Causality

Talk 3: Causal DAG Models (Causal Bayesian networks)

Yao Zhang

Intelligent Information Processing Research Group,

Faculty of Electrical Engineering and Computer Science,

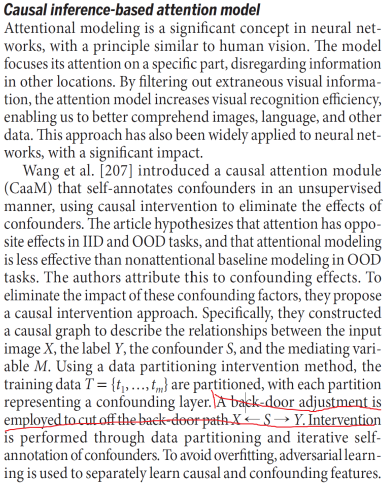
Ningbo University

Note: The following slides are primarily adapted from the course materials[1].

Nov 28, 2025

[1] C. Heinze-Deml. Causality. URL: https://stat.ethz.ch/lectures/ss21/causality.php.

Task A



Causal inference-based attention model

Attentional modeling is a significant concept in neural networks, with a principle similar to human vision. The model focuses its attention on a specific part, disregarding information in other locations. By filtering out extraneous visual information, the attention model increases visual recognition efficiency, enabling us to better comprehend images, language, and other data. This approach has also been widely applied to neural networks, with a significant impact.

Wang et al. [207] introduced a causal attention module (CaaM) that self-annotates confounders in an unsupervised manner, using causal intervention to eliminate the effects of confounders. The article hypothesizes that attention has opposite effects in IID and OOD tasks, and that attentional modeling is less effective than nonattentional baseline modeling in OOD tasks. The authors attribute this to confounding effects. To eliminate the impact of these confounding factors, they propose a causal intervention approach. Specifically, they constructed a causal graph to describe the relationships between the input image X, the label Y, the confounder S, and the mediating variable M. Using a data partitioning intervention method, the training data $T = \{t_1, \ldots, t_m\}$ are partitioned, with each partition representing a confounding layer. A back-door adjustment is employed to cut off the back-door path $X \leftarrow S \rightarrow Y$. Intervention is performed through data partitioning and iterative self-annotation of confounders. To avoid overfitting, adversarial learning is used to separately learn causal and confounding features.

Figure 1: Causal inference-based attention model in the article[2].

[2] L. Jiao et al. "Causal Inference Meets Deep Learning: A Comprehensive Survey". In: Research 7 (2024), pp. 1–41.

ETH zürich

Factorization of the joint density

- We always have:
$$f(x_1, \dots, x_p) = f(x_1) f(x_2|x_1) \dots f(x_p|x_1, \dots x_{p-1})$$
- A set of variables $X_{\mathrm{pa}(j)}$ is said to be Markovian parents of X_j if it is a minimal subset of $\{X_1, \dots X_{j-1}\}$ such that $f(x_j|x_1, \dots, x_{j-1}) = f(x_j|x_{\mathrm{pa}(j)})$
- Then
$$f(x_1, \dots, x_p) = \prod_{j=1}^{p} f(x_j|x_{\mathrm{pa}(j)})$$
"factorization property"
- We can draw a DAG accordingly
- The distribution is said to factorize according to this DAG

Last time, Cont.

- First-order Markov models: the future is independent of the past given the present

$$1 \to 2 \to \cdots \to t-1 \to t \to t+1$$

$$X_{t+1} \perp\!\!\!\perp \{X_1, X_2, \ldots, X_{t-1}\} \mid X_t$$

- In DAG models, we have a similar (local) Markov property
- Let S be any collection of nodes. Then:

$$X_S \perp\!\!\!\perp X_{V \setminus \mathrm{desc}(S) \setminus \mathrm{pa}(S)} \mid X_{\mathrm{pa}(S)} \tag{1}$$

When there are three random variables forming a collider structure (i.e., $X \to Z \leftarrow Y$), $X \perp\!\!\!\perp Y \mid Z$?
In more general cases, we need d-separation.

ETH zürich

d-separation

- A path between i to j is blocked by a set $\boldsymbol{S}$ (not containing i or j) if at least one of the following holds:
  - There is a non-collider on the path that is in $\boldsymbol{S}$; or
  - There is a collider on the path such that neither this collider nor any descendants are in $\boldsymbol{S}$
- A path that is not blocked is active

- If all paths between $i \in \boldsymbol{A}$ and $j \in \boldsymbol{B}$ are blocked by $\boldsymbol{S}$, then $\boldsymbol{A}$ and $\boldsymbol{B}$ are d-separated by $\boldsymbol{S}$. Otherwise they are d-connected given $\boldsymbol{S}$.

- Denote d-separation by $\perp$

ETH zürich

Global Markov property

- Definition:
  A distribution P with density p satisfies the global Markov property with respect to a DAG G if:

  $$\boldsymbol{A} \text{ and } \boldsymbol{B} \text{ are d-separated by } \boldsymbol{S} \text{ in } G \Rightarrow X_{\boldsymbol{A}} \perp\!\!\!\perp X_{\boldsymbol{B}} | X_{\boldsymbol{S}} \text{ in } P$$

- Theorem (Pearl, 1988):
  A distribution P with density p satisfies the global Markov property with respect to G if and only if p factorizes according to G.

ETH zürich

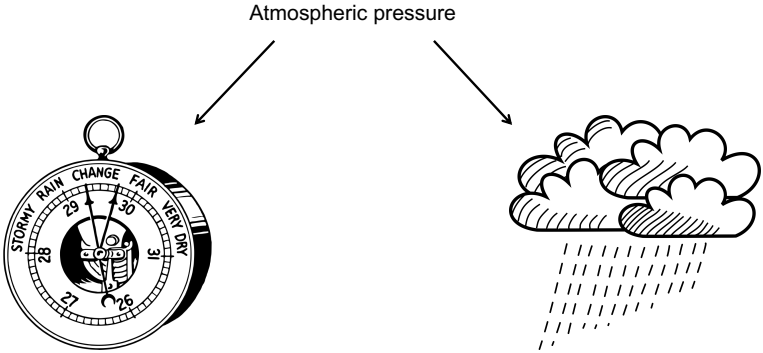
Causal graphical models

Causality
Christina Heinze-Deml
Spring 2021

Today

- Causal effects and do-operator
- Causal graphical models
- Selection bias

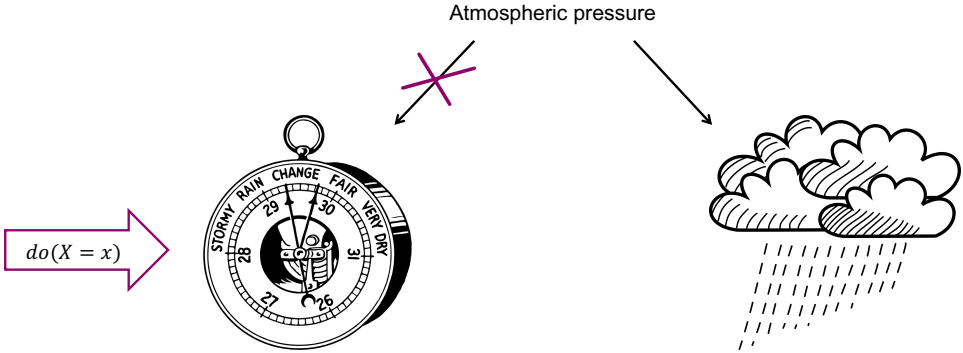
Causal effect – Example



18

Example due to Frederick Eberhardt

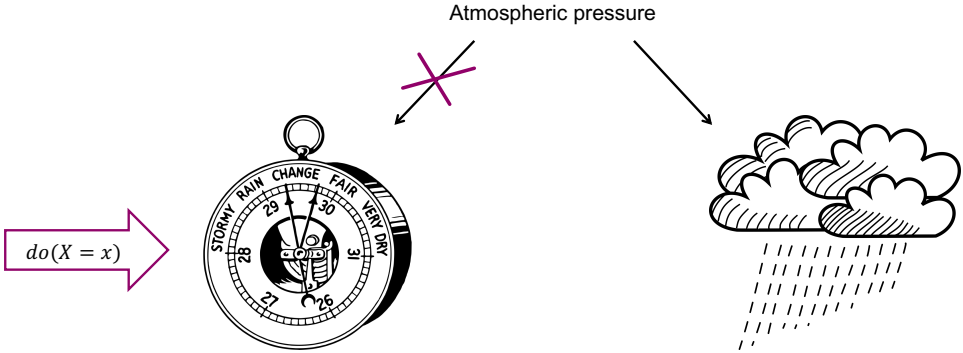
Causal effect – Example



19

Example due to Frederick Eberhardt

Causal effect – Example



X has a causal effect on Y if manipulating X changes the distribution of Y

Example due to Frederick Eberhardt

Causal effect and do-operator

- Interventional definition of causal effect:
  X has a causal effect on Y if manipulating X changes the distribution of Y

- Mathematical notion of manipulation (due to Pearl):
  - $do(X = x)$ (or shorthand $do(x)$) represents a hypothetical intervention where X is set to the value x, uniformly over the entire population
  - $p(y|do(X = x))$ is the distribution of Y after $do(X = x)$
  - $E(Y|do(X = x))$ is the expectation of Y after $do(X = x)$, etc

Conditioning on observing:	$p(y\|see(X = x)) = p(y\|x)$ (ordinary conditioning)
Intervening:	$p(y\|do(X = x))$, also written as $p^{do(X=x)}(y)$

Causal effect and do-operator

- Mathematical definition of causal effect:
  X has a causal effect on Y if $p(y|do(X = x'))$ depends on x',
  i.e., if $\exists\ a$ and b so that $p(y|do(X = a)) \neq p(y|do(X = b))$

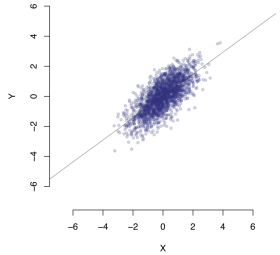
- Total average causal effect:
  $\mathrm{ACE}(x, x') = E(y|do(X = x)) - E(y|do(X = x'))$

Eg X binary, treatment $X = 1$
no treatment $X = 0$

$ACE = E(y|do(x=1)) - E(y|do(x=0))$

Classical regression models

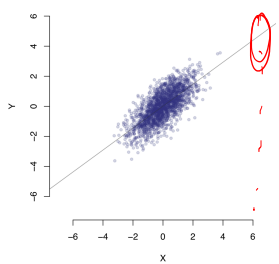
- We observe n i.i.d. observations of (X, Y)
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction

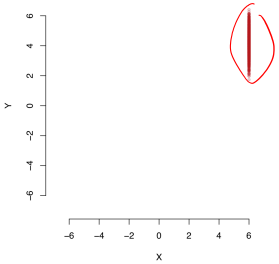


Model aspects of distribution of Y when we observe $X = x$

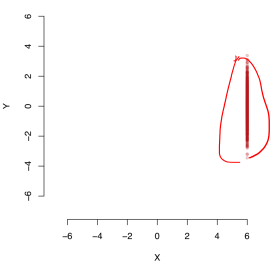
Classical regression models

- We observe n i.i.d. observations of (X, Y)
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction – but what if we set X to e.g. 6? I.e. if $do(X = 6)$?





$X \to Y$



$Y \to X$

Classical regression models

- We observe n i.i.d. observations of (X, Y)
- Goal is to model certain aspects of $p(y|x)$, for example $E(Y|X = x)$
- Useful for prediction

- Such analyses are generally not useful for policy or treatment decisions, since such decisions involve predictions in manipulated systems with post-intervention distributions different from p

Example

- Consider a rehabilitation program for prisoners. Participation in the program is voluntary.
  - $X = 1$ if prisoner participated in the program; $X = 0$ otherwise
  - $Y = 1$ if prisoner is rearrested within a year; $Y = 0$ otherwise

- $\mathrm{P}(Y = 1|X = 1)$: probability of re-arrest for prisoners who choose to participate
- $\mathrm{P}(Y = 1|do(X = 1))$: probability of re-arrest if program were compulsory for all prisoners
- Note that generally $\mathrm{P}(Y = 1|do(X = 1)) \neq \mathrm{P}(Y = 1|X = 1)$

Example

- Suppose $\mathrm{P}(Y = 1|X = 1) < \mathrm{P}(Y = 1|X = 0)$
  - Re-arrest rate among prisoners who participated in the program is lower than among those who did not participate
  - Could be due to the program, due to the intrinsic motivation of the prisoners who chose to participate, due to a mixture of these two, or….

- Suppose $\mathrm{P}(Y = 1|do(X = 1)) < \mathrm{P}(Y = 1|do(X = 0))$
  - Program lowers the re-arrest rate, i.e., program has a causal effect on the re-arrest rate
  - Manipulating X changes the distribution of Y
  - X is causal for Y

Frameworks

- Causal DAG models (Causal Bayesian networks)
- Structural equation models
- Potential outcomes

Causal Bayesian networks

- Let $G = (\boldsymbol{V}, \boldsymbol{E})$ be a DAG and P be the distribution of $X_{\boldsymbol{V}}$ with density p
- The pair (G, P) is a DAG model or a Bayesian network if

$$p(x_{\boldsymbol{V}}) = \prod_{i \in \boldsymbol{V}} p(x_i | x_{\mathrm{pa}(i)})$$

Causal Bayesian networks

- Let $G = (\boldsymbol{V}, \boldsymbol{E})$ be a DAG and P be the distribution of $X_{\boldsymbol{V}}$ with density p
- The pair (G, P) is a DAG model or a Bayesian network if

$$p(x_{\boldsymbol{V}}) = \prod_{i \in \boldsymbol{V}} p(x_i | x_{\mathrm{pa}(i)})$$

- The pair (G, P) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_{\boldsymbol{V}} | do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p\left(x_i \middle| x_{\mathrm{pa}(i)}\right) 1\{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}\}$$

33

Causal Bayesian networks

- The pair (G, P) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_{\boldsymbol{V}} \,|do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p\left(x_i \middle| x_{\mathrm{pa}(i)}\right) 1\{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}\}$$

- Modified factorization known as
  - "g-formula" (Robins)
  - "manipulation formula" (Spirtes, Glymour, Scheines)
  - "truncated factorization formula" (Pearl)

35

Causal Bayesian networks

- The truncated factorization formula implies that an intervention on X_j only changes $p(x_j|x_{\mathrm{pa}(j)})$; the other conditional distributions remain unchanged. This is also known as invariance.

Compare: $p(x_V) = \{\prod_{i \in V \setminus \{j\}} p(x_i \mid x_{pa(i)})\}\, p(x_j \mid x_{pa(j)})$

$p(x_V \mid do(x_j = x'_j)) = \{\prod_{i \in V \setminus \{j\}} p(x_i \mid x_{pa(i)})\} \cdot \mathbb{1}\{x_j = x'_j\}$

Notes week 3 - I

A : Altitude

T : Temperature

$A \rightarrow T$

$p(a,t)$: joint density of the altitude A and the average temperature T

$$p(a,t) = p(t|a)\, p(a)$$

invariant physical mechanism

→ if we change the altitude A, then we assume that the physical mechanism $p(t|a)$ responsible for producing an average temperature is still in place (invariance)

→ holds independent of $p(a)$

Causal Bayesian networks

- The truncated factorization formula implies that an intervention on X_j only changes $p(x_j|x_{\mathrm{pa}(j)})$; the other conditional distributions remain unchanged. This is also known as invariance.

"Interventions act locally"

But note: can have interventions on multiple nodes at the same time

$$p(x_v \mid do(x_j = x_j'), do(x_\ell = x_\ell')) = \prod_{i \in v \setminus \{j,\ell\}} p(x_i \mid x_{pa(i)}) \mathbb{1}\{x_j = x_j'\} \mathbb{1}\{x_\ell = x_\ell'\}$$

40

Causal Bayesian networks

- The pair (G, P) is a causal DAG model or a causal Bayesian network if for any $\boldsymbol{W} \subset \boldsymbol{V}$

$$p(x_{\boldsymbol{V}} \,|do(x_{\boldsymbol{W}} = x'_{\boldsymbol{W}})) = \prod_{i \in \boldsymbol{V} \setminus \boldsymbol{W}} p\left(x_i \middle| x_{\mathrm{pa}(i)}\right) 1\{x_{\boldsymbol{W}} = x'_{\boldsymbol{W}}\}$$

post-intervention distributions needed to define causal effects

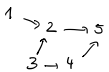
conditional distributions that can be estimated from observational data

Causal Bayesian networks

- The modified factorizations represent factorizations wrt truncated graphs $G_{\overline{W}}$, where all edges into W are removed

- See `Notes week 3 - II.pdf`

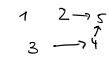
G



$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_3)\,p(x_2|x_1, x_3)\,p(x_4|x_3)\,p(x_5|x_2, x_4)$$

Consider a do-intervention on X_2:

$G_{\bar{2}\underline{2}}$

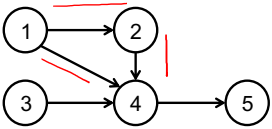


$$p(x_1, x_2, x_3, x_4, x_5 \mid do(x_2 = x_2')) = p(x_1)p(x_3)\,\mathbb{1}\{x_2 = x_2'\} \cdot p(x_4|x_3)\,p(x_5|x_2, x_4)$$

$$p(x_1, x_3, x_4, x_5 \mid do(x_2 = x_2')) = p(x_1)p(x_3)p(x_4|x_3)p(x_5|x_2, x_4)\Big|_{x_2 = x_2'}$$

Causal Bayesian networks

- $X_{\mathrm{pa}(j)}$ can be interpreted as the direct causes of X_j
- Directed edges can be interpreted as direct causal effects



- X can only be causal for Y if there is a directed path from X to Y
- Depending on the context, we might be interested in the
  - Direct causal effect
  - Indirect causal effect
  - Total causal effect
  - …

Notes week 3 - III

Consider the DAGs $1 \to 2$ and $2 \to 1$

* Any distribution P of (X_1, X_2) factorizes wrt these two DAGs:

$$p(x_1, x_2) = p(x_1)\,p(x_2|x_1) = p(x_2)\,p(x_1|x_2)$$

Assume $X_1 \not\perp\!\!\!\perp X_2$

* But the DAGs are very different when interpreted causally:

Consider a do-intervention on X_2.

The post-intervention distribution of X_1 is:

1) For the causal DAG $1 \to 2$

$$p(x_1 \mid do(x_2 = x_2')) = \underbrace{p(x_1)}_{\text{does not depend on } X_2 \text{ (no causal effect from } X_2 \text{ to } X_1)} \Big|_{x_2 = x_2'}$$

$$p(x_1 \mid do(x_2 = x_2')) = p(x_1 \mid do(x_2 = \tilde{x}_2)) = p(x_1)$$

2) For the causal DAG $2 \to 1$

$$p(x_1 \mid do(x_2 = x_2')) = p(x_1 \mid x_2)\Big|_{x_2 = x_2'} = p(x_1 \mid x_2')$$

Causal effect of X_2 on X_1

Here, $p(x_1 \mid do(x_2 = x_2')) \neq p(x_1 \mid do(x_2 = \tilde{x}_2))$, $x_2' \neq \tilde{x}_2$

Simpson's paradox

	Treatment	Placebo
Male	50/100	150/500
Female	50/500	0/100
Total	100/600	150/600

replace gender by blood pressure (BP); numbers stay the same

	Treatment	Placebo
High BP	50/100	150/500
Low BP	50/500	0/100
Total	100/600	150/600

Simpson (1951), in an example similar to this one: "*The treatment can hardly be rejected as valueless to the race when it is beneficial when applied to males and to females.*"

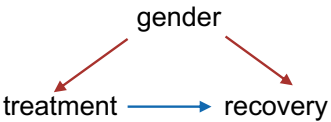
⇒ control for gender, use the treatment

Simpson (1951), in an example similar to this one: "…, *yet it is the combined table which provides what we would call the sensible answer…*"

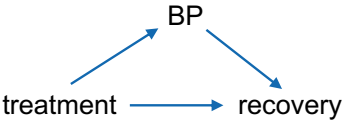
⇒ don't control for BP, don't use the treatment

Simpson's paradox and causal diagrams

- Same numbers, different conclusions….
  - Must use additional information: "story behind the data", causal assumptions
- Consider total causal effect of treatment on recovery
  - Possible scenarios:



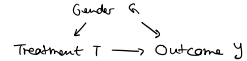
gender is a confounder; control for gender



BP is an intermediate variable; don't control for BP

Notes week 3 - IV

1)



* Observational distribution:

$$P(G, T, Y) = P(G)\,P(T \mid G)\,P(Y \mid T, G)$$

* Interventional distribution under $do(T=1)$

("force all people to take the treatment")

$$P(G, T, Y \mid do(T=1)) = P(G)\,\mathbb{1}_{T=1}\,P(Y \mid T, G)$$

Marginalize:

$$P(G, Y \mid do(T=1)) = P(G)\,P(Y \mid G, T=1)$$

$$P(Y \mid do(T=1)) = \sum_g P(G=g)\,P(Y \mid G=g, T=1)$$

both parts known from observation; conditional on gender

2)

* Observational distribution:

$$P(Y, T, B) = P(T)\,P(B \mid T)\,P(Y \mid B, T)$$

* Jnt. distribution:

$$P(y, \tau, B \mid do(\tau=1)) = \mathbf{1}\{\tau=1\}\, P(B \mid \tau)\, P(y \mid B, \tau)$$

Marginalize:

$$P(y \mid do(\tau=1)) = \sum_b P(B=b \mid \tau=1)\, P(y \mid B=b, \tau=1)$$

$$= P(y \mid \tau=1)$$

also known from observation — unconditional wrt B

Causal DAGs

- Causal DAGs imply strong assumptions, allowing us to estimate post-intervention distributions from observational data

- How do we know the causal DAG?
  - Now: assume it is given, e.g. from background knowledge
  - Later: consider learning causal DAG (under some assumptions)
  - In any case, causal DAG provides clear framework to state causal assumptions for analysis
    - Allows for an honest debate about such assumptions
    - Can draw several possible causal DAGs, conduct the analysis for each of them and perform a sensitivity analysis

Discussion

Any comments or questions?

We may not always find an answer, and since we're not very familiar with causality, we will need to dedicate more time to this topic.