Machine Learning

Directed Acyclic Graph (DAG)

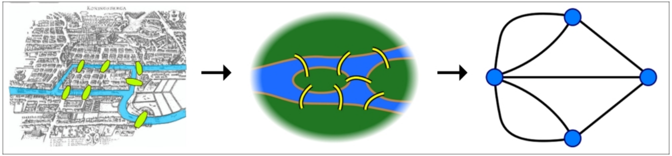
Kia Nazarpour

Based on original slides by Hang Tao

Context

Directed Acyclic Graph (DAGs) are used to encode researchers' *a priori* assumptions about the relationships between and among variables in causal structures.

1. More expressive than mathematical representation?
2. Enable clear communication
3. Inform us about how to avoid bias due to confounding



Learning Outcomes

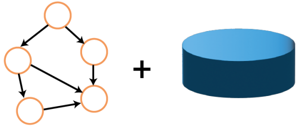
1. Acknowledge the key motivations behind the use of DAGs
2. Remember the notion of statistical independence in the context of DAGs
3. Able to write the chain rule for conditional probabilitie
4. Able to draw and interpret a simple DAG

References:

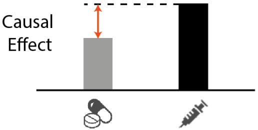
1. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2008. (Section 9.1)

Causal Machine Learning

Causal Reasoning

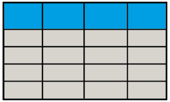


Given by Experts or Causal Discovery + **Observed data**



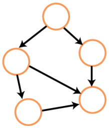
Goal: To quantify the effect of an intervention

Causal Discovery



Structured data

Causal Graph

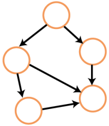


Goal: To determine the causal structure

Causal Representation Learning

Unstructured data

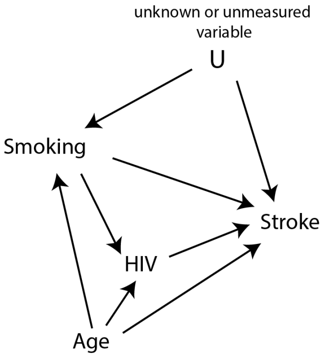
Causal variable



Goal: To find relevant variable to solve causal tasks

DAG: Key concepts

A DAG is a model of causality.

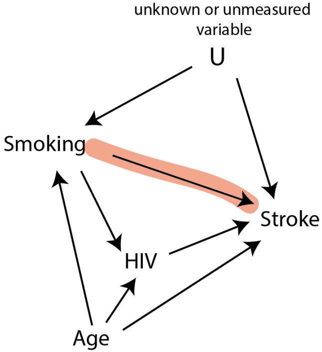


Am J Epidemiol, 177(4):292–298

DAG: Key concepts

A DAG is a model of causality.

Example: What is the (total) effect of smoking on stroke risk?



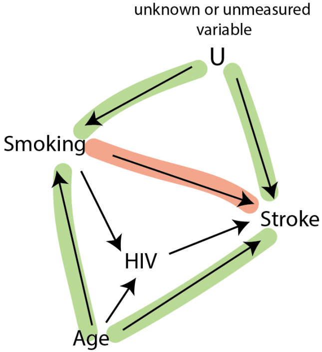
Am J Epidemiol, 177(4):292–298

DAG: Key concepts

A DAG is a model of causality.

Example: What is the (total) effect of smoking on stroke risk?

> Age and U are confounders - control for their effects



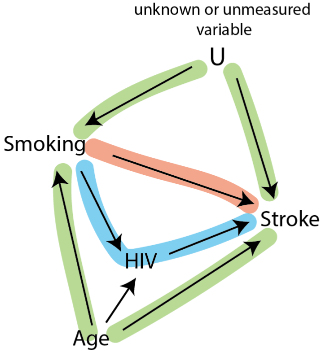
Am J Epidemiol, 177(4):292–298

DAG: Key concepts

A DAG is a model of causality.

Example: What is the (total) effect of smoking on stroke risk?

> Age and U are confounders - control for their effects
>
> HIV is a mediator - do not control



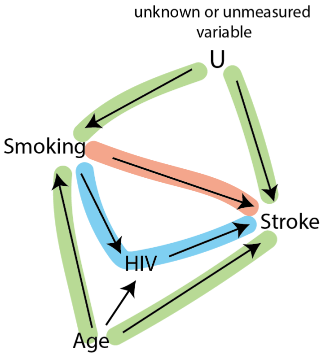
Am J Epidemiol, 177(4):292–298

DAG: Key concepts

A DAG is a model of causality.

Example: What is the (total) effect of smoking on stroke risk?

> Age and U are confounders - control for their effects
>
> HIV is a mediator - do not control



Am J Epidemiol, 177(4):292–298

Treatment Effect: What will be the outcome if a patient is given a particular treatment? (causal reasoning)

Statistical Independence of two variables

- Two variables x and y are independent if

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

- Equivalently, two variables x and y are independent if

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

- We will use $x \perp y$ to denote the independence of x and y.

Statistical Independence of many variables

- If $\{x_1, \ldots, x_n\} \perp \{y_1, \ldots, y_m\}$ then

$$p(x_1, \ldots, x_n, y_1, \ldots, y_m) = p(x_1, \ldots, x_n)p(y_1, \ldots, y_m)$$

- Independence implies factorisation.

- For example, suppose $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$. If $\{x, y\} \perp z$,

$$p(x, y, z) = p(x, y)p(z).$$

Statistical Independence of many variables

- If $\{x_1, \ldots, x_n\} \perp \{y_1, \ldots, y_m\}$ then

$$p(x_1, \ldots, x_n, y_1, \ldots, y_m) = p(x_1, \ldots, x_n) p(y_1, \ldots, y_m)$$

- Independence implies factorisation.

- For example, suppose $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $z \in \mathcal{Z}$. If $\{x, y\} \perp z$,

$$p(x, y, z) = p(x, y) p(z).$$

- The original domain is $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, but after factorisation, the domain we need to consider, $\mathcal{X} \times \mathcal{Y}$ and $\mathcal{Z}$, is much smaller than $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$.

Mutual independence vs pairwise independence

- The variables x_1, x_2, x_3 are mutually independent if

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

- If $x_1 \perp x_2$, $x_2 \perp x_3$, and $x_1 \perp x_3$, then x_1, x_2, x_3 are pairwise independent.

- Mutual independence implies pairwise independence, but the converse is not necessarily true.

Conditional independence

- The variables x and y are conditionally independent given z if

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

- In this case, we write $(x \perp y) \mid z$.

- The sets of variables $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_m\}$ are conditionally independent given $\{z_1, \dots, z_t\}$ if

$$\begin{aligned} p\big((x_1, \dots, x_n, y_1, \dots, y_m)|z_1, \dots, z_t\big) \\ = p(x_1, \dots, x_n|z_1, \dots, z_t)p(y_1, \dots, y_m|z_1, \dots, z_t). \end{aligned} \tag{1}$$

Testing independence

- By definition of marginalisation,

$$p(x|z) = \sum_y p(x, y|z)$$
$$p(y|z) = \sum_x p(x, y|z)$$

- Check if

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

for all x, y, and z.

- The above algorithm is slow. In general, testing independence is a hard problem.

“Chain rule” of conditional probabilities

- Any joint probability $p(x_1, x_2, \ldots, x_n)$ can be factorised in any order.
- Not relying on independence and true for any distribution
- For example,

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \ldots, x_{n-1}).$$

- Or

$$p(x_1, x_2, \ldots, x_n) = p(x_n)p(x_{n-1}|x_n)p(x_{n-2}|x_{n-1}, x_n) \cdots p(x_1|x_2, \ldots, x_n).$$

Proof - "Chain rule" of conditional probabilities

$$
\begin{aligned}
p(x_1, x_2, \ldots, x_n) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_{n-1}|x_1, \ldots, x_{n-2})p(x_n|x_1, \ldots, x_{n-1}) \\
&= p(x_1)\frac{p(x_2, x_1)}{p(x_1)}p(x_3|x_1, x_2)\cdots p(x_{n-1}|x_1, \ldots, x_{n-2})p(x_n|x_1, \ldots, x_{n-1})
\end{aligned}
$$

Proof - "Chain rule" of conditional probabilities

$$
\begin{aligned}
p(x_1, x_2, \ldots, x_n) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_{n-1}|x_1, \ldots, x_{n-2})p(x_n|x_1, \ldots, x_{n-1}) \\
&= p(x_1)\frac{p(x_2, x_1)}{p(x_1)}p(x_3|x_1, x_2)\cdots p(x_{n-1}|x_1, \ldots, x_{n-2})p(x_n|x_1, \ldots, x_{n-1}) \\
&= p(x_1)\frac{p(x_2, x_1)}{p(x_1)}\frac{p(x_3, x_2, x_1)}{p(x_2, x_1)}\cdots\frac{p(x_{n-1}, x_1, \ldots, x_{n-2})}{p(x_{n-2}, x_{n-3}, \ldots, x_1)}\frac{p(x_n, x_1, \ldots, x_{n-1})}{p(x_{n-1}, x_{n-2}, \ldots, x_1)}
\end{aligned}
$$

Applying independence

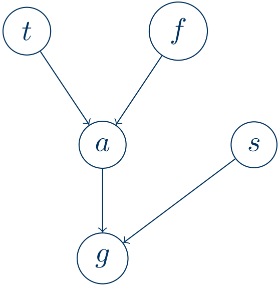
- Every Thursday there is an alarm testing (t).
- The alarm (a) goes off when there is fire (f).
- If the alarm goes off, people in the building should meet at the front door (g) on the ground floor.
- People gathers in front the building when there is a strike (s).

Applying independence

- Alarm testing is independent of a fire ($t \perp f$).
- A strike is independent of what happens in the building ($s \perp \{a, f, t\}$).
- People gathering is independent of fire and alarm testing if we know whether the alarm goes off or whether there is a strike ($g \perp \{f, t\} \mid s, a$).
- Combining the above, we have

$$\begin{aligned} p(a, t, f, s, g) &= p(t)p(f|t)p(a|f, t)p(s|a, f, t)p(g|s, a, f, t) \\ &= p(t)p(f)p(a|f, t)p(s)p(g|s, a) \end{aligned}$$

A (directed) graph representation



$$p(a, t, f, s, g) = p(t)p(f)p(a|f, t)p(s)p(g|s, a)$$

A (directed) graph representation

- Each vertex is a variable.
- A parent has edges pointing from itself to its children.
- The graph is directed and acyclic.
- A distribution factorises according to a graph if

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

- Instead of describing independencies, the graph describes a factorisation.

Two objects

- Graph
- Probability distribution
  - A probability distribution has a set of independencies.
  - A probability distribution can factorise according to a graph.
- Can we read off independencies from a graph?

Basic structures

chain	$x \rightarrow y \rightarrow z$	$x \perp z \mid y$
common cause	$x \leftarrow y \rightarrow z$	$x \perp z \mid y$
v-structure	$x \rightarrow y \leftarrow z$	$x \perp z$ $x \not\perp z \mid y$

Basic structures - Chain

chain $\quad x \rightarrow y \rightarrow z \quad x \perp z \mid y$

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

Basic structures - Common Cause

common cause $\quad x \leftarrow y \rightarrow z \quad$ $x \perp z \mid y$

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(y)p(x|y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

Basic structures - v-structure

v-structure

$$x \rightarrow y \leftarrow z$$

$$x \perp z$$
$$x \not\perp z \mid y$$

$$p(x,z) = \sum_y p(x,y,z) = \sum_y p(x)p(z)p(y|x,y) = p(x)p(z)$$

Basic structures - v-structure

v-structure $\quad x \rightarrow y \leftarrow z \quad$ $\begin{array}{c} x \perp z \\ x \not\perp z \mid y \end{array}$

If $x \perp z \mid y$,

$$
\begin{aligned}
p(x, z) &= \sum_y p(x, y, z) = \sum_y p(y) p(x, z|y) = \sum_y p(y) p(x|y) p(z|y) \\
&= \sum_y p(x|y) p(y, z).
\end{aligned}
$$

But

$$
p(x, z) = p(x) p(z) = \sum_y p(x) p(y, z).
$$

This can hold only when $p(x|y) = p(x)$, but x and y are not independent; a contradiction.