Causality

Pedro A. Ortega

Computational & Biological Learning Lab
University of Cambridge

18th February 2010

Why is causality important?

The future of machine learning is to **control** (the world).

Examples

- Classical example:

  "Do smokers get lung cancer?" versus "Do smokers have lung cancer?"

- Programming:

$$y \leftarrow f(x) \quad \text{versus} \quad y = f(x).$$

- Physics:

$$a \leftarrow \frac{F}{m} \qquad \text{versus} \qquad F = ma.$$

- Statistics is about measuring **correlation** of events.
- Causality is about the **functional dependency** of events.
- Most of science is driven by the need of **causal** understanding.

Why is causality . . .

. . . easy?

- It is intuitive: we reason in causal terms.
- Statistics can deal with it (given the right assumptions).

. . . difficult?

- Confounders impede the isolation of the functional dependency of interest.
- The concepts of causation are not fully formalized.
- Because it behaves like conditional probabilities under certain circumstances; in fact quite often because we tend to model causally!

Current status

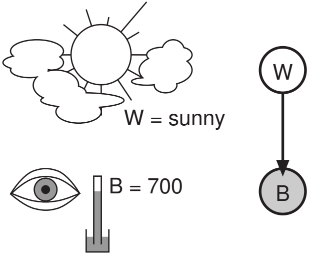
- Historically studied by many philosophers (e.g. Hume).
- Banned from statistical vocabulary at the beginning of the 20th century (Pearson, Russell, . . .).
- Exception: Randomized controlled trial (Fisher?).

Today, still in infancy state:

- Significant progress in causal understanding at beginning of the 90's.
- No consensus in formalization of causal notions.
- Many good (but confusing and mutually inconsistent) formalizations (Pearl, Spirtes, Shafer, Dawid, . . .).
- No measure-theoretic formalization.
- But we are slowly getting there!
- Compare to the history of probability!

Barometer example

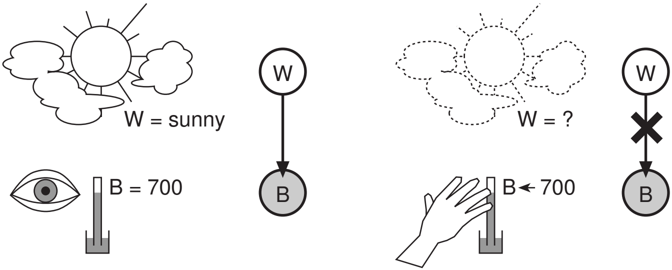
A barometer allows predicting the weather.



- If we **read** B, then can infer W. (Observation)

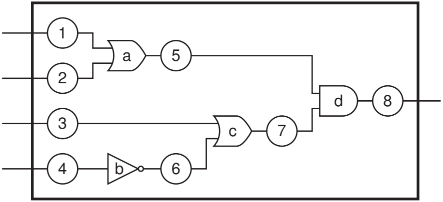
Barometer example

A barometer allows predicting the weather.



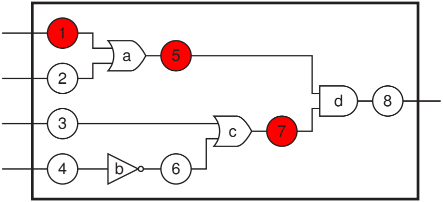
- If we **read** B, then can infer W. (Observation)
- If we **set** B, then we cannot infer W. (Intervention)
- We have to distinguish between **seeing** and **doing**.

Seeing versus doing



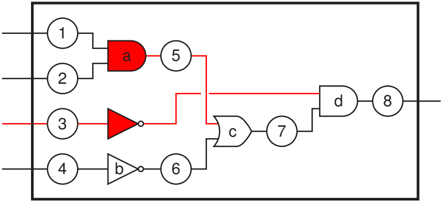
- Assume a circuit connecting **observable** quantities.
- Circuit represents a system **embedded** in Nature.
- Nature & system **determine** values of observable quantities.
- **No control** over the inputs $\Rightarrow$ uncertainty.
- Statistician can act only **inside** of the system.

Seeing versus doing



- **Seeing = Observing = Measuring**.
- **Seeing** is the act of recording the value of observable quantities.
- **Seeing** is passive: the causal flow is undisturbed.
- Collected data allows constructing a truth table.

Seeing versus doing

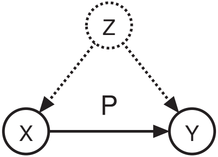


- **Doing** = **Manipulating** = **Intervening**.
- **Doing** is the act of changing the functional dependency amongst observable quantities.
- **Doing** is active: the causal flow is disturbed.
- Knowing the blueprint is crucial to predict the resulting functional dependencies after interventions.

The essence of causal discovery

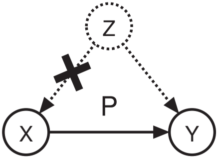
- How does X affect Y?
- Collect data $\Longrightarrow$ obtain $P(X, Y) \Longrightarrow$ compute $P(Y|X)$?

The essence of causal discovery



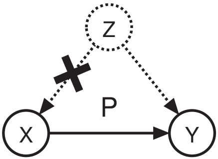
- How does X affect Y? ($\leftarrow$ What does this even mean?)
- Collect data $\Longrightarrow$ obtain $P(X, Y)$ $\Longrightarrow$ compute $P(Y|X)$? **No!**
- **There might be a confounder!** What do we do now?

The essence of causal discovery



- How does X affect Y? ($\leftarrow$ What does this even mean?)
- Collect data $\implies$ obtain $P(X, Y) \implies$ compute $P(Y|X)$? **No!**
- **There might be a confounder!** What do we do now?
- Idea: decouple X from confounders.
- How: manipulate $X \implies$ intervene P (e.g. randomization).

The essence of causal discovery



- How does X affect Y? ($\leftarrow$ What does this even mean?)
- Collect data $\Longrightarrow$ obtain $P(X, Y)$ $\Longrightarrow$ compute $P(Y|X)$? **No!**
- **There might be a confounder!** What do we do now?
- Idea: decouple X from confounders.
- How: manipulate X $\Longrightarrow$ intervene P (e.g. randomization).
- But: Now P has changed into (say) Q!

Intervention of a probability distribution 1

Problem:

- If the intervention transforms P into Q, how can we ever say something about P using Q?
- Under invariance assumptions, we can!

Intervention of a probability distribution 2

Example:

1. Determine the "blueprint",

$$\begin{aligned} P(X,Y,Z) &= P(X)P(Y|X)P(Z|X,Y) \\ &= P(X)P(Y|X,Z)P(Z|X) \\ &= P(X|Y)P(Y)P(Z|X,Y) \\ &= P(X|Y,Z)P(Y)P(Z|Y) \\ &= P(X|Z)P(Y|X,Z)P(Z) \leftarrow \quad \text{(causal decomposition)} \\ &= P(X|Y,Z)P(Y|X)P(Z) \end{aligned}$$

2. Replace $P(X|Z)$ by $Q(X)$:

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

3. Collect data from $Q(X,Y,Z)$ and compute $Q(Y|X)$.

Intervention of a probability distribution 3

What have we achieved?

- Note that $Q(Y|X) \neq P(Y|X)$.
- By decoupling X from Z, we have **isolated the functional dependency** mapping X into Y.
- $Q(Y|X)$ reflects the right dependency, whereas $P(Y|X)$ doesn't!
- Analogy: we cannot understand the effect of X on Y in

$$Y \leftarrow f(X, Z)$$

if $X \leftarrow g(Z)$ in the collected data, because

$$Y \leftarrow f(X, g^{-1}(X)) = h(X),$$

and $h(X) \neq f(X, Z)$!

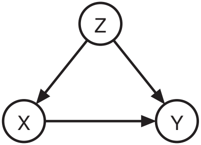
Stop.

Formalizations of causal inference 1

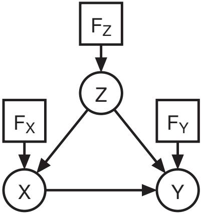
A non-exhaustive list:

- Pearl:
  - structural equations
  - represented in DAGs with causal meaning
  - do-calculus
- Dawid:
  - Augmented DAGs (influence diagrams)
  - decision variables determine regime of operation
- Shafer:
  - Probability tree
  - Moivrean events (sets of leaves) (=measure-theoretic events)
  - Humean events (sets of edges) (transformations)

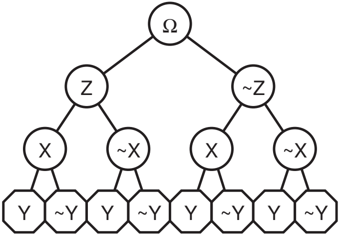
Formalizations of causal inference 2



Causal DAG
(Pearl)



Augmented DAG
(Dawid)



Probability Tree
(Shafer)

Causality based on structural equations (Pearl)

Description

- Causal theory specifies:
  1. functional dependencies,
  2. probability distribution.
- Probabilities can be conditioned in two ways:
  1. evidential (Bayesian): $P(Y|X = x)$;
  2. interventional (causal): $P(Y|\mathrm{do}(X = x))$.

Causal theory

- $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ (observed variables)
- $\mathcal{U} = \{U_1, U_2, \ldots, U_m\}$ (unobserved variables)
- $P(\mathcal{U})$ (prob. over unobserved variables)
- $\mathcal{F} = \{X_i = f_i(\mathcal{X},\mathcal{U})\}_{i=1}^{n}$ (inducing partial order over $\mathcal{X}$)

Causality based on structural equations (Pearl)

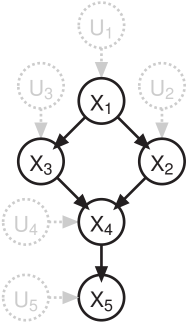
Description

- Causal theory specifies:
  1. functional dependencies,
  2. probability distribution.
- Probabilities can be conditioned in two ways:
  1. evidential (Bayesian): $P(Y|X = x)$;
  2. interventional (causal): $P(Y|\mathrm{do}(X = x))$.

Causal theory

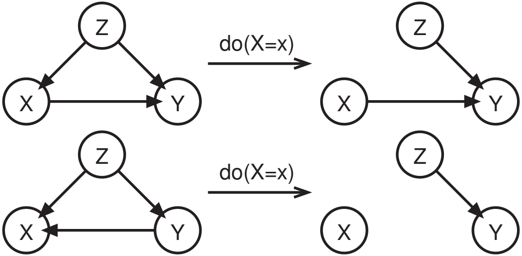
- $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$ (observed variables)
- $\mathcal{U} = \{U_1, U_2, \ldots, U_m\}$ (unobserved variables)
- $P(\mathcal{U})$ (prob. over unobserved variables)
- $\mathcal{F} = \{X_i = f_i(\mathcal{X}, \mathcal{U})\}_{i=1}^{n}$ (inducing partial order over $\mathcal{X}$)
- A causal theory can be represented as a DAG.

Example causal theory (Pearl)



$$
\begin{aligned}
X_1 &= f_1(U_1) \\
X_2 &= f_2(X_1, U_2) \\
X_3 &= f_3(X_1, U_3) \\
X_4 &= f_4(X_2, X_3, U_4) \\
X_5 &= f_5(X_4, U_5)
\end{aligned}
$$

The do-operator (Pearl)



- Handy notation for interventions that mimicks conditions.
- do($X = x$) means "replace the equation for X by $X = x$".
- do($X = x$) corresponds to $Q(X) = \delta_x(X)$.
- Easy graphical interpretation (remove parent links).

Can we infer causal relations from observations?

- "To find out what happens
  if you kick the system,
  you have to kick the system."
- Experiment is impossible or too costly.
- E.g. can we replace $P(Y|\mathrm{do}(X = x))$ by $P(Y|X = x)$?
- Calculus to manipulate expressions with do-operations.

Can we infer causal relations from observations?

- "To find out what happens
  if you kick the system,
  you have to kick the system."
- Experiment is impossible or too costly.
- E.g. can we replace $P(Y|\text{do}(X = x))$ by $P(Y|X = x)$?
- Calculus to manipulate expressions with do-operations.
- Do-calculus
- complete

Do-calculus (Pearl)

Let G be the causal DAG representing a causal theory.

Rules

- Insertion/deletion of observations:

$$P(y|\text{do}(x), z, w) = P(y|\text{do}(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$$

- Action/observation exchange:

$$P(y|\text{do}(x), \text{do}(z), w) = P(y|\text{do}(x), z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X},\underline{Z}}}$$

- Insertion/deletion of actions:

$$P(y|\text{do}(x), \text{do}(z), w) = P(y|\text{do}(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X},\overline{Z(W)}}}$$

where $Z(W)$ are Z-nodes not ancestors of W-nodes in $G_{\overline{X}}$.

Simpson's paradox

Two different recommendations with same data!

- Males and females take drug, then check recovery rate.

	Drug	No-drug
Males	18/30 (60%)	7/10 (70%)
Females	2/10 (20%)	9/30 (30%)
Totals	20/40 (50%)	16/40 (40%)

Simpson's paradox

Two different recommendations with same data!

- Males and females take drug, then check recovery rate.

	Drug	No-drug
Males	18/30 (60%)	7/10 (70%)
Females	2/10 (20%)	9/30 (30%)
Totals	20/40 (50%)	16/40 (40%)

- Patients take drug, blood pressure is measured, then check recovery rate.

	Drug	No-drug
High	18/30 (60%)	7/10 (70%)
Low	2/10 (20%)	9/30 (30%)
Totals	20/40 (50%)	16/40 (40%)

Simpson's paradox

Two different recommendations with same data!

- Males and females take drug, then check recovery rate.

	Drug	No-drug
Males	18/30 (60%)	7/10 (70%)
Females	2/10 (20%)	9/30 (30%)
Totals	20/40 (50%)	16/40 (40%)

- Patients take drug, blood pressure is measured, then check recovery rate.

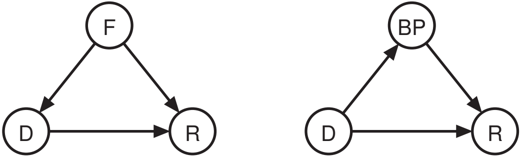
	Drug	No-drug
High	18/30 (60%)	7/10 (70%)
Low	2/10 (20%)	9/30 (30%)
Totals	20/40 (50%)	16/40 (40%)

- First case: consult separate tables.
- Second case: consult aggregated table.

Simpson's paradox 2

Why?

- There is a confounder!
- The correct probability to compute is $P(R|\text{do}(D))$.
- The two cases have different causal models.
- For the second case: $P(R|\text{do}(D)) = P(R|D)$.



Simpson's paradox 3

1. Assumptions:

$$P(R|\text{do}(D), F) < P(R|\text{do}(\neg D), F)$$
$$P(R|\text{do}(D), \neg F) < P(R|\text{do}(\neg D), \neg F)$$

2. From intervened graph:

$$P(F|\text{do}(D)) = P(F|\text{do}(\neg D)) = P(F)$$

3. Calculating:

$$\begin{aligned} P(R|\text{do}(D)) &= P(R|\text{do}(D), F)P(F|\text{do}(D)) + P(R|\text{do}(D), \neg F)P(\neg F|\text{do}(D)) \\ &= P(R|\text{do}(D), F)P(F) + P(R|\text{do}(D), \neg F)P(\neg F) \\ P(R|\text{do}(\neg D)) &= P(R|\text{do}(\neg D), F)P(F) + P(R|\text{do}(\neg D), \neg F)P(\neg F) \end{aligned}$$

4. Using the assumptions:

$$P(R|\text{do}(D)) < P(R|\text{do}(\neg D)).$$

Conclusions

- Causality is about **functional dependencies**.
- Understanding functional dependencies
  is essential for **control**.
- **Ask the right question**:
  correlation or functional dependency?
- Key operation to isolate functional dependencies:
  decoupling of control variables (doing).
- There are causal formalisms that work in practice!

Questions?