(Y \mid Z)$$



$$P(X,Y|Z) = \sum_{Z'} P(Z' \mid Z)P(X \mid Z')P(Y \mid Z')$$

Recall Naïve Bayes

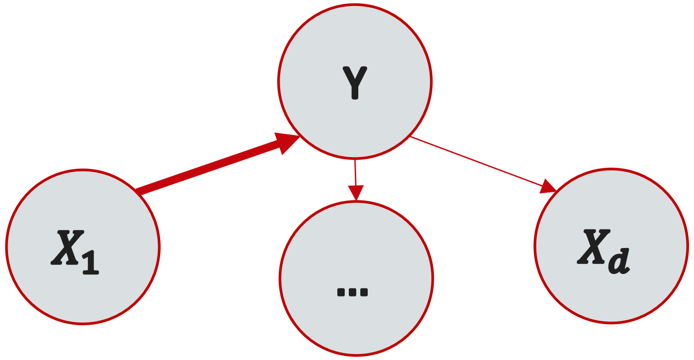
- Conditional independence of X_is | Y allows for efficient computation of $P(X \mid Y)$:



$$P(X|Y) = P(X_1, \dots, X_d|Y)$$
$$= \frac{\prod_i P(X_i \mid Y)\, P(Y)}{P(Y)}$$
$$= \prod_i P(X_i \mid Y)$$

Recall Naïve Bayes

- Could we switch the direction of **one** of the arrows?



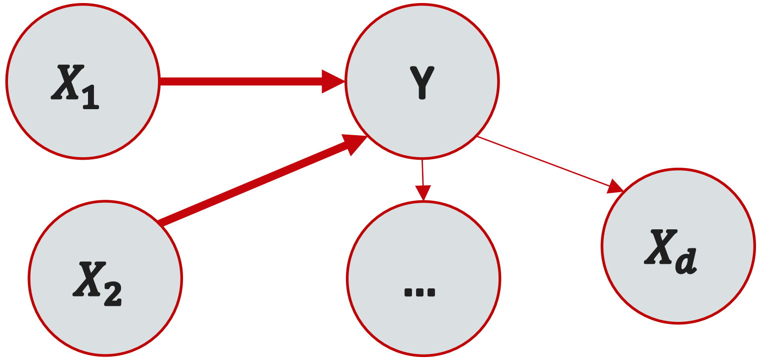
$$P(X|Y) = P(X_1, \dots, X_d|Y)$$

$$= \frac{P(X_1)P(Y|X_1)\prod_{i=2} P(X_i \mid Y)}{P(Y)}$$

$$= \prod_i P(X_i \mid Y)$$

Recall Naïve Bayes

- Could we switch the direction of **two** of the arrows?



$$P(X|Y) = P(X_1, \dots, X_d|Y)$$

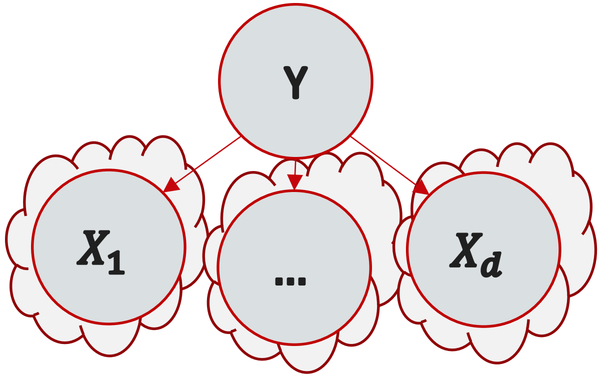
$$= \frac{P(X_1)P(X_2)P(Y|X_1,X_2)\prod_{i=3} P(X_i \mid Y)}{P(Y)}$$

Now we need $X_1 \perp X_2$

$$= \frac{P(X_1)P(X_2)}{P(X_1,X_2)} P(X_1, X_2 \mid Y) \prod_{i=3} P(X_i \mid Y)$$

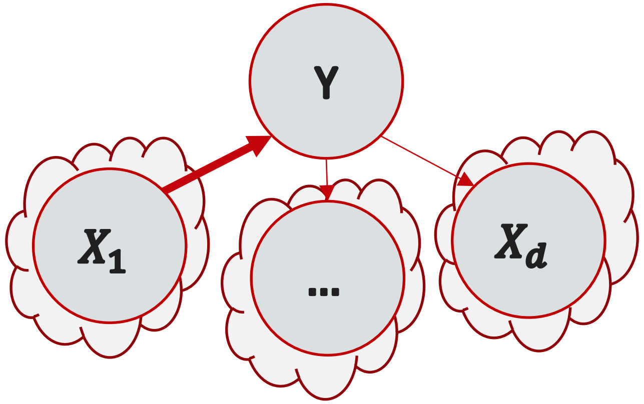
What happened?

Naïve Bayes



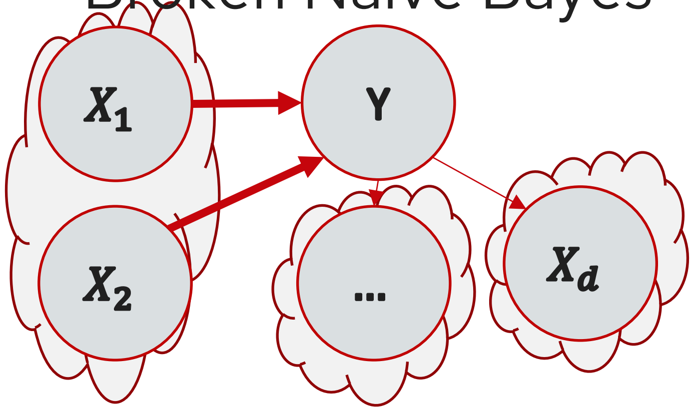
$$P(X|Y) = \prod_i P(X_i \mid Y)$$

Modified Naïve Bayes



$$P(X|Y) = \prod_i P(X_i \mid Y)$$

Broken Naïve Bayes



$$P(X|Y) = \frac{P(X_1)P(X_2)}{P(X_1, X_2)} P(X_1, X_2 \mid Y) \prod_{i=3} P(X_i \mid Y)$$

Intuitively: Ignoring graph structure can **double-count evidence.**

Questions about Conditional Independence?

Directed Graphical Models: Bayesian Networks

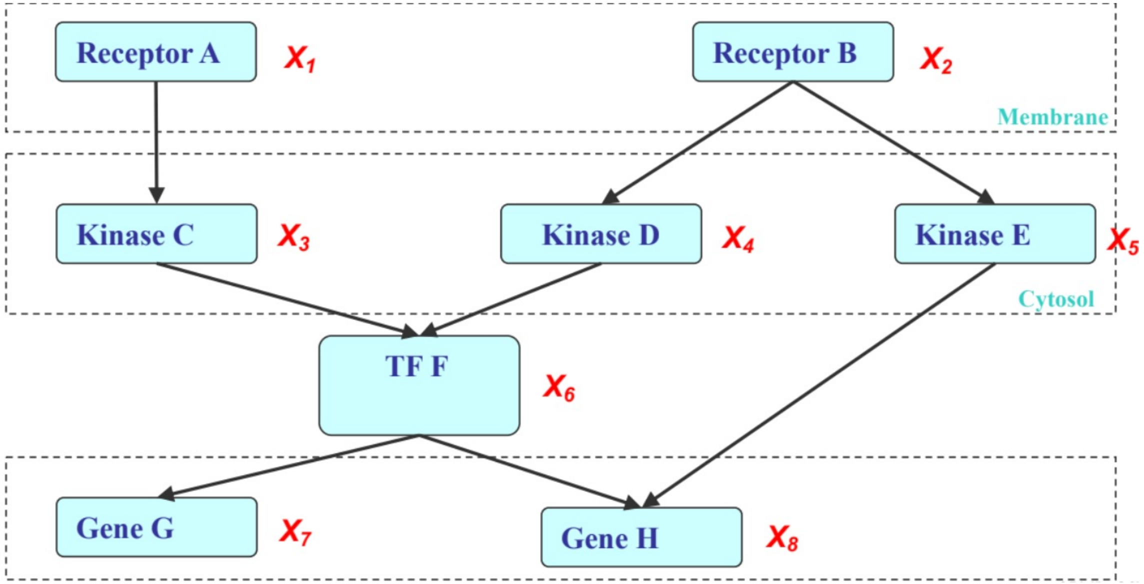
Two types of Graphical Models

- Directed edges give causality relationships (e.g. Bayesian Network)

- Undirected edges give correlations between variables (e.g. Markov Random Field)

Representing Multivariate Distributions

- If $X_i s$ are conditionally independent, the joint can be factored to a product of simpler terms, e.g.



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2) P(X_3|X_1)P(X_4|X_2)P(X_5|X_2) P(X_6|X_3, X_4)P(X_7|X_6)P(X_8|X_6, X_5)$$

- Special case: If $X_i s$ are independent: $P(X_i \mid \cdot) = P(X_i)$

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)P(X_6)P(X_7)P(X_8)$$

Example: The Dishonest Casino

- Suppose a casino has two dice:
  - Fair dice: P(1) = P(2) = … = P(6) = 1/6
  - Loaded dice: P(1) = P(2) = P(3) = P(4) = P(5) = 1/10, **P(6) = ½**
- Suppose the dealer switches between die every 20 times
- Game:
  - You bet $1
  - You roll
  - Dealer rolls (maybe with fair dice, maybe with loaded dice)
  - Highest number wins $2

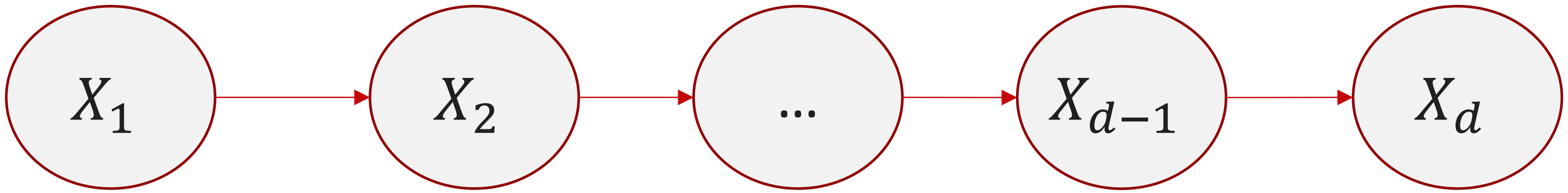
Do you play at the dishonest casino?

Fundamental Questions at the dishonest casino

- Representation
  - Can we build a model of how this game works?
- Learning
  - Can we learn how "loaded" is the loaded dice? How often does the dealer change from fair to loaded and back?
- Inference
  - After observing a sequence of rolls, can we say what portion of the sequence was generated with a fair die vs a loaded one? How likely are we to sit at the table?

A Simple Directed PGM

- Markov Chain
- **Markov property:** “The future state depends only on the present state, and not on past states”
- Parameters:
  - Transition Probability Matrix: $M_{ij} = P(X_t = j \mid X_{t-1} = i)$
  - Initial State Distribution: $\pi_i = P(X_1 = i)$



$$\mathrm{P(X)} = P(X_1) \prod_{t=2} P(X_t \mid X_{t-1})$$

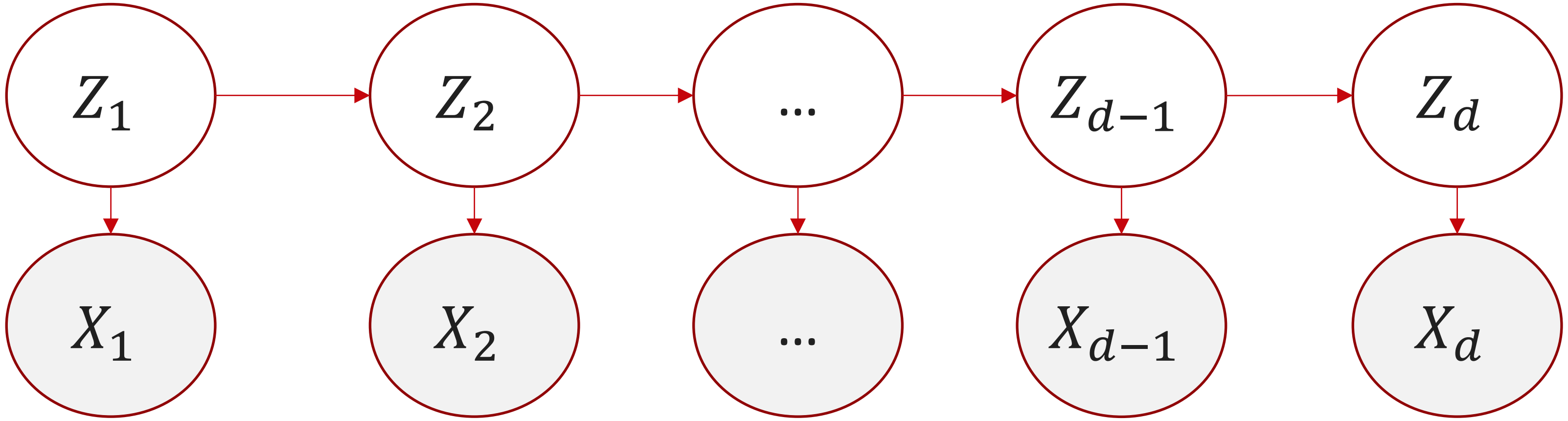
Hidden Markov Model (HMM)

- Markov chain but **underlying drivers not observed**
- Parameters:

  Observation ("Emission") Probability $\mathrm{E_{kj}} = P(X_t = k \mid Z_t = j)$

  Transition Probability Matrix: $M_{ij} = P(Z_t = j \mid Z_{t-1} = i)$

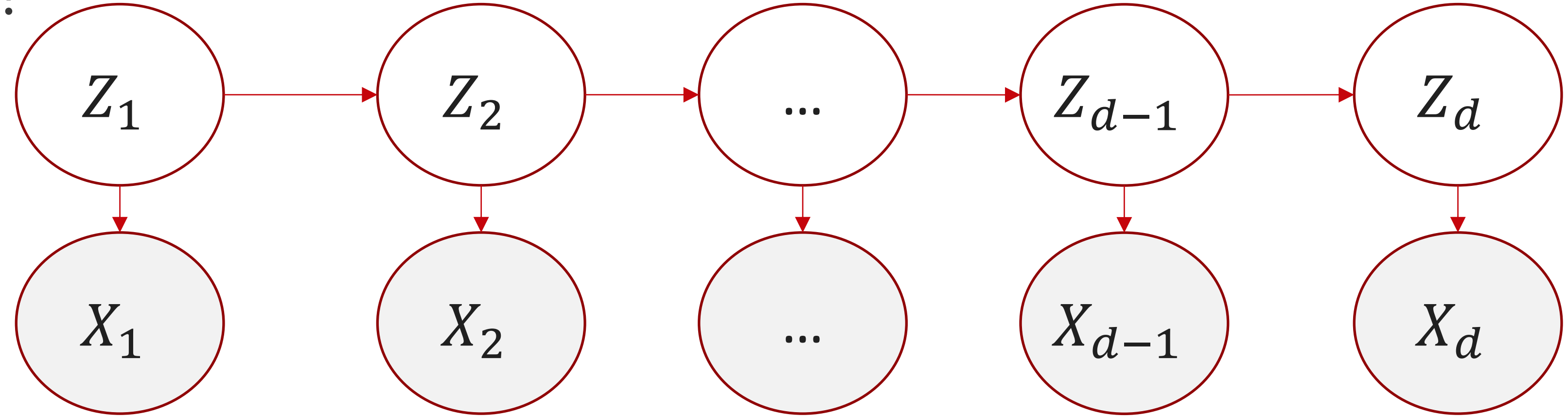
  Initial State Distribution: $\pi_i = P(Z_1 = i)$



$$\mathrm{P(X, Z)} = P(Z_1) \prod_{t=2} P(Z_t \mid Z_{t-1}) \prod_{t'} P(X_t \mid Z_t)$$

Dishonest Casino as HMM

- Z_t: Dice being used by dealer (fair or loaded)
- Observation Probability Matrix: Probability of dice roll, given Z_t
- Transition Probability Matrix: How often dealer switches die.
- Initial State Distribution: What do we believe the dealer started with?



$$\mathrm{P(X, Z)} = P(Z_1) \prod_{t=2} P(Z_t \mid Z_{t-1}) \prod_{t\prime} P(X_t \mid Z_t)$$

Bayesian Network (BN)

- A BN is a **directed acyclic graph** whose nodes represent the random variables and whose edges represent direct influence of one variable on another
- Provides the skeleton for representing a joint distribution compactly in a **factorized** way
- Compact representation of a set of **conditional independence** assumptions
- We can view the graph as encoding a **generative sampling process** executed by nature.

Bayesian Network (BN)

Factorization Theorem:

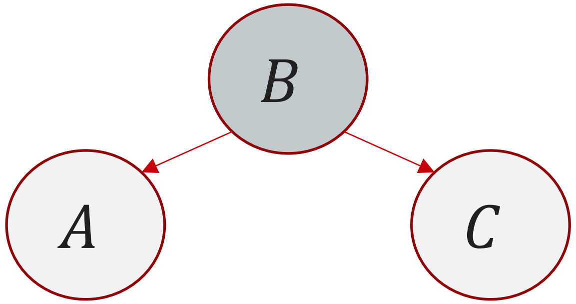
Given a DAG, the most general form of the probability distribution that is consistent with the graph factors according to:

$$P(X) = \prod_i P(X_i \mid X_{\pi_i})$$

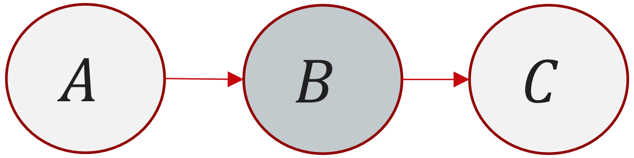
where X_{π_i} is the set of parents of X_i.

Bayesian Network: Local Structures

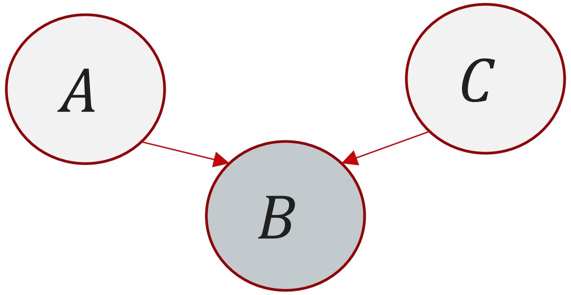
- Common parent
  - Knowing B **decouples** A and C
  - $A \perp C \mid B$



- Cascade
  - Knowing B **decouples** A and C
  - $A \perp C \mid B$



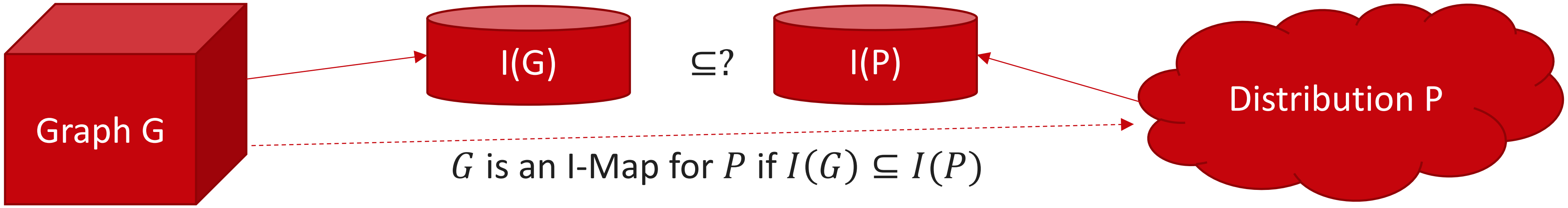
- V-structure
  - Knowing B **couples** A and C
  - A can “explain away” C



Three foundational building blocks for creating complex BNs

I-Maps

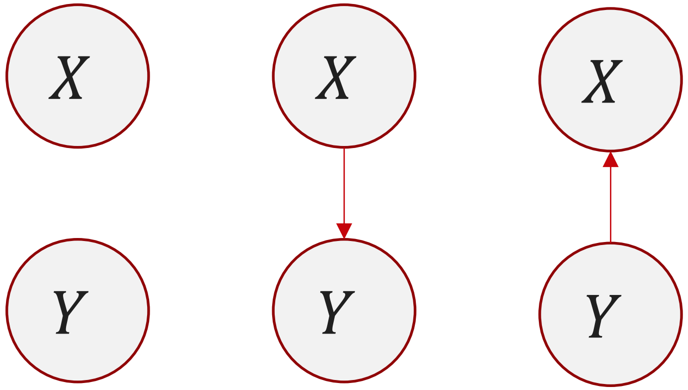
- Independence set: Let P be a distribution over X. We define $I(P)$ to be the set of independences $(X \perp Y \mid Z)$ that hold in P.
- I-Map: Let G be any graph object with an associated independence set $I(G)$. We say that G is an **I-map** for an independence set I if $I(G) \subseteq I$.
- I-Map of Distribution: We say G is an I-map for P if G is an I-map for $I(P)$, when we use $I(G)$ as the associated independence set.



Why does the graph get special privileges?

Facts about I-Maps

- For G to be an I-map of P, it is necessary that G does not mislead us regarding any independencies in P.
  - Any independence that G asserts must also hold in P. Conversely, P may have additional independencies that are not reflected in G.
  - "We must be able to use G to estimate P".
- Example



G_0 $G_{X \to Y}$ $G_{Y \to X}$

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

P_1

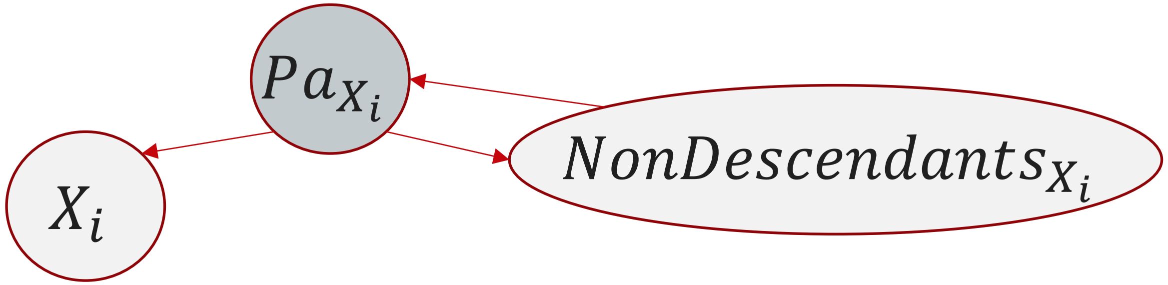
X	Y	P(X,Y)
0	0	0.4
0	1	0.3
1	0	0.2
1	1	0.1

P_2

From I(G) to local Markov assumptions of BNs

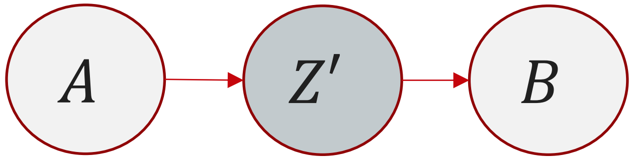
- In a BN, each node is independent of its non-descendants given its parents.
- Let Pa_{X_i} denote the parents of X_i in G and $NonDescendants_{X_i}$ denote the variables in the graph that are not descendants of X_i. Then G encodes the following set of *local conditional independence assumptions* $I_l(G)$:

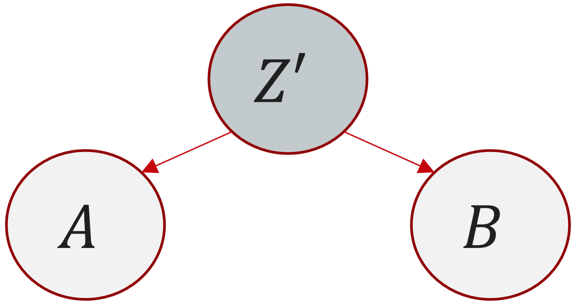
$$I_l(G) = \{X_i \perp NonDescendants_{X_i} | Pa_{X_i}: \quad \forall\, i\}$$

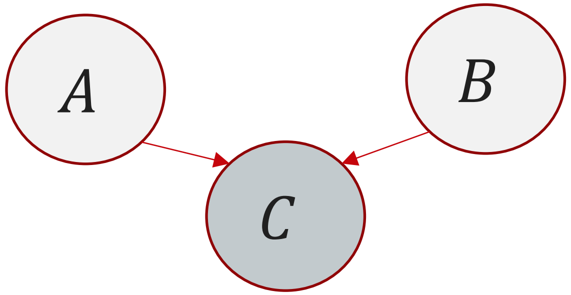


Graph separation

- D-separation criterion for Bayesian networks [Pearl, 1988]
  - D for "directed" edges
  - **Definition:** A set of nodes X is d-separated (conditionally independent) from a set of nodes Y given a conditioning set Z iff every path between any nodes in X and any node in Y is **blocked** by Z.
    - A path between nodes A and B is **blocked** by Z if it contains at least one of the following structures:
      - Chain: $A \rightarrow Z' \rightarrow B$ for $Z' \in Z$
      - Fork: $A \leftarrow Z' \rightarrow B$ for $Z' \in Z$
      - Collider: $A \rightarrow C \leftarrow B$ for $C \notin Z$ AND no descendant of C is in Z

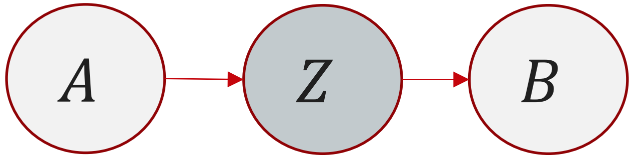


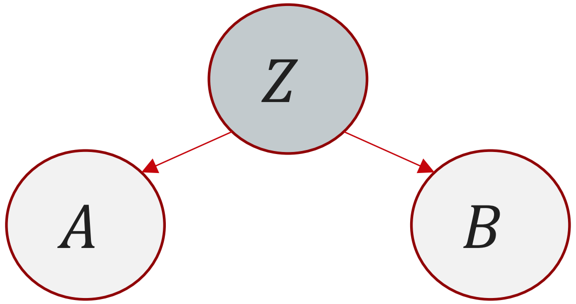


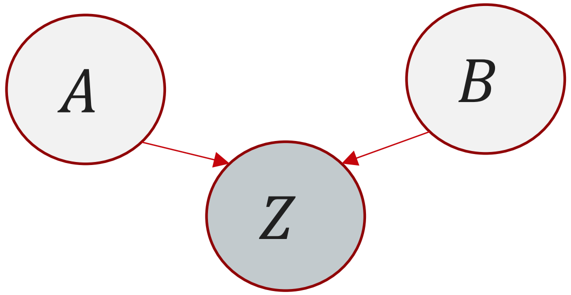


Active Trails

- Causal: $A \rightarrow Z \quad \rightarrow B$
  - Active iff Z is not observed.
- Common Cause: $A \leftarrow Z \quad \rightarrow B$
  - Active iff Z is not observed.
- Collider: $A \rightarrow Z \leftarrow B$
  - Active iff Z OR one of Z's descendants is observed.

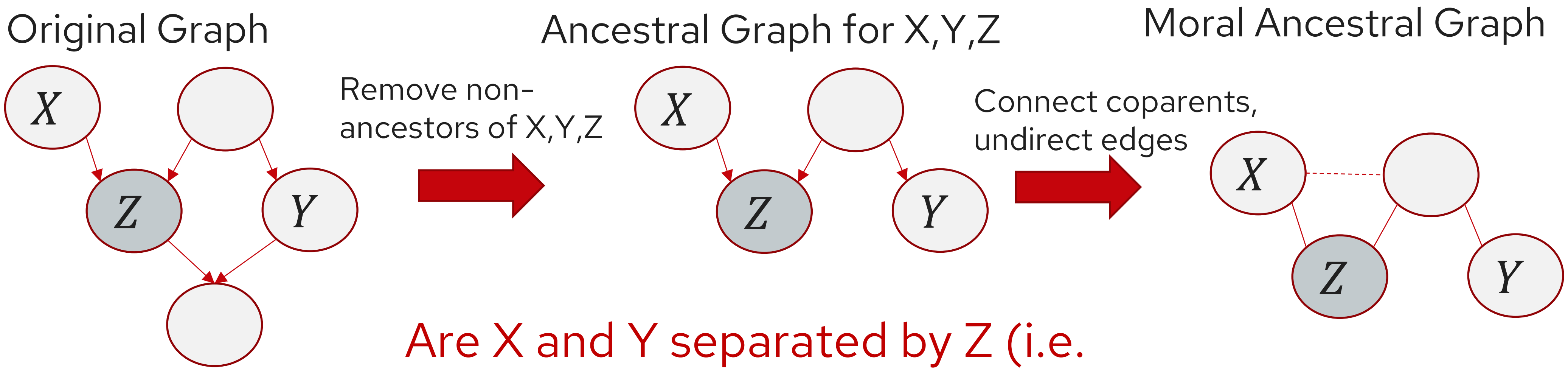






An alternate definition of D-separation

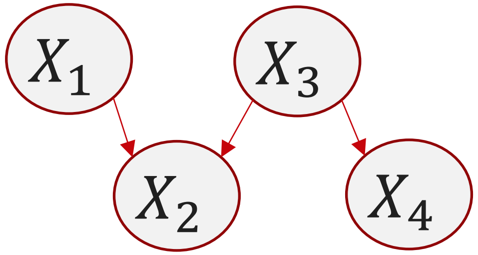
- <u>MAG Definition of D-Separation</u>
  - Variables X and Y are D-separated given Z if they are separated in the **m**oralized **a**ncestral **g**raph.
- Example:



Are X and Y separated by Z (i.e. removing Z disconnects X and Y)?

Example

- What is the I(G) of this graph?



- $X_1 \perp X_3$
- $X_1 \perp X_4$

- $X_1 \perp X_3 \mid X_4$
- $X_2 \perp X_4 \mid X_3$

Quantitatively Specifying Probability Distributions

Equivalence Theorem:

For a graph G,

Let D_1 denote the family of all distributions that satisfy $I(G)$.

Let D_2 denote the family of all distributions that factor according to G

$$P(X) = \prod_i P(X_i \mid X_{\pi_i})$$

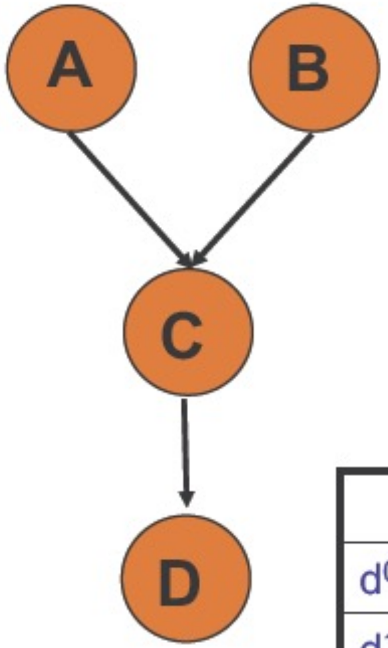
Then $D_1 = D_2$.

Conditional Probability Tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)



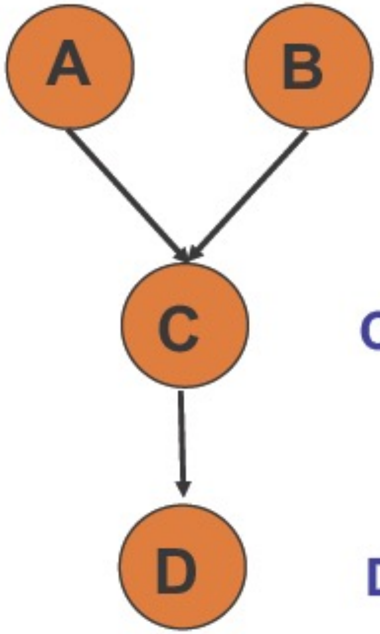
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Conditional Probability Density Functions (CPDs)

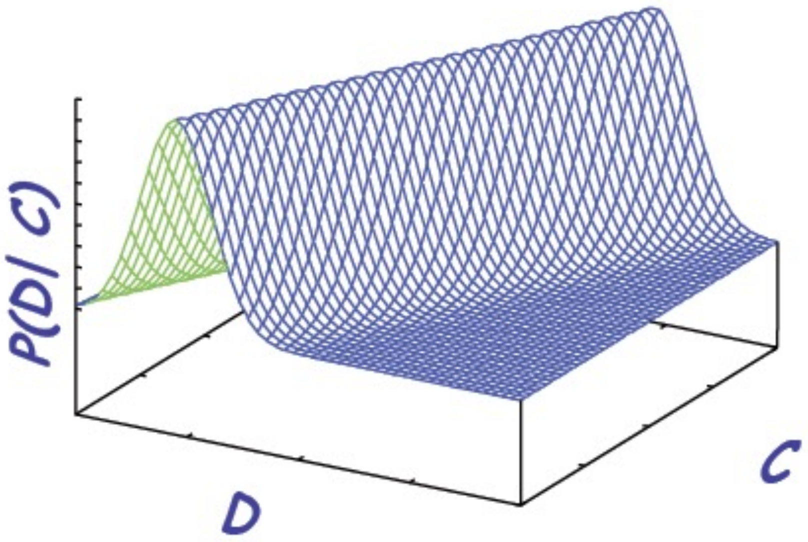
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

$P(a,b,c.d) =$
$P(a)P(b)P(c|a,b)P(d|c)$



$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_d+C, \Sigma_d)$



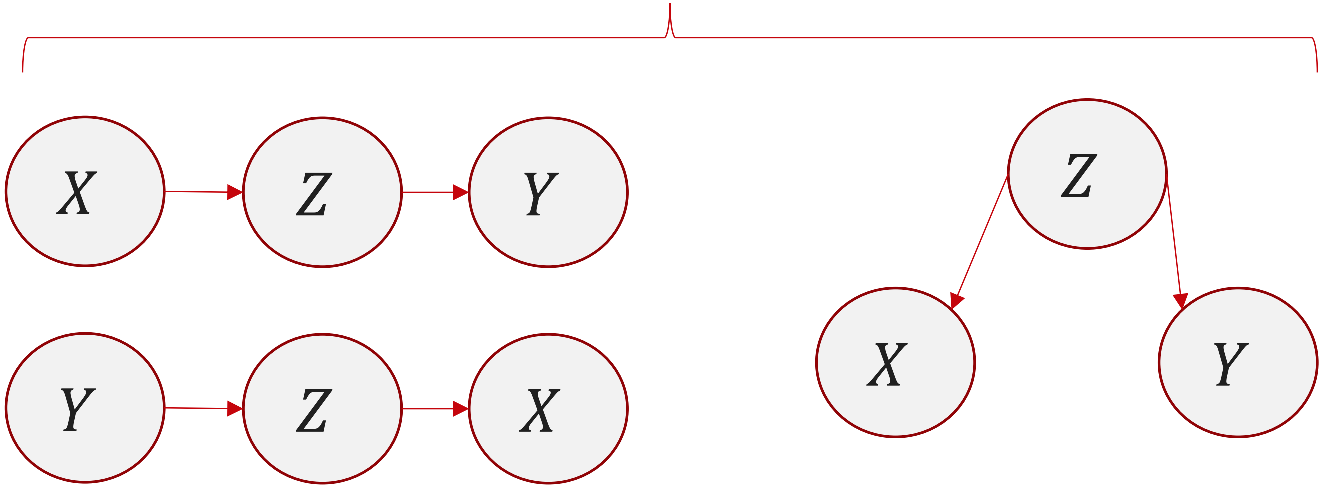
Summary of BN semantics

- A Bayesian Network is a pair (G, P) where P factorizes over G and where P is specified as a set of CPDs associated with G's nodes.

Uniqueness of BNs

- Very different BN graphs can be equivalent (in that they encode the same set of conditional independence assertions).

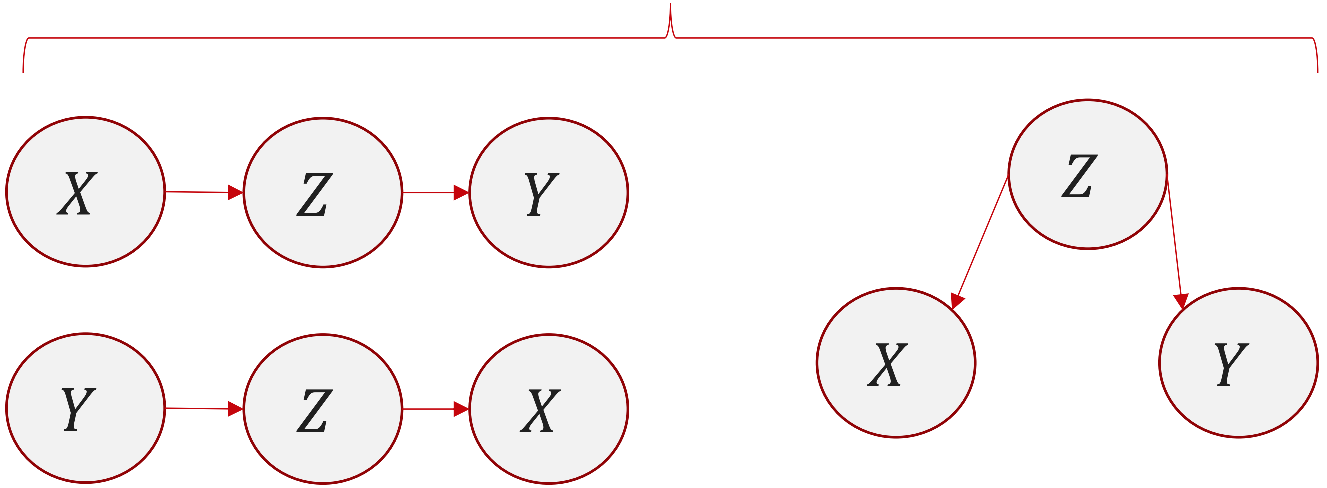
$$X \perp Y \mid Z$$



I-equivalence

- Definition of I-Equivalence: Two BN graphs G_1 and G_2 over X are *I-equivalent* if $I(G_1) \ = \ I(G_2)$.

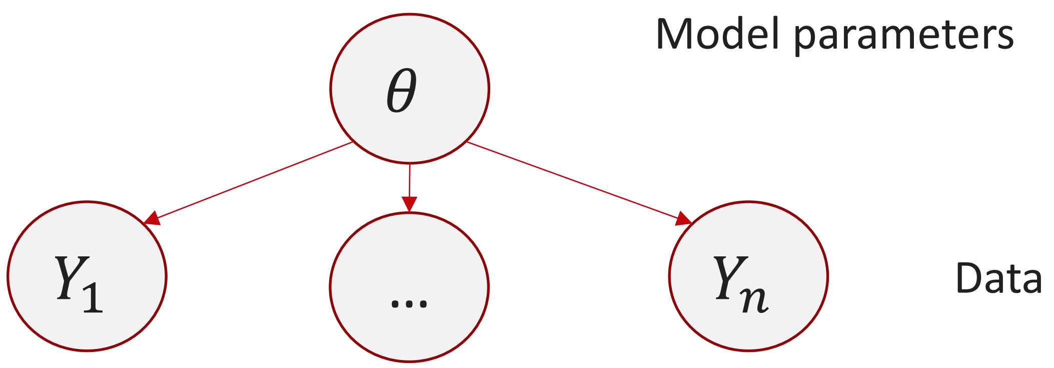
$$X \perp Y \mid Z$$



How can we distinguish structures when learning?

Simple BNs

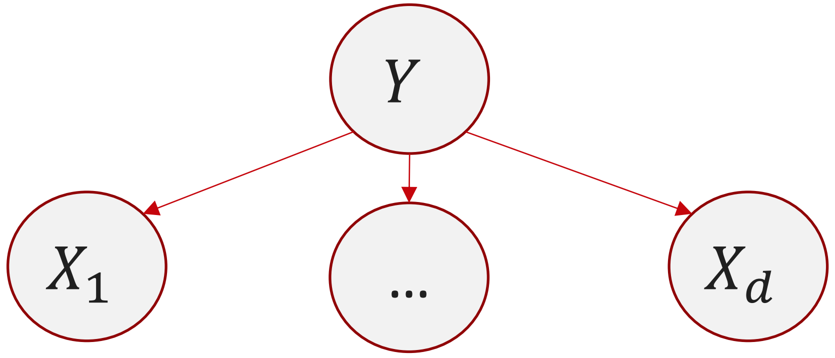
- IID Observations



$$P(Y;\theta) = P(\theta)\prod_i P(Y_i \mid \theta)$$

Simple BNs

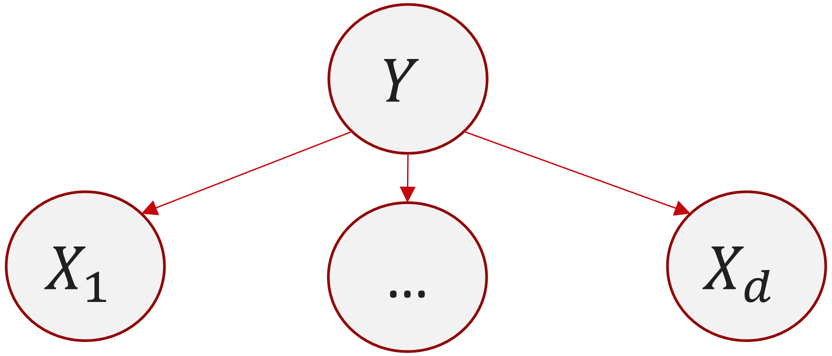
- Naïve Bayes



$$P(X \mid Y) = P(Y) \prod_i P(X_i \mid Y)$$

Notation: "Plate"

- Naïve Bayes with Streamlined Notation



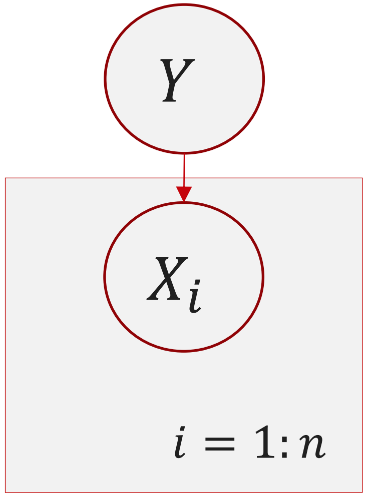


Plate notation

Variables within a plate are replicated
in a conditionally independent manner

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON