732A96/TDDE15 Advanced Machine Learning

Graphical Models

Jose M. Peña
IDA, Linköping University, Sweden

Lectures 5: Causal Inference

Contents

- Causal Inference
- Causal Models
- Interventions
- Truncated Factorization
- Causal Effect Identifiability
- Back-Door Criterion
- Front-Door Criterion
- *do*-Calculus

Literature

- Main source
  - Pearl, J. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press, 2009. Chapters 1-3.
- Additional source
  - Pearl, J. *Causality: Models, Reasoning, and Inference* (1st ed.). Cambridge University Press, 2000. Chapters 1-3.
  - Pearl, J. *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press, 2009. Epilogue chapter.

Causal Inference

- We want to compute the causal effect of an **intervention**, e.g.

$$p(cholesterol|\boldsymbol{do}(exercise)).$$

- **Intervention**: Fixing the value of a variable (for the whole population) so that it is no longer governed by its natural causes.
- **Observation**: Focus on the subpopulation that attains a particular value for a variable, e.g.

$$p(cholesterol|exercise).$$

- **Randomized controlled trials**: Gold standard for assessing causal effects, but they are not always feasible, e.g. the treatment/intervention may be too costly or prohibited due to ethical considerations.
- Can we compute causal effects from **observational** data and, thus, **without** performing interventions ? Yes, but not always.

Causal Inference

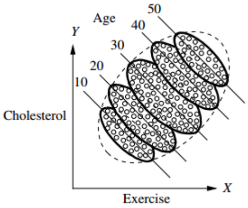


Figure 1.1: Results of the exercise-cholesterol study, segregated by age

- $p(cholesterol|\mathbf{do}(exercise)) = f(p(cholesterol, exercise, age))$?
- $p(cholesterol|\mathbf{do}(exercise)) = p(cholesterol|exercise)$?

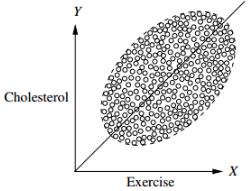


Figure 1.2: Results of the exercise-cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

Causal Inference

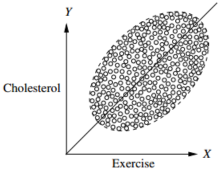


Figure 1.2: Results of the exercise-cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

- Due to the **confounder** Age,

$$p(cholesterol|\mathbf{do}(exercise)) \neq p(cholesterol|exercise).$$

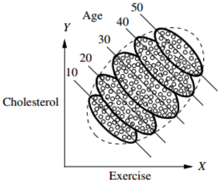
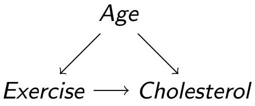


Figure 1.1: Results of the exercise-cholesterol study, segregated by age



- Instead,

$$p(cholesterol|\mathbf{do}(exercise)) = \sum_{age} p(cholesterol|exercise, age)p(age).$$

Causal Inference

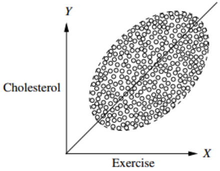


Figure 1.2: Results of the exercise-cholesterol study, unsegregated. The data points are identical to those of Figure 1.1, except the boundaries between the various age groups are not shown

- Due to the **confounder** Age,

$$p(cholesterol|\boldsymbol{do}(exercise)) \neq p(cholesterol|exercise).$$

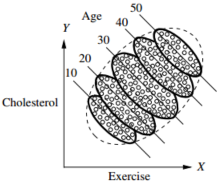
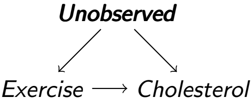


Figure 1.1: Results of the exercise-cholesterol study, segregated by age



Unobserved

Exercise ⟶ *Cholesterol*

- Now,

$$p(cholesterol|\boldsymbol{do}(exercise)) \neq f(p(cholesterol, exercise)).$$

Causal Models

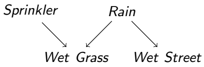
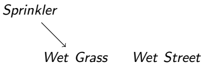
- A causal structure over a set of variables V is a DAG over V.
- A causal model consists of a causal structure, a set of functions $x_i = f_i(pa_i, u_i)$ for each $X_i \in V$, and a distribution $p(u_i)$ for each U_i.
- The functions are also called structural equations, which is **different** from algebraic equations since the equality sign should be read as an assignment or determination, i.e. it is asymmetric.
- The error, noise or disturbance terms U_i are assumed to be **independent** one of another. They may be seen as representing unmodeled or unobserved causes.
- Note that $f_i(pa_i, u_i)$ and $p(u_i)$ together define a conditional distribution $p(x_i|pa_i)$. Then, a causal model defines a distribution over V:

$$p(v) = \prod_i p(x_i|pa_i).$$

- Note that a causal model is also known as a Bayesian network.
- A causal model can be obtained from knowledge of the physics behind the phenomenon being modeled, from interventional experiments such as randomized control trials, or from passive observations. In the latter case, recall that the true model may not be uniquely identified due to structure equivalence.

Interventions

- Intervening on a variable $X_i \in V$ aims to modify the **natural** causal mechanism of X_i. For simplicity, we only consider interventions that set X_i to a fixed value x_i', and denote it as $do(x_i')$.
- Assume that the causal model at hand consists of a DAG G over V and a set of structural equations $x_i = f_i(pa_i, u_i)$ for all $X_i \in V$ together with a set of distributions $p(u_i)$ or, alternatively, a set of conditional distributions $p(x_i|pa_i)$ for all $X_i \in V$.
- The result of an intervention can be represented by modifying the given causal model:
    - **Delete** the equation corresponding to X_i.
    - Replace x_i with x_i' in the remaining equations.
    - **Delete** from G the directed edges into X_i.

Original	After $do(r_1)$
	
$p(s) = (0.3, 0.7)$ $p(r) = (0.5, 0.5)$ $p(wg\|r_0, s_0) = (0.1, 0.9)$ $p(wg\|r_0, s_1) = (0.7, 0.3)$ $p(wg\|r_1, s_0) = (0.8, 0.2)$ $p(wg\|r_1, s_1) = (0.9, 0.1)$ $p(ws\|r_0) = (0.1, 0.9)$ $p(ws\|r_1) = (0.7, 0.3)$	$p(s) = (0.3, 0.7)$ $p(wg\|s_0) = p(wg\|r_1, s_0) = (0.8, 0.2)$ $p(wg\|s_1) = p(wg\|r_1, s_1) = (0.9, 0.1)$ $p(ws) = p(ws\|r_1) = (0.7, 0.3)$ $p(s, wg, ws) = p(s)p(wg\|s)p(ws)$
$p(s, r, wg, ws) = p(s)p(r)p(wg\|s, r)p(ws\|r)$	

Truncated Factorization

- Either representation of an intervention results in a truncated factorization

$$p(v|do(x_i')) = \begin{cases} \prod_{j \neq i} p(x_j|pa_j) & \text{if } x_i = x_i' \\ 0 & \text{otherwise.} \end{cases}$$

- Note that

$$\prod_{j \neq i} p(x_j|pa_j) = p(v)/p(x_i'|pa_i) = p(v)p(pa_i)/p(x_i', pa_i)$$
$$= p(v \setminus \{x_i'\} \setminus pa_i | x_i', pa_i) p(pa_i).$$

- **Adjustment for direct causes:** Let $X_i, Y \in V$ st $Y \notin Pa_i$. Then,

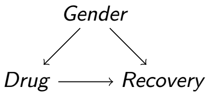
$$p(y|do(x_i')) = \sum_{pa_i} p(y|x_i', pa_i) p(pa_i).$$

- The goal of the above is to eliminate spurious (i.e., non-causal) dependencies between cause and effect.
- Note that if Y is not a descendant of X_i, then $Y \perp_G X_i | Pa_i$ and thus, as expected,

$$p(y|do(x_i')) = \sum_{pa_i} p(y|x_i', pa_i) p(pa_i) = \sum_{pa_i} p(y|pa_i) p(pa_i) = p(y).$$

- Things get more complicated when some variables in Pa_i are unobserved, since it prevents estimation of $p(y|x_i', pa_i)$ and $p(pa_i)$.

Truncated Factorization



	Drug=1	Drug=0
Gender=1	81/87 recovered	234/270 recovered
Gender=0	192/263 recovered	55/80 recovered

- Average causal effect:

$$E[R|do(D=1)]-E[R|do(D=0)] = p(R=1|do(D=1))-p(R=1|do(D=0))$$

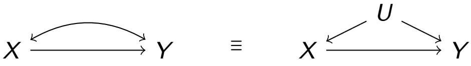
which can also be interpreted as the fraction of the population that recovers if everyone takes the drug compared to when no one takes the drug. Moreover, adjusting for the direct causes gives

$$\begin{aligned} p(R=1|do(D=1)) &= p(R=1|D=1,G=1)p(G=1)+p(R=1|D=1,G=0)p(G=0) \\ &= (81/87)(87+270)/700+(192/263)(263+80)/700 = 0.832 \end{aligned}$$

$$\begin{aligned} p(R=1|do(D=0)) &= p(R=1|D=0,G=1)p(G=1)+p(R=1|D=0,G=0)p(G=0) \\ &= (234/270)(87+270)/700+(55/80)(263+80)/700 = 0.7818 \end{aligned}$$

Causal Effect Identifiability

- Given a causal structure which may include unobserved variables, the causal effect $p(y|do(x_i'))$ is **identifiable** if it can be computed uniquely from any positive probability distribution over the observed variables.
- Positivity ensures that the effect is well defined.
- Therefore, $p(y|do(x_i'))$ is identifiable if Y, X_i, and Pa_i are observed, i.e. measured. The effect is computed by adjusting for the parents.
- $p(y|do(x_i'))$ is not identifiable in the **bow graph**:



- Proof: We construct two causal models M_1 and M_2 st $p_1(x,y) = p_2(x,y)$ but $p_1(y|do(x')) \neq p_2(y|do(x'))$. Specifically, let X, Y and U be binary, and take

M_1	M_2
$u = Uniform(0,1)$	$u = Uniform(0,1)$
$x = u$	$x = u$
$y = XOR(x,u)$	$y = 0.$

Back-Door Criterion

- A set of variables Z satisfies the **back-door criterion** with respect to an ordered pair of sets of variables (X, Y) in a causal structure G which may include unobserved variables if
    - Z contains no descendants of X, and
    - Z blocks every path between X and Y that contains an arrow into X.

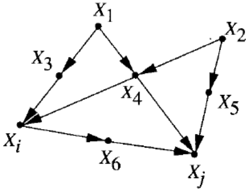


Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$.

- If Z satisfies the back-door criterion with respect to (X, Y), then

$$p(y|do(x)) = \sum_z p(y|x, z)p(z).$$

- The role of Z is to block only the paths entering X through the back-door. Several such sets Z may exist but some may be preferred, e.g. due to their size. Note that $Z = Pa_X$ always satisfies the criterion (what we called adjustment for the direct causes) but it may include latent variables.

Front-Door Criterion

- A set of variables Z satisfies the **front-door criterion** with respect to an ordered pair of sets of variables (X, Y) in a causal structure G which may include unobserved variables if
  - Z blocks all the directed paths from X to Y,
  - there is no unblocked back-door path from X to Z, and
  - all the back-door paths from Z to Y are blocked by X.

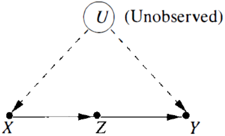


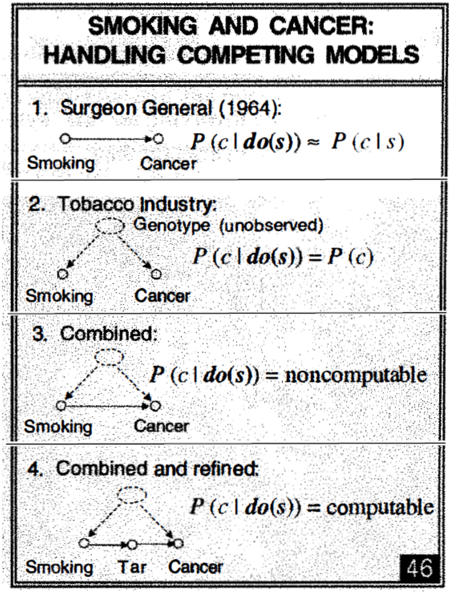
Figure 3.5 A diagram representing the front-door criterion. A two-step adjustment for Z yields a consistent estimate of $P(y \mid \hat{x})$.

- Note that Figure 3.5 is Figure 3.4 with $U = \{X_1, \ldots, X_5\}$. Note that the back-door criterion does not help here.
- If Z satisfies the front-door criterion with respect to (X, Y), then

$$p(y|do(x)) = \sum_z p(z|x) \sum_{x'} p(y|x', z)p(x').$$

Front-Door Criterion

- The effect of smoking on lung cancer: Non-identifiable vs identifiable.



SMOKING AND CANCER: HANDLING COMPETING MODELS

1. Surgeon General (1964):

$P(c \mid do(s)) \approx P(c \mid s)$

2. Tobacco Industry:

$P(c \mid do(s)) = P(c)$

3. Combined:

$P(c \mid do(s)) =$ noncomputable

4. Combined and refined:

$P(c \mid do(s)) =$ computable

46

do-Calculus

- Three rules whose repeated application together with standard probability manipulations, aims to transform a causal effect into an expression that only involves observational quantities:
  - Rule 1 (**insertion/deletion of observations**)

    $$p(y|do(x), \boldsymbol{z}, w) = p(y|do(x), w) \text{ if } Y \perp_{G_{\overline{X}}} Z|X \cup W$$

    $G_{\overline{X}}$ is G after deleting all the (bi)directed edges into X, i.e. simulate $do(x)$.
  - Rule 2 (**intervention/observation exchange**)

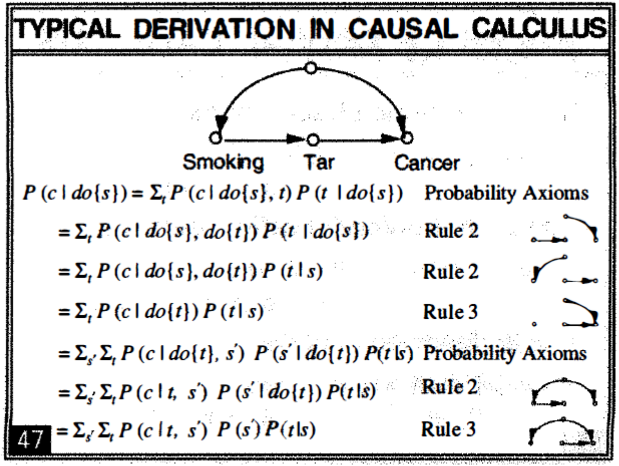
    $$p(y|do(x), \boldsymbol{do(z)}, w) = p(y|do(x), \boldsymbol{z}, w) \text{ if } Y \perp_{G_{\overline{X}\underline{Z}}} Z|X \cup W$$

    $G_{\overline{X}\underline{Z}}$ is G after deleting all the directed edges into X and all the directed edges out of Z.
  - Rule 3 (**insertion/deletion of interventions**)

    $$p(y|do(x), \boldsymbol{do(z)}, w) = p(y|do(x), w) \text{ if } Y \perp_{G_{\overline{X}\,\overline{Z(W)}}} Z|X \cup W$$

    where $Z(W)$ are the nodes in Z that are not ancestors of W in $G_{\overline{X}}$.
- The rules are sound and **complete**, i.e. all identifiable effects will be identified.
- There is a sound and complete algorithm to apply the rules.

do-Calculus



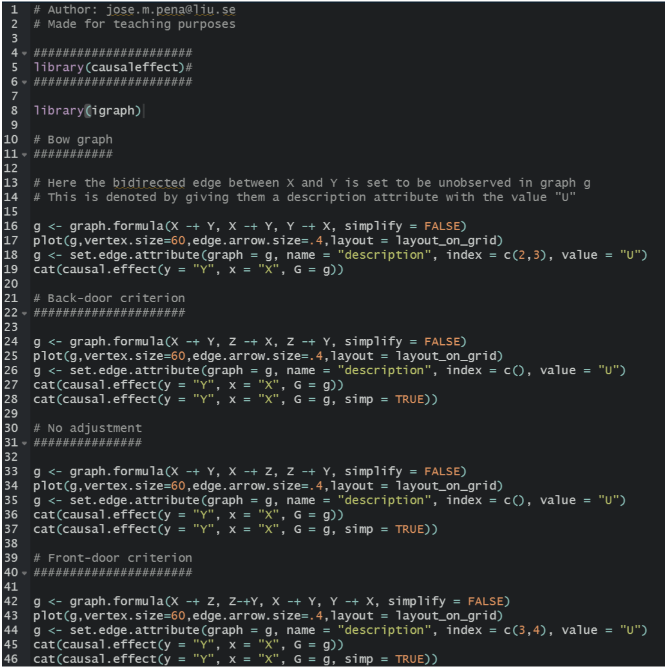
Smoking
Tar
Cancer

TYPICAL DERIVATION IN CAUSAL CALCULUS

$P(c \mid do\{s\}) = \Sigma_t P(c \mid do\{s\}, t)\, P(t \mid do\{s\})$	Probability Axioms
$= \Sigma_t P(c \mid do\{s\}, do\{t\})\, P(t \mid do\{s\})$	Rule 2
$= \Sigma_t P(c \mid do\{s\}, do\{t\})\, P(t \mid s)$	Rule 2
$= \Sigma_t P(c \mid do\{t\})\, P(t \mid s)$	Rule 3
$= \Sigma_{s'} \Sigma_t P(c \mid do\{t\}, s')\ P(s' \mid do\{t\})\, P(t \mid s)$	Probability Axioms
$= \Sigma_{s'} \Sigma_t P(c \mid t,\ s')\ P(s' \mid do\{t\})\, P(t \mid s)$	Rule 2
$= \Sigma_{s'} \Sigma_t P(c \mid t,\ s')\ P(s')\, P(t \mid s)$	Rule 3

47

do-Calculus



```
# Author: jose.m.pena@liu.se
# Made for teaching purposes

######################
library(causaleffect)#
######################

library(igraph)

# Bow graph
###########

# Here the bidirected edge between X and Y is set to be unobserved in graph g
# This is denoted by giving them a description attribute with the value "U"

g <- graph.formula(X -+ Y, X -+ Y, Y -+ X, simplify = FALSE)
plot(g,vertex.size=60,edge.arrow.size=.4,layout = layout_on_grid)
g <- set.edge.attribute(graph = g, name = "description", index = c(2,3), value = "U")
cat(causal.effect(y = "Y", x = "X", G = g))

# Back-door criterion
#####################

g <- graph.formula(X -+ Y, Z -+ X, Z -+ Y, simplify = FALSE)
plot(g,vertex.size=60,edge.arrow.size=.4,layout = layout_on_grid)
g <- set.edge.attribute(graph = g, name = "description", index = c(), value = "U")
cat(causal.effect(y = "Y", x = "X", G = g))
cat(causal.effect(y = "Y", x = "X", G = g, simp = TRUE))

# No adjustment
###############

g <- graph.formula(X -+ Y, X -+ Z, Z -+ Y, simplify = FALSE)
plot(g,vertex.size=60,edge.arrow.size=.4,layout = layout_on_grid)
g <- set.edge.attribute(graph = g, name = "description", index = c(), value = "U")
cat(causal.effect(y = "Y", x = "X", G = g))
cat(causal.effect(y = "Y", x = "X", G = g, simp = TRUE))

# Front-door criterion
######################

g <- graph.formula(X -+ Z, Z-+Y, X -+ Y, Y -+ X, simplify = FALSE)
plot(g,vertex.size=60,edge.arrow.size=.4,layout = layout_on_grid)
g <- set.edge.attribute(graph = g, name = "description", index = c(3,4), value = "U")
cat(causal.effect(y = "Y", x = "X", G = g))
cat(causal.effect(y = "Y", x = "X", G = g, simp = TRUE))
```

Summary

- Causal Inference
- Causal Models
- Interventions
- Truncated Factorization
- Causal Effect Identifiability
- Back-Door Criterion
- Front-Door Criterion
- *do*-Calculus
- Recommended readings on the importance of causality for ML and AI:
  - Darwiche, A. Human-Level Intelligence or Animal-Like Abilities ? *Communications of the ACM*, 61:56-67, 2018.
  - Pearl, J. The Seven Tools of Causal Inference with Reflections on Machine Learning. *Communications of the ACM*, 62:54-60, 2019.

Thank you