

Bayesian Networks I: BN Representation

For Self-learning Purposes

Jiashu Chen

February 6, 2024

Contents

1	Joint Distribution	1
1.1	Covid Example: 12 Data Sample	1
1.2	Probability	2
1.3	Summary: What is the Goal?	2
2	Independence	4
2.1	Importance of Independence	4
2.2	Statistical Independence	4
2.3	Conditional Independence	5
2.4	Conditional Independence with Chain Rule	6
2.5	Summary: What is the Goal?	6
3	Graphical Visualization	7
3.1	Parent Notation	7
3.2	Graph	7
3.3	Bayesian Networks	8
3.4	Covid Example	8
3.5	Summary: What is the Goal?	8
4	From Factorization to Independence: I-map	9
4.1	Local Independence from BN	9
4.2	$I_l(G)$ and $I(P)$: Example	10
4.3	Summary: What is the Goal?	10
5	Minimal I-map	11
5.1	Multiple I-map	11
5.2	Minimal I-map	11
5.3	Summary: What is the Goal?	12
6	Global Independence	13
6.1	Global Conditional Independence	13
6.2	Trails	15

6.3	d-separation	18
6.4	Summary: What is the Goal?	19
7	P-map	20
7.1	$I(G) \stackrel{?}{=} I(P)$	20
7.2	P-map	20
7.3	Summary: What is the Goal?	21
8	Independence Equivalence	22
8.1	I-equivalence	22
8.2	I-equivalence class: PDAG	23
8.3	Summary: What is the Goal?	23
9	Naive Bayes Model	25

Chapter 1

Joint Distribution

1.1 Covid Example: 12 Data Sample

#	Covid	Mask	Social Distancing
1	True	False	True
2	False	True	True
3	True	False	False
4	False	True	True
5	False	True	False
6	False	True	False
7	True	False	True
8	False	True	True
9	False	True	True
10	False	False	False
11	False	True	False
12	True	True	True

$$P(COVID = F \mid Mask = T, SocialDistancing = T) = ? \quad (1.1)$$

- Marginal Probability

$$P(M = T, D = T) = \sum_{C=T,F} P(C, M = T, D = T) = \frac{6}{12} \quad (1.2)$$

- Conditional Probability

$$\begin{aligned} P(C = F \mid M = T, D = T) &= \frac{P(C = F, M = T, D = T)}{P(M = T, D = T)} \\ &= \frac{5/12}{6/12} \end{aligned} \quad (1.3)$$

- Joint Probability Distribution

The key is to find the joint probability distribution for C, M, D, i.e., $P(C, M, D)$, with 8 parameters p_1, \dots, p_8 , where $\sum_{i=1}^8 p_i = 1$

So, the number of parameters to obtain $P(C, M, D)$ is $2^3 - 1 = 7$

Key Takeaways:

If we have the joint distribution, we can calculate any probability using conditional distribution and marginal distribution.

1.2 Probability

- Joint probability distribution

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

- Marginal probability distribution

Marginalize out some variables

$$P(X_1) = \sum_{x_2, \dots, x_n} P(X_1, X_2, \dots, X_n)$$

- Conditional probability distribution

$$P(X_1 | X_2 \dots X_n) = \frac{P(X_1 X_2 \dots X_n)}{P(X_2 \dots X_n)}$$

- Number of parameters of a distribution

Given discrete random variables $X_1 X_2 \dots X_n$ that take $\alpha_1, \alpha_2, \dots, \alpha_n$ values, respectively,

- the number of parameters to represent $P(X_1 X_2 \dots X_n)$ is

$$\alpha_1 \alpha_2 \dots \alpha_n - 1$$

- the number of parameters to represent $P(X_1 | X_2 \dots X_n)$ is

$$(\alpha_1 - 1) \alpha_2 \dots \alpha_n$$

1.3 Summary: What is the Goal?

The goal is to find the joint probability distribution from data. So that we could answer any probabilistic inquiries.

$$P(C, M, D) = ?$$

#	Covid	Mask	Social Distancing	Flu	Cough	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
...
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

However, in the real world, we often have much more variables. Suppose we have 12 variables in the covid example, all binary except for season with 4 values, then the number of required parameters to calculate joint probability is $2^{11} * 4 - 1 = 8191$

If we only have 1000 data points, then we have to set at least 7191 parameters to zero, which is problematic.

Chapter 2

Independence

So we understand the importance of joint probability distribution, the problem now is to deal with the large number of parameters while calculating joint probability distribution.

2.1 Importance of Independence

For Covid, Mask, Social Distancing, using conditional probability

$$P(C, M, D) = P(C | M, D)P(M, D)$$

Suppose M and D are independent, denoted by $M \perp D$,

$$P(M, D) = P(M)P(D)$$

$$\text{Thus, } P(C, M, D) = P(C | M, D)P(M)P(D)$$

$$\text{Number of parameters} = 2^2 + (2 - 1) + (2 - 1) = 6$$

$$P(C | M, D) : (2 - 1) * 2 * 2 = 2^2$$

$$P(M) : M \text{ takes two values} = 2$$

$$P(D) : D \text{ takes two values} = 2$$

2.2 Statistical Independence

Random variables X and Y are independent, i.e. $X \perp Y$, if

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

$$\text{or } P(X | Y) = P(X) \text{ or } P(Y | X) = P(Y)$$

Covid Example

#	Covid	Mask	Social Distancing	Flu	Cough	Fever	Ventilation	Season	ConGestion	Difficulty Breathing	DRug	Allergy
1	True	False	True	False	True	True	True	Spring	True	True	False	False
2	False	True	True	False	False	True	False	Summer	False	False	True	False
3	True	False	False	True	True	False	False	Fall	False	True	True	False
4	False	True	True	False	False	True	False	Winter	True	True	False	True
...
1000	True	True	True	False	True	False	True	Spring	False	False	True	True

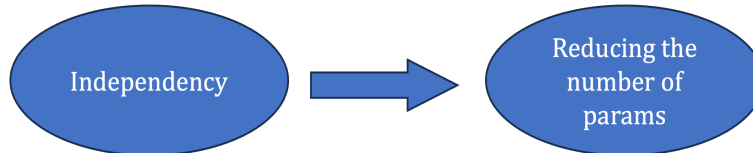
Suppose all 12 random variables are mutually independent:

$$C \perp M, C \perp D, \dots, R \perp A$$

$$P(C, M, D, U, \dots, A) = P(C)P(M)P(D) \cdots P(A)$$

$$\text{Number of parameters} = 11 + 3 = 14$$

The previous number was 8191.



Key Takeaways:

with M,D being independent, the number of parameters reduced from 7 to 6. Therefore, independence could reduce the number of parameters.

However, the independence assumption does not always hold. For instance, *Mask* $\not\perp$ *Social distancing*

2.3 Conditional Independence

- Bother, Sister, Parent Example

- Consider a brother(B) and a sister(S)

$$P(S = \text{white} \mid B = \text{white}) = 0.9$$

$$P(S = \text{white} \mid B = \text{black}) = 0.1$$

Therefore, $S \not\perp B$

- Consider brother(B), sister(S), and Parent(Pa)

$$P(S = \text{white} \mid Pa = \text{white}) = 0.95$$

$$P(S = \text{white} \mid Pa = \text{white}, B = \text{white}) = 0.95$$

$$P(S = \text{white} \mid Pa = \text{white}, B = \text{black}) = 0.95$$

Therefore, knowing the brother does not matter if we know the parent

S is conditionally independent of B given Pa, i.e. $(S \perp B \mid Pa)$

- Conditional Independence Definition

Consider the set of random variables X, Y, Z, We say that X is conditionally independent of Y given Z in a distribution P, denoted by $P \models (X \perp Y \mid Z)$ if

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z), \forall x, y, z$$

- The variables in set Z are said to be observed
- The set of all probability independencies in P is denoted by $I(P)$

$$P \models (X \perp Y \mid Z) \text{ if and only if}$$

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

- How can Conditional Independence Help

- Using Covid, Fever, and PRC test example

$$P(F, P, C) = P(F | P, C)P(P, C)$$

- if $P(F \perp P | C) \in I(P)$, i.e. fever and prc test are conditionally independent given covid, then

$$P(F, P, C) = P(F | C)P(P, C)$$

- The number of parameters = $2 + (2 * 2 - 1) = 5$

Key Takeaways:

Conditional independence can reduce the number of parameters.

2.4 Conditional Independence with Chain Rule

- We could factor joint distribution into CPDs using chain rule

$$\begin{aligned} P(X_1, \dots, X_n) &= P(X_n | X_{n-1}, \dots, X_1) \dots P(X_3 | X_2, X_1)P(X_2 | X_1)P(X_1) \\ &= \prod_{i=1}^n p(X_i | X_{i-1}, \dots, X_1) \end{aligned}$$

Suppose all random variables take binary values, then the number of parameters for

$$p(X_i | X_{i-1}, \dots, X_1) = 2^{(i-1)}$$

- Consider joint distribution $P(X_1, X_2, X_3, X_4)$. Using the chain rule, we could get

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

- If $(X_4 \perp X_1, X_2 | X_3) \in I(P)$, then

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

- The joint distribution becomes

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

- The total number of parameters reduced from $2^4 - 1 = 15$ to $2 + 2^2 + 2 + 1 = 9$

2.5 Summary: What is the Goal?

- Now we know that to calculate joint probability distribution, we just need to find the conditional independencies and factorize the joint probability distribution. But, there are two more problems

- How to find conditional independencies $I(P)$ from data?
- Given the conditional independencies, how to factorize the joint distribution?

- **To visualize the factorization of joint probability distribution is Bayesian Networks!**

- What is the goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- $P(C)P(M | C)P(D)$: 4 params
- $P(C | M)P(M)P(D)$: 4 params
- ...
- $P(C | M)P(M | D)P(D)$: 5 params
- $P(M | C, D)P(C)P(D)$: 6 params

Different factorization leads to different number of parameters.

Chapter 3

Graphical Visualization

3.1 Parent Notation

- In previous example, If $(X_4 \perp X_1, X_2 \mid X_3) \in I(P)$, then

$$P(X_4 \mid X_3, X_2, X_1) = P(X_4 \mid X_3)$$

- For each X_i , we define the parents of X_i , denoted by Pax_i , as the set of variables that X_i is conditioned on in the factorization.

$$Pax_4 = X_3, Pax_3 = X_1, X_2, Pax_2 = X_1, Pax_1 = \emptyset$$

- Then the joint distribution could be written as

$$P(X_1, X_2, X_3, X_4) = \prod_{i=1}^4 P(X_i \mid Pax_i)$$

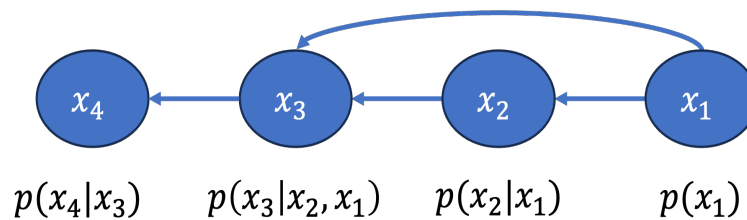
3.2 Graph

Now we could construct the directed graph $G = (V, E)$ with vertices $V = X_1, X_2, X_3, X_4$ and where E is the set of directed edges from Pax_i to X_i for $i = 1, 2, 3, 4$

Directed Acyclic Graph (DAG)

Bayesian Networks: visualizing the way we factorize our joint probability distribution.

Lack of an edge indicates conditional independence.



3.3 Bayesian Networks

- Factorization Definition (Chain rule for Bayesian Networks)
The distribution P factorizes according to the DAG G , if

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i \mid \text{Pax}_i^G)$$

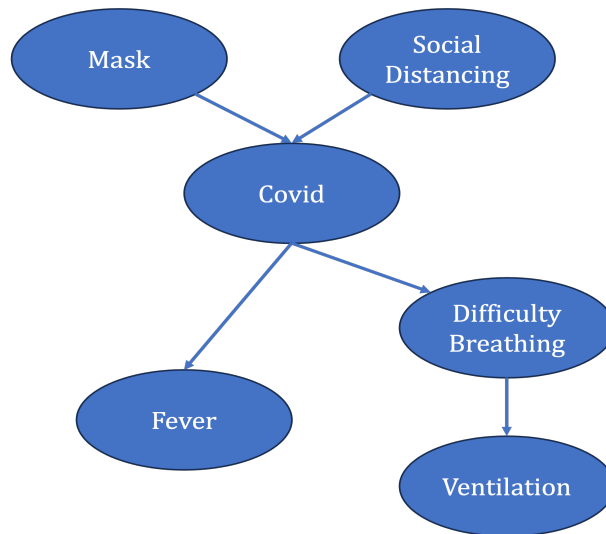
- Bayesian Networks Definition
Given the random variables $V = X_1, X_2, \dots, X_n$, a Bayesian Network(BN) is a pair $B = (G; P_B)$, where
 - G is directed acyclic graph (**BN structure**) with node set V .
 - P_B is a probability function that factorizes according to G and is specified as a set of conditional probability distributions (CPDs) $P_B(X_i \mid \text{Pax}_i)$ for all $X_i \in V$ (**BN parameters**).

3.4 Covid Example

- Assume following joint probability distribution for Mask, Social distancing, Covid, Fever, Difficulty Breathing, and Ventilation

$$P(M, D, C, B, F, V) = P(F \mid C)P(V \mid B)P(B \mid C)P(C \mid M, D)P(M)P(D)$$

- Corresponding Bayesian Networks



BN structure is a representation of how the joint probability distribution of a set of random variables can be factorized.

3.5 Summary: What is the Goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- First, finding the conditional independence $I(P)$.
- Use graphical visualization (BN), then determine which graph is the correct factorization.

Chapter 4

From Factorization to Independence: I-map

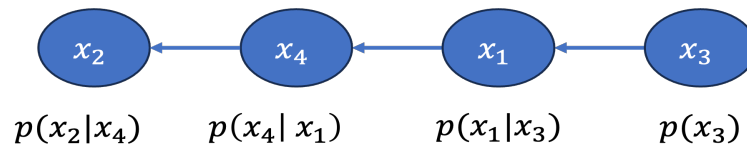
4.1 Local Independence from BN

- The algebraic way of finding independence from factorization is complicated. The easier way is to use Bayesian Networks.

Assume we have the following factorization

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

Equivalent BN



Then, $(X_2 \perp X_1, X_3 | X_4), (X_4 \perp X_3 | X_1) \in I(P)$

Lack of edges indicates conditional independence.

$$(X_2 \perp NonDescendants_{X_2} | Pax_2), (X_4 \perp NonDescendants_{X_4} | Pax_4) \in I(P)$$

$NonDescendants_{X_i}$ are all the nodes excluding the descendants of X_i .

These conditional independencies are known as the local independencies of the graph because they are conditioned on X_i 's parents.

- Local Independence Definition

Given the graph G, the set of **local (Markov) independencies**, denoted by $I_l(G)$, consists of

$$(X_i \perp NonDescendants_{X_i} | Pax_i^G) \forall i$$

- Independence-map (I_map) Definition

$$G \text{ is an I_map for } P \text{ if } I_l^G \subseteq I(P)$$

So, G being an I_map for P means that P satisfies the local independencies of G.

- I_map Theorem

Let G be a DAG and P be a joint distribution over a set of random variables. P factorizes according to G , if and only if G is a I_map for P .

4.2 $I_l(G)$ and $I(P)$: Example

- Consider the joint distribution P over X_1, X_2, \dots, X_4 , where

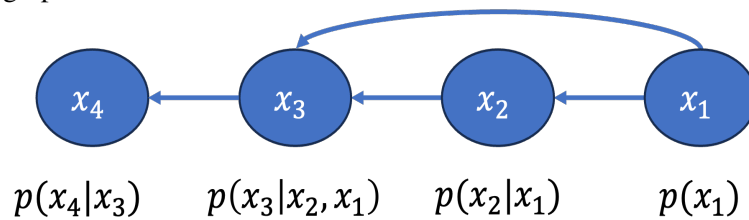
$$I(P) = (X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}$$

- Factorization 1

- * Does P satisfy the following factorization?

$$P(X_1, \dots, X_4) = P(X_4 \mid X_3)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- * Equivalent graph G



imposing local independencies:

$$I_l(G) = (X_4 \perp X_2, X_1 \perp X_3)$$

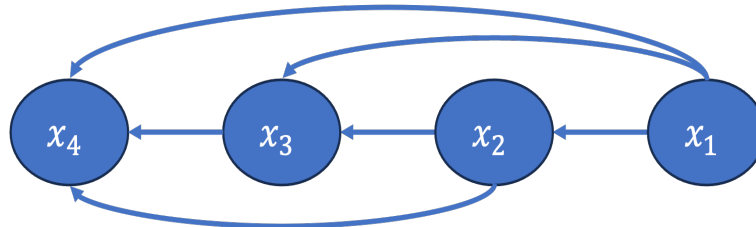
Therefore, $I_l(G) \subseteq I(P)$, G is an i-map for P .

- Factorization 2

- * Does P satisfy the following factorization?

$$P(X_1, \dots, X_4) = P(X_4 \mid X_3, X_2, X_1)P(X_3 \mid X_2, X_1)P(X_2 \mid X_1)P(X_1)$$

- * Equivalent graph G



imposing local independencies:

$$I_l(G) = \emptyset$$

Therefore, $I_l(G) \subseteq I(P)$, G is an i-map for P .

- There could be more than one i_map for P .

4.3 Summary: What is the Goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- First, finding the conditional independence $I(P)$
- For each factorization, we could draw a bayesian nets, and we could write down local independencies I_l^G . If $I_l^G \subseteq I(P)$, then G is i_map for P
- There could be several i_map for P , the problem now is which one to choose?

Chapter 5

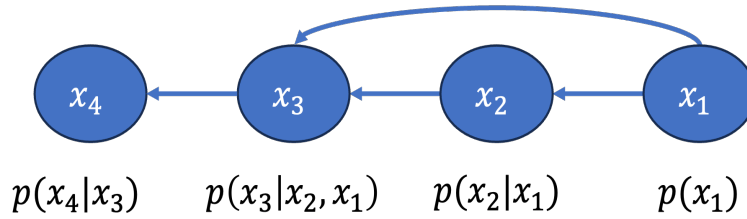
Minimal I-map

5.1 Multiple I-map

- Consider the joint distribution P over X_1, X_2, \dots, X_4 , where

$$I(P) = (X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}$$

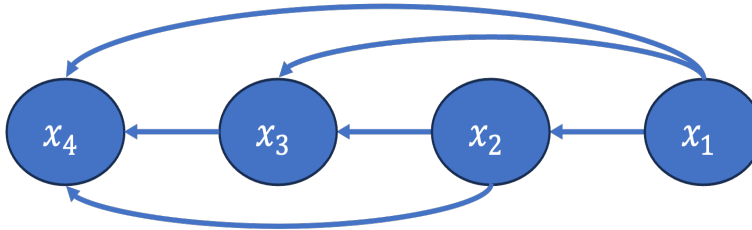
- Graph1



$$I_l(G_1) = (X_4 \perp X_2, X_1 \perp X_3)$$

Therefore, $I_l(G_1) \subseteq I(P)$, G_1 is an i-map for P .

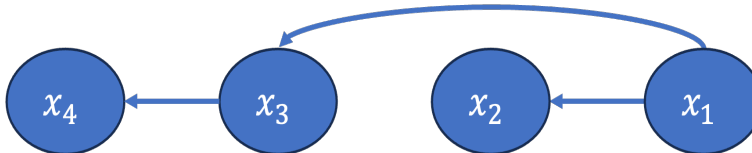
- Graph2



$$I_l(G_2) = \emptyset$$

Therefore, $I_l(G_2) \subseteq I(P)$, G_2 is an i-map for P .

- Graph3



$(X_3 \perp X_2 \mid X_1) \in I_l(G_3)$ $(X_3 \perp X_2 \mid X_1) \notin I(P)$ Therefore, G_3 is not an i-map for p .

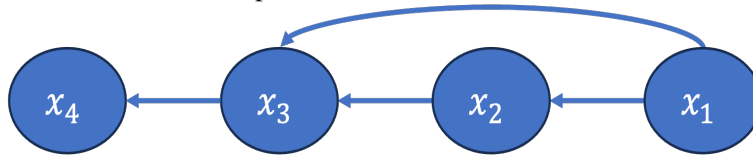
5.2 Minimal I-map

- Definition

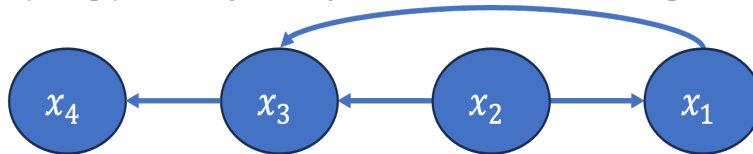
A graph G is a minimal I-map for P if it is an I-map for P , and if the removal of any edge from G makes it not an I-map.

– Minimal I-maps are not Unique

* G_1 is a minimal I-map



* By simply reversing one edge, G_2 is also a minimal I-map.



5.3 Summary: What is the Goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- First, finding the conditional independence $I(P)$
- For each factorization, we could draw a bayesian nets, and we could write down local independencies I_l^G . If $I_l^G \subseteq I(P)$, then G is i_map for P
- There could be several i_map for P , the problem now is which one to choose? The minimal I-maps.

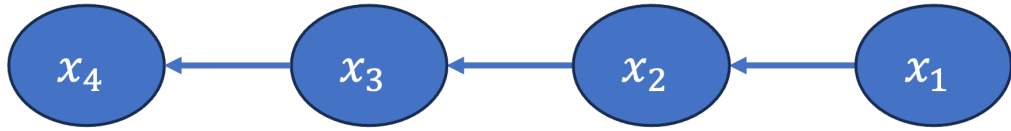
Chapter 6

Global Independence

6.1 Global Conditional Independence

- Consider the joint distribution P over X_1, X_2, \dots, X_4 , factorizing as

$$P(X_1, \dots, X_4) = P(X_4 | X_3)P(X_3 | X_2)P(X_2 | X_1)P(X_1)$$



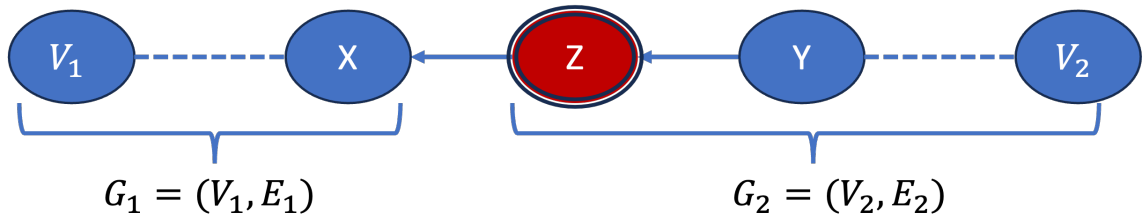
Equivalent graph G , imposing the local independencies:

$$I_l(G) = (X_4 \perp X_1, X_2 | X_3), (X_3 \perp X_1 | X_2)$$

- What if we condition X_4 on X_2

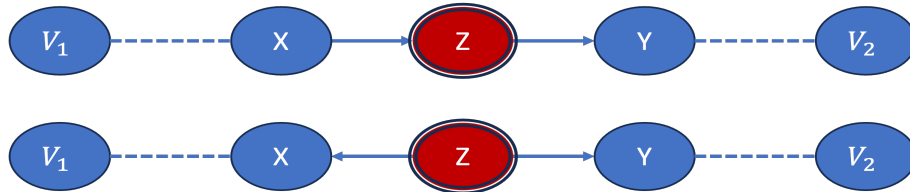
$$\begin{aligned}
 P(X_4, X_1 | X_2) &= \frac{P(X_4, X_1, X_2)}{P(X_2)} \\
 &= \frac{\sum_{X_3} P(X_4, X_3, X_2, X_1)}{P(X_2)} \\
 &= \frac{\sum_{X_3} P(X_4 | X_3)P(X_3 | X_2)P(X_2 | X_1)P(X_1)}{P(X_2)} \\
 &= \frac{\sum_{X_3} P(X_4 | X_3)P(X_3 | X_2)P(X_2)P(X_1 | X_2)}{P(X_2)} \\
 &= \frac{\sum_{X_3} P(X_4, X_3, X_2)P(X_1 | X_2)}{P(X_2)} \\
 &= \frac{P(X_4, X_2)P(X_1 | X_2)}{P(X_2)} \\
 &= P(X_4 | X_2)P(X_1 | X_2)
 \end{aligned} \tag{6.1}$$

- This type of conditional independence holds for a general graph including node V_1 and V_2 that are connected by some path, but become disconnected if node Z is removed.



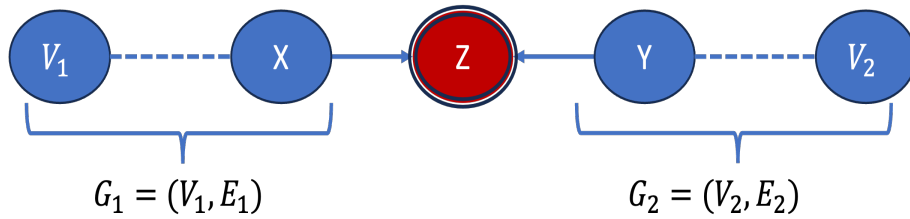
Hence, $V_1 \perp V_2 \mid Z$

It holds true for the following graphs:



The observed node Z blocks information flows between X and Y , and consequently V_1 and V_2

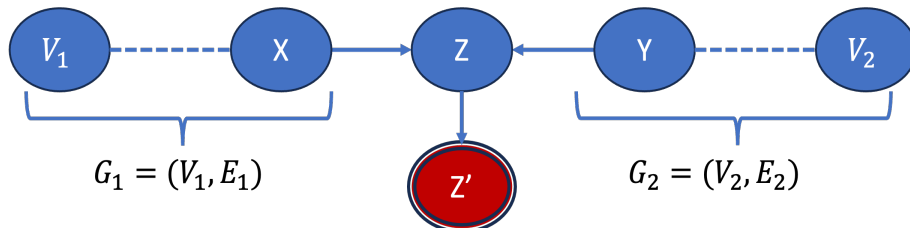
- What about $X \rightarrow Z \leftarrow Y$, known as the v-structure



$$\begin{aligned}
 P(V_1, V_2) &= \sum_{v \neq V_1, V_2} \prod_v P(v \mid Pa_v) \\
 &= \sum_{v \neq V_1, V_2} \prod_{v \in V_1} P(v \mid Pa_v) \prod_{v \in V_2} P(v \mid Pa_v) P(Z \mid X, Y) \\
 &= \sum_{v \neq V_1, V_2, Z} \prod_{v \in V_1} P(v \mid Pa_v) \prod_{v \in V_2} P(v \mid Pa_v) \sum_Z P(Z \mid X, Y) \\
 &= \sum_{v \neq V_1, V_2, Z} P(V_1) P(V_2) \\
 &= P(V_1) P(V_2)
 \end{aligned} \tag{6.2}$$

Hence, $V_1 \perp V_2$

- The conclusion does not hold if Z had an observed descendant



6.2 Trails

- In all of the cases, there are two nodes, v_1 and v_2 .

- information flows between them, they can be dependent.



- information flow is blocked, they are globally conditionally independent.



- A sequence of consecutively connected nodes $x_1 \Leftarrow x_2 \Leftarrow \dots \Leftarrow x_n$ in a graph is called a trail.
- A trail is **active** if information flows and **inactive** if the flow is blocked.
- Three types of inactive trails

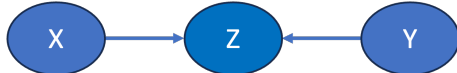
- Indirect, causal effect



- Common cause



- V-structure



- Inactive trail \Leftrightarrow Global conditional independence

- Four types of active trails

- Indirect, causal effect



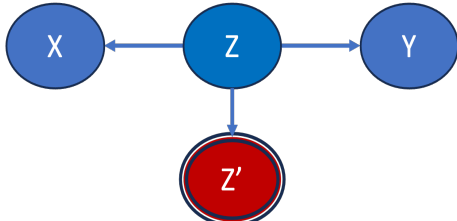
- Common cause



- V-structure



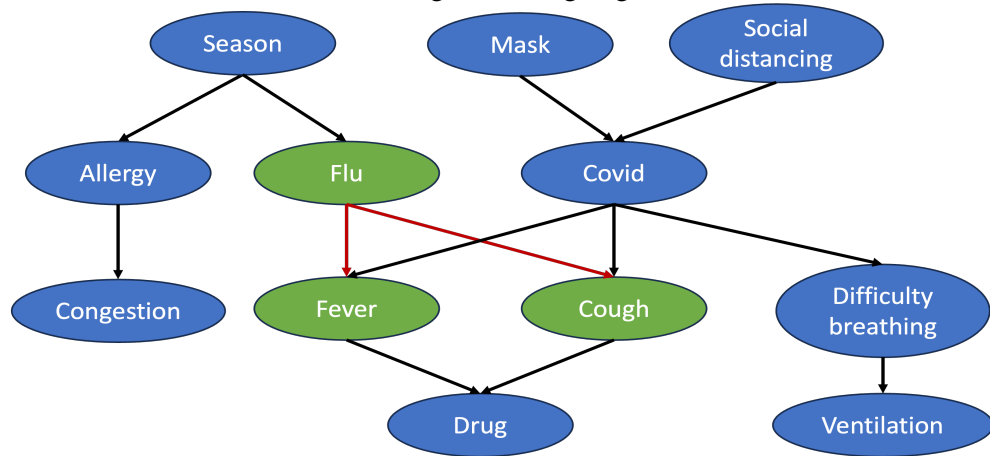
- V-structure



- Common Cause Example: Consider the trail Fever \Leftarrow Flu \rightarrow Cough in the covid example

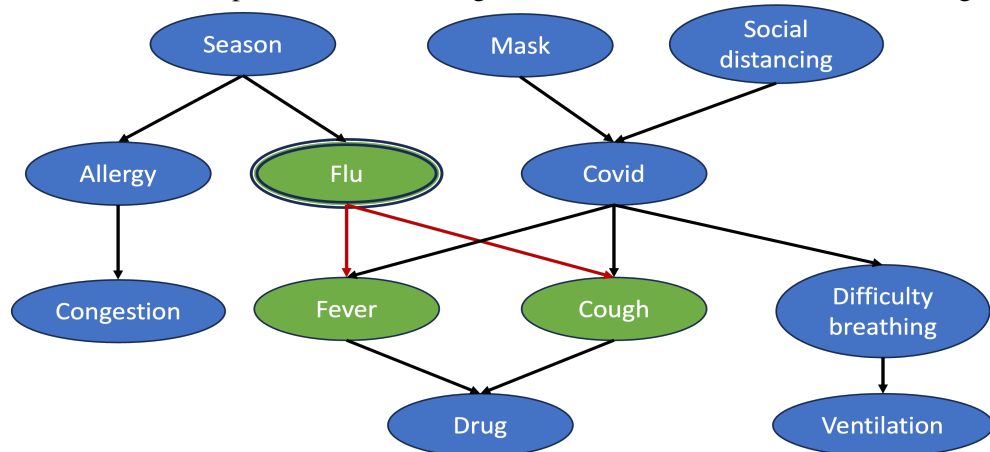
- If $Z = \emptyset \Rightarrow$ active trail

Fever and cough are dependent: a person with fever gives us information about that person having flu or coughing.



- If $Z = \text{Flu} \Rightarrow$ inactive trail

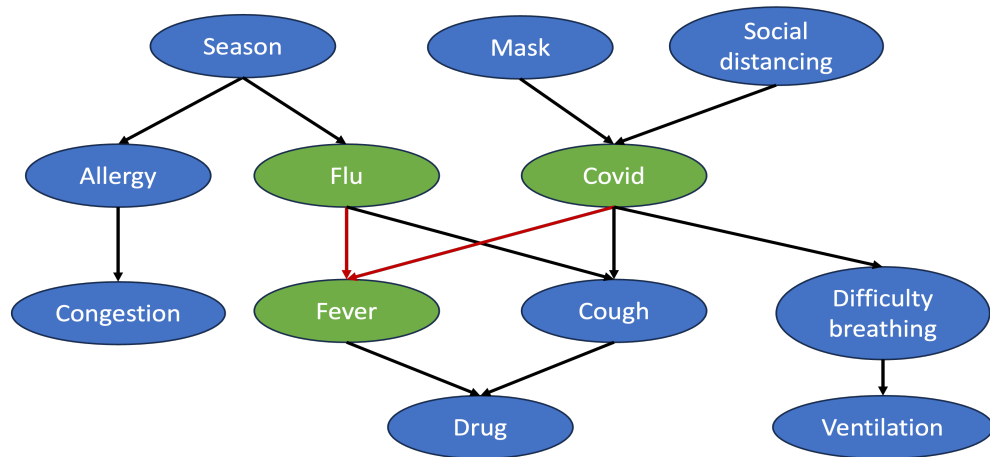
Once we know the person has flu, fever gives no additional information about cough.



- V-structure Example: Consider the trail $\text{Flu} \rightarrow \text{Fever} \leftarrow \text{Cough}$ in the covid example

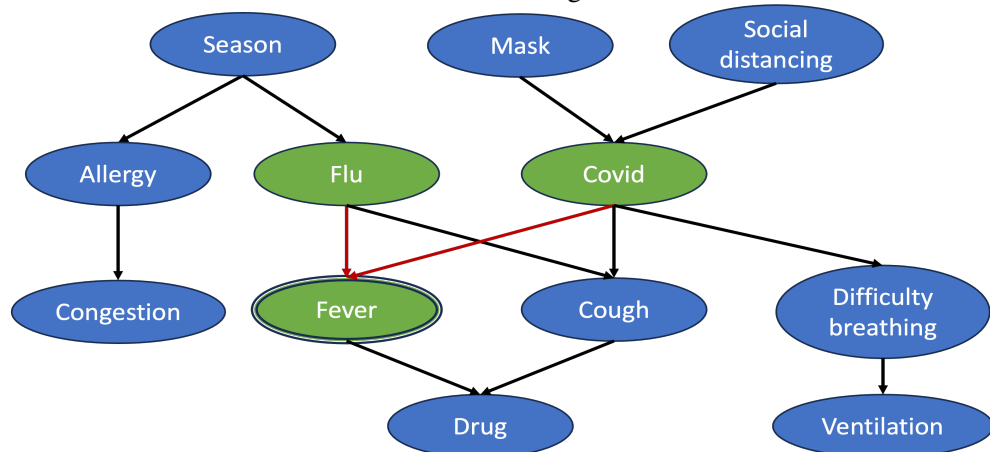
- If $Z = \emptyset \Rightarrow$ inactive trail

having flu, does not make a person more or less likely to have covid. So, no information flows along the trail.

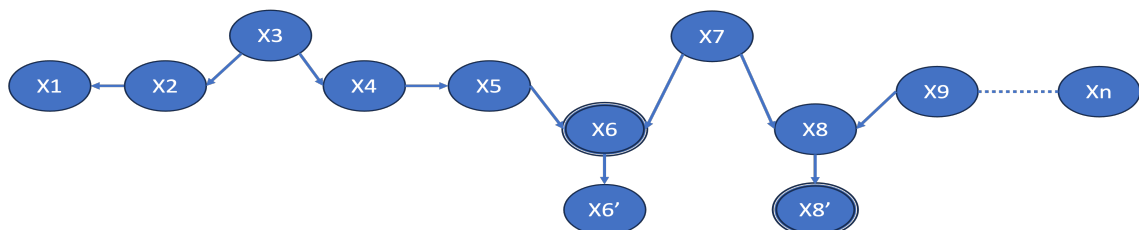


- If $Z = \text{Fever} \Rightarrow$ active trail

Flu and covid become dependent. If a person is known to have fever, then she/he is likely infected with covid if not having flu and vice versa.



- Active trails



Put all the nodes on a horizontal line. Move common cause above the line and v-structure below the line. In an active trail, **exactly the nodes at the bottom lines are observed**. Only sinks could be observed.

- Definition

- Active trail

Let Z be the set of observed variables. Trail $x_1 \Leftarrow \dots \Leftarrow x_n$ in G is active given Z if: 1. For any v-structure $x_{i-1} \rightarrow x_i \leftarrow x_{i+1}$, either x_i or one of its descendants are in Z . 2. No other node along the trail is in Z .

- In order to have global conditional independence, all trails must be inactive (no information flow).
- Inactivity implies independence: Given observed Z , if there is no active trail between the nodes X and Y in graph G , then X and Y are independent given Z under any probability distribution that factorize according to G .

6.3 d-separation

- D(dependence)-separation definition

In a graph G , two sets of nodes X and Y are d-separated given the set Z , denoted

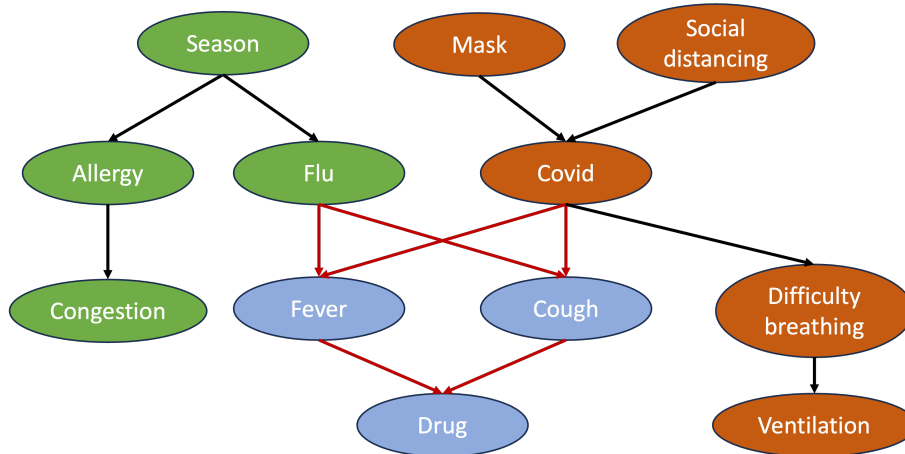
$$d\text{-sep}_G(X; Y \mid Z)$$

, if there is no active trail between any node $x \in X$ and any node $y \in Y$ given Z .

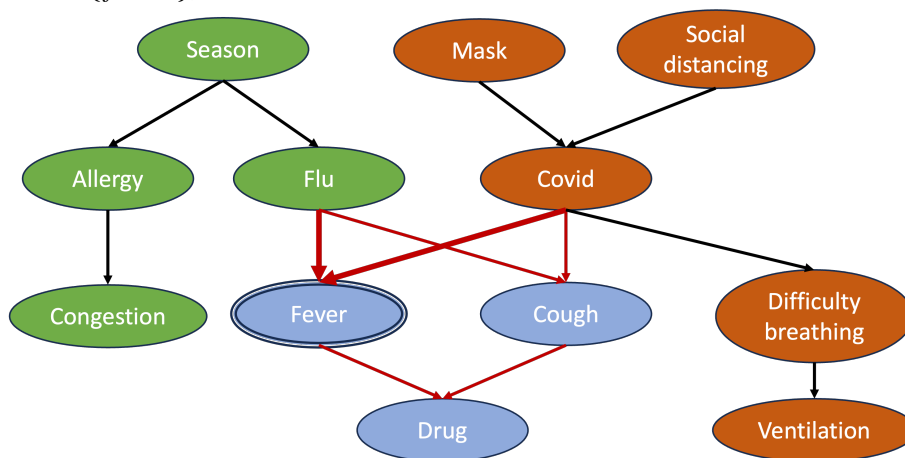
- Covid example

Consider the sets $X = \text{Season, Allergy, Flu, Congestion}$ and $Y = \text{Mask, Social distancing, COVID, Difficulty breathing, Ventilation}$.

- If $Z = \emptyset$, all trails are inactive



- If $Z = \{\text{fever}\}$, there exists an active trail.



- Definition: Global Markov Independence

The set of independencies resulting from d-separation is defined as the set of global markov independencies.

$$I(G) = (X \perp Y \mid Z) : d - sep_G(X; Y \mid Z)$$

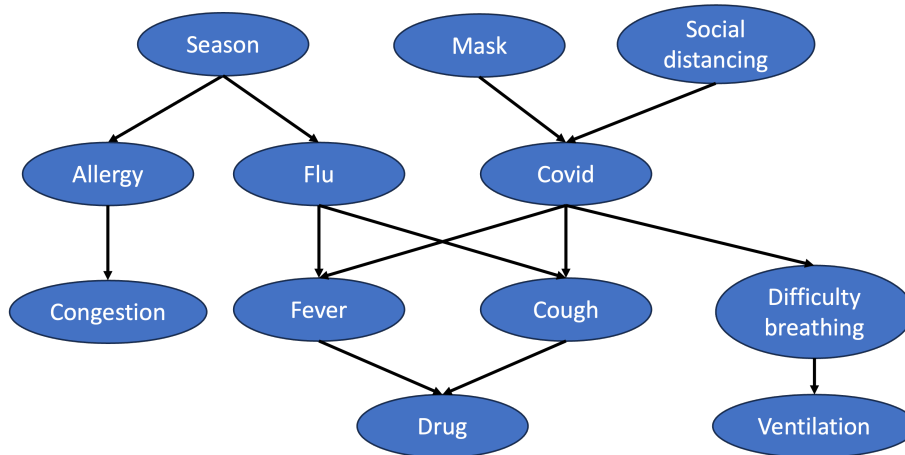
- Theorem

If a distribution P factorizes according to G , then $I(G) \subseteq I(P)$.

$$I_l(G) \subseteq I(P) \Rightarrow I(G) \subseteq I(P)$$

i.e., if the local independencies are satisfied by P , so are the global.

- Example



- Assume that the true joint probability distribution P of random variables satisfies the previously stated independencies.
- Is the graph an i-map for P ? Does P factorize according to the graph? yes
- Does the factorization impose any independence in addition to the previously stated one? yes, but P satisfied them all.

6.4 Summary: What is the Goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- First, finding the conditional independence $I(P)$
- For each factorization, we could draw a bayesian nets, and we could write down local independencies I_l^G . If $I_l^G \subseteq I(P)$, then G is I_map for P
- There could be several i_map for P , which one to choose? The minimal I-maps
- Does P satisfy all global independencies imposed by the minimal I-maps? yes

Chapter 7

P-map

7.1 $I(G) \stackrel{?}{=} I(P)$

- Recall from last chapter: If a distribution P factorizes according to G , then $I(G) \subseteq I(P)$. It means that all independencies implied by G are also included in P .
- The question now is can P have an independence not included in G ? The answer is Yes.
- Can we conclude $I(P) = I(G)$? No, but almost yes.
- Theorem of Completeness

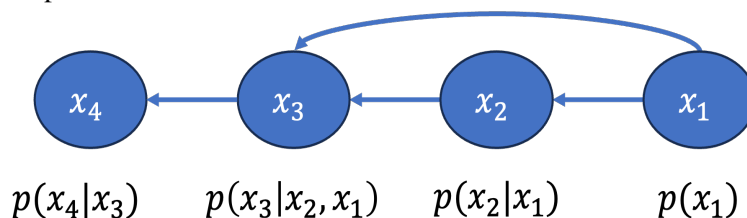
For almost all distributions P that factorize over G , except for a set of measure zero in the space of CPD parameterizations, $I(G) = I(P)$

7.2 P-map

- Definition: Graph G is a perfect map (P-map) for P if $I(P) = I(G)$.
- Consider the joint distribution P over X_1, X_2, \dots, X_4 , where

$$I(P) = (X_4 \perp X_2, X_1 \mid X_3) \text{ and its derivations}$$

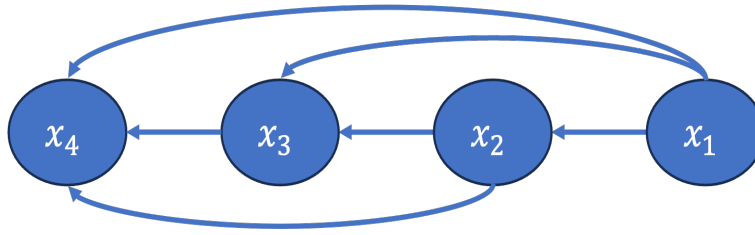
– Graph1



$$I_l(G_1) = (X_4 \perp X_2, X_1 \perp X_3)$$

Therefore, $I(G_1) = I(P)$, G_1 is an p-map for P .

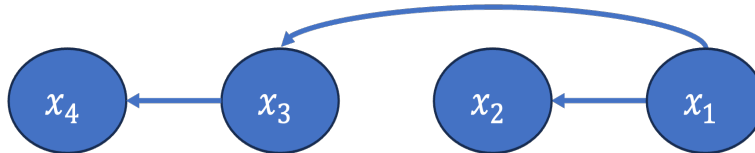
– Graph2



$$I_l(G_2) = \emptyset$$

Therefore, $I(G_2) \subseteq I(P)$, G_2 is not a p-map for P .

– Graph3



$(X_3 \perp X_2 \mid X_1) \in I_l(G_3)$ $(X_3 \perp X_2 \mid X_1) \notin I(P)$ Therefore, G_3 is not an i-map for p and not a p-map for p .

7.3 Summary: What is the Goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- First, finding the conditional independence $I(P)$
- For each factorization, we could draw a bayesian nets, and we could write down local independencies I_l^G . If $I_l^G \subseteq I(P)$, then G is I_map for P
- There could be several i_map for P , which one to choose? The minimal I-maps
- Does P satisfy all global independencies imposed by the minimal I-maps? yes
- Do the minimal I-maps satisfy all independencies imposed by P ? Are the minimal I-maps a P-map? Depends on $I(P)$

Chapter 8

Independence Equivalence

From previous chapters, we understand that there could be multiple minimal I-maps, is there a class of minimal I-maps?

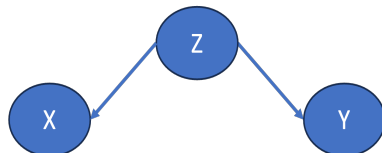
8.1 I-equivalence

- Using the Bayes rule, 3 types of structures have equal distributions. For all three graphs, $I(G) = X \perp Y \mid Z$

– $P(X, Y, Z) = P(X)P(Z \perp X)P(Y \perp Z)$



– $P(X, Y, Z) = P(Z)P(X \perp Z)P(Y \perp Z)$

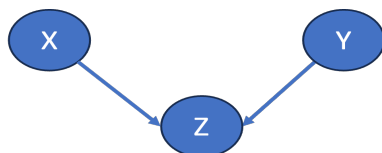


– $P(X, Y, Z) = P(Y)P(Z \perp Y)P(X \perp Z)$



- V-Structure

Suppose we have $I(G) = X \perp Y$, $P(X, Y, Z) = P(X)P(Y)P(Z \perp X, Y)$, it cannot be converted to another structure using the bayes rule.



- Definition: Independence-equivalence

Two graphs G_1 and G_2 defined over the same variables are I-equivalent if $I(G_1) = I(G_2)$. If two graphs have the same skeleton and set of v-structures, then they are i-equivalent.

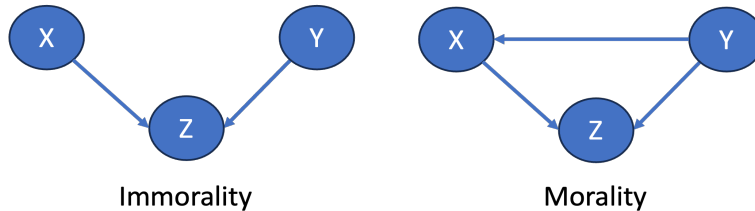
- General trail

In general, the arrows in a trail can be reversed as long as a new v-structure is not produced.

- Immorality

A v-structure $X \rightarrow Z \leftarrow Y$ is an immorality if X and Y are not linked. If there is a link, it is called a covering edge.

The graphs G_1 and G_2 are i-equivalent if and only if they have the same skeleton and the same set of immoralities.



8.2 I-equivalence class: PDAG

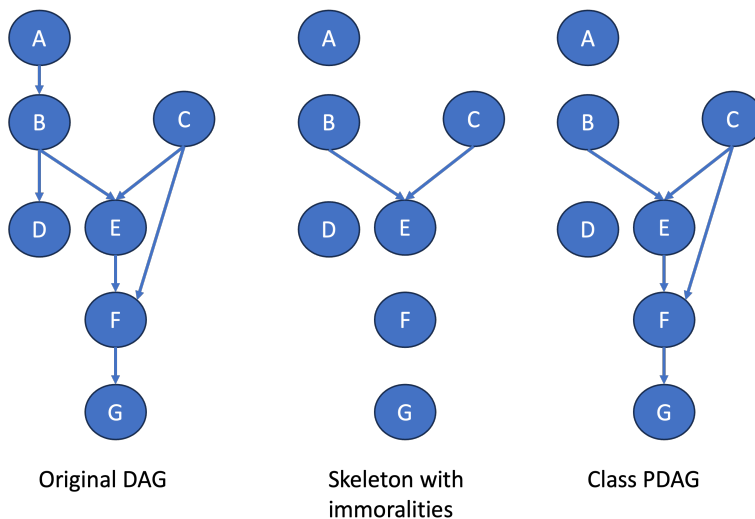
- Class PDAG of a DAG g is a PDAG that

- has the same skeleton as G
- includes a directed edge only if all of the i-equivalent graphs to G also have that directed edge.

- How to obtain the class PDAG of a DAG G

- obtain the skeleton of G .
- find the immoralities of G and draw them in the skeleton.
- find the orientations of the other edges by obeying the following rules: 1. do not create extra immorality. 2. do not create a directed cycle.

- Example



8.3 Summary: What is the Goal?

The goal is to find from data, the correct factorization of the joint probability distribution.

- First, finding the conditional independence $I(P)$

- For each factorization, we could draw a bayesian nets, and we could write down local independencies I_l^G . If $I_l^G \subseteq I(P)$, then G is I_map for P
- There could be several i_map for P, which one to choose? The minimal I-maps
- Does P satisfy all global independencies imposed by the minimal I-maps? yes
- Do the minimal I-maps satisfy all independencies imposed by P? Are the minimal I-maps a P-map?
Depends on I(P)
- Is there a class of minimal I-maps? Yes

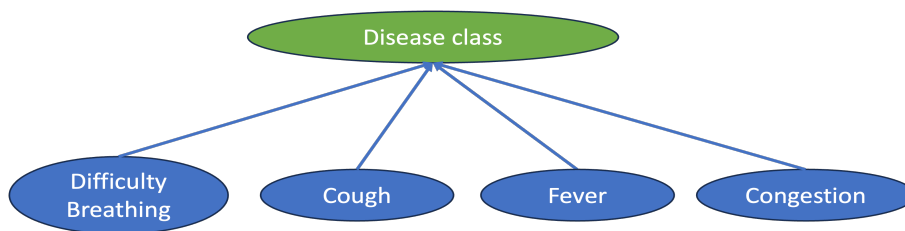
Chapter 9

Naive Bayes Model

In the Covid example, suppose we want to detect the disease type based on the symptoms. What is the simplest BN that can do this?

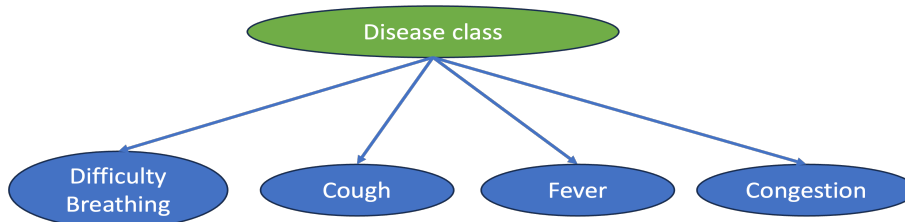
Assume Disease class Y = Allergy, Flu, COVID

- Option 1



This BN will lead to 32 parameters, $P(\text{Disease} | B, C, F, G)$.

- Option 2



This BN will lead to 15 parameters.

- Local dependencies: $I(G) = \{(G \perp F, O, B \mid Y), (F \perp O, G, B \mid Y), (O \perp F, G, B \mid Y), (B \perp F, O, G \mid Y)\}$
- So the symptoms are mutually independent given the disease class.
- By specifying the CPDs, we have the probability of each symptoms given the disease.
- Total number of params is 15: 2 for Y and 3 for each symptom.
- $P(Y, X_1, \dots, X_n)$ are factorized as $P(Y) \prod_{i=1}^n P(X_i \mid Y)$
- With joint probability distribution, we could calculate the probability of each disease given the symptoms.