Causal Inference in Statistics

Chapter 1 and 2

presented by Kun Woong Kim

Seoul National University

2020.1.31.

Contents

Introduction

Definition (Cause)

A variable X is a *cause* of a variable Y
if Y in any way relies on X for its value.

Why study **causation**?

- We need to make sense of data to guide actions and to learn from our success and failures.

e.g. Is malaria transmitted by mosquitoes or air?

SCM

SCM (Structural Causal Models)

- U : A set of exogenous variables
  (external; not to explain how they are caused)
- V : A set of endogenous variables
  (descendant of an exogenous variable)
- F : A set of functions which assign values of variables in V based on the other variables.

If we know U, then we can perfectly determine V using F.

e.g. SCM 2.2.1 (School Funding, SAT scores, College Acceptance)

$$U = \{U_X, U_Y, U_Z\},\ V = \{X, Y, Z\},\ F = \{f_X, f_Y, f_Z\}$$

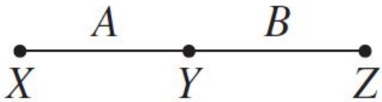
$$f_X : X = U_X,\ f_Y : Y = \frac{x}{3} + U_Y,\ f_Z : Z = \frac{y}{16} + U_Z$$

DAGs

Why **graphs**?

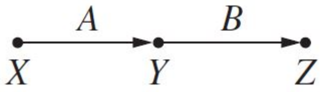
- Graphs help us to capture the probabilistic information visually that is embedded in a SCM.

(Mathematical) graph is a collection of nodes (X, Y, Z) and edges (A, B).

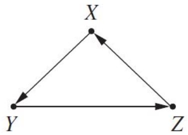


DAGs

Directed Acyclic Graphs (DAGs)



If edges are arrows, then they are **directed**.
When a directed path exists from a node to itself : cyclic



No cycles in a graph : **acyclic**.

- X is a parent (direct cause) of Y, Z is a child of Y.
- X and Y are ancestors of Z, Y and Z are descendants of X.

Product Decomposition

$$P(x_1, x_2, ..., x_n) = \prod_i P(x_i|pa_i)$$

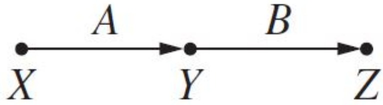
where pa_i stands for the values of parents of variable X_i.

Causal Markov condition

A variable is conditionally independent of its non-descendants given its parent variables.

Product Decomposition

e.g.



$$P(X = x, Y = y, Z = z)$$
$$= P(X = x)P(Y = y|X = x)P(Z = z|Y = y)$$

Graphical Rules (Chains, Forks, Colliders)

d-separation

Chains

The configuration of variables —
three nodes and two edges, with one edge directed into and one edge directed out of the middle variable — is called a **chain**.

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

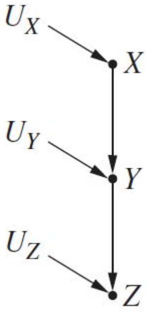
$$P(X,Y,Z) = P(X)P(Y|X)P(Z|X,Y) \qquad \text{(Bayes')}$$

$$\implies P(Z=z|Y=y) = P(Z=z|X=x,Y=y) \;\; \forall x,y,z$$

Thus Z and X are independent, conditional on Y

Chains

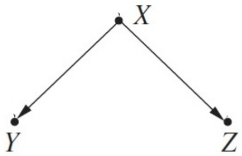
Rule 1 (Conditional Independence in Chains)
Two variables, X and Z, are conditionally independent given Y, if there is only one unidirectional path between X and Z and Y is any set of variables that intercepts that path.



Z and X are independent, conditional on Y
For all $x, y, z, P(Z = z|X = x, Y = y) = P(Z = z|Y = y)$

Forks

The configuration of variables —
three nodes with two arrows emanating from the middle variable —
is called a **fork**.



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

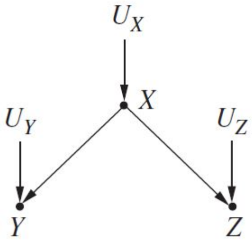
$$P(X, Y, Z) = P(X)P(Y|X)P(Z|X, Y) \qquad \text{(Bayes')}$$

$$\implies P(Z = z|X = x) = P(Z = z|X = x, Y = y) \;\; \forall x, y, z$$

Thus Z and Y are independent, conditional on X

Forks

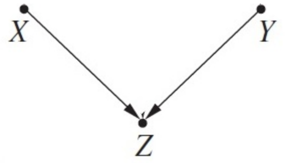
Rule 2 (Conditional Independence in Forks)
If a variable X is a common cause of variables Y and Z,
and there is only one path between Y and Z,
then Y and Z are independent conditional on X.



Y and Z are independent, conditional on X
For all $x, y, z, P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

Colliders

The configuration contains a **collider** node, if one node receives edges from two other nodes.



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

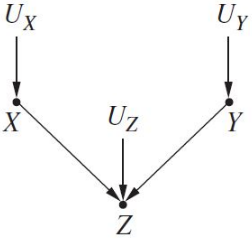
$$P(X, Y, Z) = P(X)P(Y|X)P(Z|X, Y) \qquad \text{(Bayes')}$$

$$\implies P(Y = y) = P(Y = y|X = x) \ \ \forall x, y, z$$

Thus X and Y are independent.

Colliders

Rule 3 (Conditional Independence in Colliders)
If a variable Z is the collider node between two variables X and Y, and there is only one path between X and Y, then X and Y are unconditionally independent but are dependent conditional on Z and any descendants of Z.



X and Y are dependent, conditional on Z
For some $x, y, z, P(Y = y|X = x, Z = z) \neq P(Y = y|Z = z)$
e.g. $Z = X + Y$

d-separation

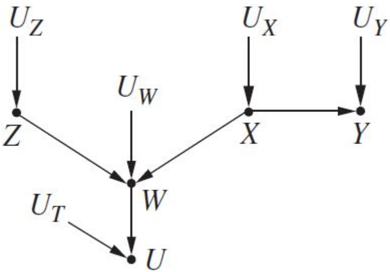
Definition (A blocked path)

A path p is blocked by a variable B

if and only if

p contains

1. a **chain** $A \to B \to C$ or a **fork** $A \leftarrow B \to C$ such that the middle node B is conditioned on, or
2. a **collider** $A \to B \leftarrow C$ such that the collision node B is not conditioned, and no descendant of B is conditioned.



d-separation

- Two nodes A and B are

$$\begin{cases} d\text{-separated iff every path between them is } \textit{blocked}. \\ d\text{-connected iff even one path between them is } \textit{unblocked}. \end{cases}$$

Remark

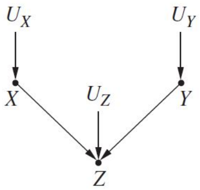
If X and Y are *d*-separated conditional on Z,
then X is statistically independent of Y given Z.

Model Testing and Causal Search

- How to test models locally?

e.g.

Suppose we believe
S (a data set) might have generated G (a graph; a model).



G : Two variables X and Y are independent conditional on Z.
S : No, they are not.
$\rightarrow$ Reject G as a possible causal model for S.