Intro to Causality for Computer Scientists

Instructor and Copyright: Bruno Ribeiro

Causality vs Data-driven Modeling

Introduction to Structural Causal Modeling

Understanding What Probabilities Really Are

In this course, we have often used libraries to sample random variables

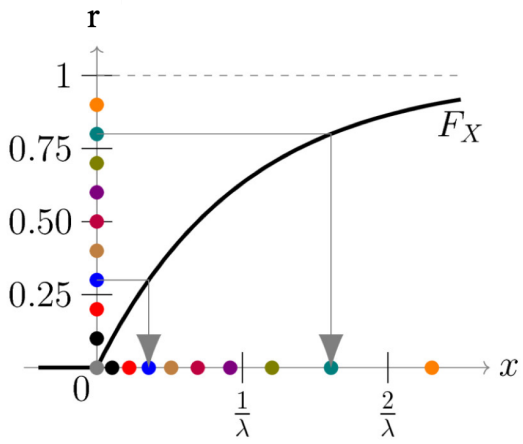
```
import numpy as np

[...]

ramdom_exp_values = np.random.exponential(my_lambda)

ramdom_normal_values = np.random.normal(average,std)
```

- **But how do these libraries work?**

Data Generation Process (i)

Inversion transform sampling method

Wikipedia example for exponential distribution $P(X \leq x) = F_X(x) = 1 - e^{-\lambda x}$ with inverse $x = F_X^{-1}(r) = -\frac{1}{\lambda}\ln(1-r)$:



- This is the most fundamental technique for generating sample values of random variables
- It uses the cumulative distribution function (CDF) of the random variable
- The method depends on the fact that, for any random variable X, the CDF, $F_X(x) = P(X \leq x)$, is a non-decreasing function of x that outputs a number in the interval [0,1]
- Let F_X^{-1} be the inverse of F_X, i.e., $x = F_X^{-1}(F_X(x))$.
- Let $r \sim \text{Uniform}(0,1)$ be a random uniform value in the interval [0,1]

    - This is obtained by a pseudorandom number generator

- Then,

$$x_{\text{sample}} = F_X^{-1}(r)$$

  is a random sample with distribution $P(X = x)$.

Data Generation Process (ii)

Any probability distribution

$$P(Y)$$

can be described as the data generated by the inverse transform sampling

$$Y = F_Y^{-1}(U),$$

where

$$F_Y^{-1} \text{ is a deterministic function}$$

and

$$U \sim \text{Uniform}(0, 1)$$

is some independent uniform random noise.

Notation: $a \sim b$ means a is "sampled from distribution" b

Data Generation Process (iii)

Conditional Distributions

Any **conditional** probability distribution

$$P(Y|X = x)$$

can be described as

$$Y = F_{Y|X}^{-1}(x, U)$$

for

$$U \sim \text{Uniform}(0, 1).$$

Data Generation Process & Simpson's Paradox

Consider the following supervised learning task. Doctors prescribe two different treatments (A and B) to patients with kidney stones. Our goal is to predict which treatment we should ascribe to a patient (even a Naïve Bayes classifier can do this simple task). Let $T \in \{A, B\}$ denote the prescribed treatment. And let $Y \in \{0, 1\}$ be the success (1) or failure (0) of the treatment. In our dataset, we have 700 patients ascribed treatment, equally balanced between A and B.

	Treatment	
	A	B
Y=1	(78%) 273	(83%) 289
Y=0	77	61

- Which treatment is more effective: A or B?

Alice, the person in charge of applying machine learning at the hospital, investigated the data a little further and identified that doctors find treatment A more invasive and tend to only prescribe it in more severe cases. Her new data shows the following

	Size of kidney stone			
	Small		Large	
	Treat. A	Treat. B	Treat. A	Treat. B
Y=1	(93%) 81	(87%) 234	(73%) 192	(69%) 55
Y=0	6	36	71	25

- Which treatment is more effective: A or B?

- Now treatment A seems to be more effective

Describing Joint Probability Distributions

Let $Y, T, S \in \{0, 1\}$ be three binary random variables.

Consider the following interpretation.

- Y = treatment positive outcome {0,1}
- T = treatment {A,B}
- S = kidney stone size.

Suppose we use hospital information as training data for our statistical model: $\mathcal{D} = \{(y_i, t_i, s_i)\}_{i=1}^n$ for each patient i.

The *chain rule* of probability) states

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A).$$

Hence, the joint probability distribution of (Y, T, S) is

$$P(Y, T, S)$$

and can be decomposed as

$$P(Y|T, S)P(T|S)P(S)$$

or

$$P(S|Y, T)P(Y|T)P(T)$$

or

$$P(S|T, Y)P(T|Y)P(Y)$$

or ...

On the Data Generation Process

A joint probability distribution is simply a way to assign probabilities to joint events

Q: Should we use conditional distributions to interpret how the data was generated?

A: Never, because for any of the following data generation processes **describes the training data equally well:**

1. A data generation process based on the decomposition $P(Y|T,S)P(T|S)P(S)$:

$$S = F_S^{-1}(U_S), \quad (1)$$
$$T = F_T^{-1}(S, U_T), \quad (2)$$
$$Y = F_Y^{-1}(T, S, U_Y), \quad (3)$$

where $U_S, U_T, U_Y \sim \text{Uniform}(0,1)$ are uniform variables sampled indepedently.

1. A data generation process based on the decomposition $P(Y|T,S)P(S|T)P(T)$:

$$T = F_T^{-1}(U_T), \quad (4)$$
$$S = F_S^{-1}(T, U_S), \quad (5)$$
$$Y = F_Y^{-1}(T, S, U_Y), \quad (6)$$

where $U_S, U_T, U_Y \sim \text{Uniform}(0,1)$ sampled indepedently.

1. A data generation process based on the decomposition $P(Y|T,S)P(S)P(T)$ that assumes $P(S|T) = P(S)$:

$$T = F_T^{-1}(U_T), \quad (7)$$
$$S = F_S^{-1}(U_S), \quad (8)$$
$$Y = F_Y^{-1}(T, S, U_Y), \quad (9)$$

where $U_S, U_T, U_Y \sim \text{Uniform}(0,1)$ sampled indepedently.

Hence, we cannot predict what happens if we force $T = A$ in the data generation process (force treatment to be "A"):

$$P(Y, S|do(T = A))$$

- `do()` notation: The DO notation asks what would happen if we forced a variable to have a certain value. This is the notation developed by Judea Pearl (Turin Award Winner 2011).

Alternative notation: An alternative notation for $P(Y, S|do(T = A))$ is

$$P(Y(T = A), S(T = A))$$

which is the notation used by Guido Imbens (Nobel Prize Winner 2021).

1. In the data generation process

$$S = F_S^{-1}(U_S), \quad (10)$$
$$T = F_T^{-1}(S, U_T), \quad (11)$$
$$Y = F_Y^{-1}(T, S, U_Y), \quad (12)$$

the "do" operation is forcing $T = A$, hence the data is generated as

$$S = F_S^{-1}(U_S), \quad (13)$$
$$T = A, \quad (14)$$
$$Y = F_Y^{-1}(T, S, U_Y). \quad (15)$$

1. In a different data generation process, the "do(T=A)" operation gets the following data

$$T = A, \quad (16)$$
$$S = F_S^{-1}(T, U_S), \quad (17)$$
$$Y = F_Y^{-1}(T, S, U_Y). \quad (18)$$

2. In yet another data generation process, the "do(T=A)" operation gets the following data

$$T = A, \tag{19}$$
$$S = F_S^{-1}(U_S), \tag{20}$$
$$Y = F_Y^{-1}(T, S, U_Y). \tag{21}$$

Q: Which data generation process is more likely to describe our hospital data?

	Size of kidney stone			
	Small		Large	
	Treat. A	Treat. B	Treat. A	Treat. B
Y=1	(93%) 81	(87%) 234	(73%) 192	(69%) 55
Y=0	6	36	71	25

The Dangers of Data-driven Machine Learning

	Treatment	
	A	B
Y=1	(78%) 273	(83%) 289
Y=0	77	61

In our data, we found that given treatment B ($T = B$), patients are more likely to recover ($Y = 1$) than with treatment $T = A$:

$$P(Y = 1|T = B) > P(Y = 1|T = A)$$

Is the above enough evidence to say that treatment B is better than A?

The "Simple Statistical Model" fallacy

- In another hospital, it is possible that $P(Y = 1|T = B) \approx 1$ and $P(Y = 1|T = A) \approx 0$, which would allow us to build a simple predictive model
- Still, even under this scenario, we could still have $P(Y = 1|do(T = B)) \approx 0$.
  - Using model simplicity to justify our classifier's decisions is an example of *associational machine learning*
    - Occam's raisor: the simplest explanation is likely the true explanation
      - Occam's raisor is a misleading principle for explaining cause and effect

Causal Execution Directed Acyclic Graph

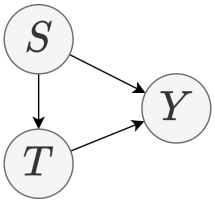
We could describe the data generation process of this problem using the following random variables:

- S = Kidney stone size
- T = Treatment type
- Y = Treatment outcome

$$S = F_S^{-1}(U_{\text{stone size}}), \tag{22}$$
$$T = F_T^{-1}(S, U_{\text{treatment}}), \tag{23}$$
$$Y = F_C^{-1}(T, S, U_{\text{outcome}}), \tag{24}$$

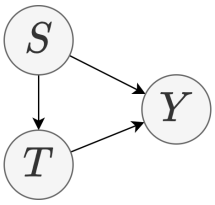
where $U_{\text{stone size}}, U_{\text{treatment}}, U_{\text{outcome}} \in [0, 1]$ are independent variables.

The above data generation can be described by an execution graph, called the **causal Directed Acyclic Graph (DAG)**:



Confounder variables

- We say `kidney stone size` (S) is a **confounder variable**, which is a common cause for both Treatment T and outcome Y



Another data generation process for $P(Y, T, S)$:

$$S = F_S^{-1}(U_{\text{Zeus}}), \tag{25}$$
$$T = F_T^{-1}(U_{\text{Zeus}}), \tag{26}$$
$$Y = F_Y^{-1}(U_{\text{Zeus}}), \tag{27}$$

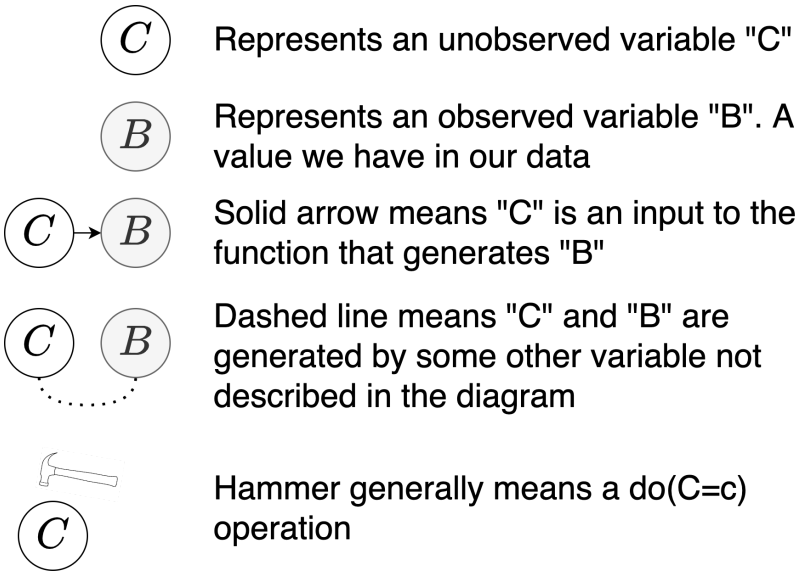
where $U_{\text{Zeus}} \sim \text{Uniform}(0, 1)$ a decision of the Greek god Zeus.

Q: From data alone, can we tell which data generation process is the correct one?

No. From data alone, we cannot tell which data generation process is the correct one

Causal Directed Acyclic Graph (Causal DAG)

The DAG graph notation is as follows:



Structural Causal Models (SCMs)

Structural Causal Modeling (SCM) is a formal way to describe what we know about the data generation process.

- Structural Causal Modeling is a combination of data generation equations and their graphical representation

- Think of SCM as a description of the *code* that generated the data

(Galles & Pearl (1998)) shows that any data generation process can be described through a causal DAG.

- The variables $X_1, \ldots, X_n$ are the endogenous variables
  - Endogenous variables are real quantities that one could measure
- The variables $U_1, \ldots, U_m$ are called exhogenous variables, $m \geq n$.
  - Exhogenous variables are not explicitly modeled in our task (often they cannot be measured)
- Directed Acyclic Graph (DAG) G with endogenous variables as vertices $X_1, \ldots, X_n$
  - May also include exhogenous variables $U_1, \ldots, U_m$ as vertices
- Semantics: Parents = direct causes
  - $\mathrm{PA}(H)$ are the parents of variable H in the causal DAG (described next).
- We define a vertex X_i as

$$X_i := f_i(\mathrm{PA}(X_i), U_i), \quad i \in \{1, \ldots, n\}$$

where U_i are denoted as noise variable (or just exhogenous variables)

Independence between Cause and Mechanism (ICM) (Lemeire & Dirkx 2006), (Janzing & Scholkopf 2010)

- Independence between Cause and Mechanism (ICM) generally assumes that:

1. The mechanisms $\{f_i\}_{i=1}^n$ do not depend on the exhogeneous variables $U_1, \ldots, U_m$
2. The exhogeneous variables $U_1, \ldots, U_m$ are independent.

Causal Effects

What happens with T if we force $do(T = B)$, i.e., we "force" treatment B on patients (regardless of their kidney stone condition).

- This "forcing" is called:
  - An **intervention** if it is done before our data is collected (e.g., to a new person).
    - Example: Clinical trials. Volunteers in the trial are forced to either take the drug or take the placebo.
  - **Counterfactual reasoning** if it is done after the data is collected. That is, we consider an alternative reality that goes against some fact in our data.

Consider the SCM:

$$S = F_S^{-1}(U_{\text{Zeus}}), \tag{28}$$
$$T = F_T^{-1}(U_{\text{Zeus}}), \tag{29}$$
$$Y = F_Y^{-1}(U_{\text{Zeus}}), \tag{30}$$

where $U_{\text{Zeus}} \sim \mathrm{Uniform}(0, 1)$ a decision of the Greek god Zeus.

Now let's see what happens to Y if we set $do(T = B)$:

$$S = F_S^{-1}(U_{\text{Zeus}}), \tag{31}$$
$$T = B, \tag{32}$$
$$Y = F_Y^{-1}(U_{\text{Zeus}}). \tag{33}$$

Q: Under this data generation process, does forcing treatment B changes the probability of a favorable outcome?

A: No.

Now consider another data generation process (SCM) that could generate the same data:

$$S = F_S^{-1}(U_{\text{Zeus}}), \tag{34}$$
$$T = F_T^{-1}(U_{\text{Zeus}}), \tag{35}$$
$$Y = F_Y^{-1}(T, U_{\text{Zeus}}). \tag{36}$$

Q: Could forcing treatment B (that is, $do(T = B)$) change the probability of patient outcome?

$$S = F_S^{-1}(U_{\text{Zeus}}), \tag{37}$$
$$T = B, \tag{38}$$
$$Y = F_Y^{-1}(T, U_{\text{Zeus}}). \tag{39}$$

A: Yes.

Structural Causal Models

(Galles & Pearl (1998)) shows that any data generation process can be described through a causal DAG.

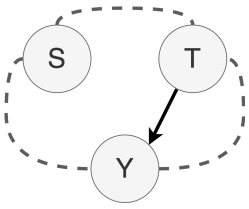
- In our previous equations, the variables $U_{\text{Zeus}}, U_Y, U_T, U_S$ are called exhogenous variables
  - **Exhogenous variables** are not explicitly modeled in our task (often they cannot be measured)
- The variables S, T, Y are the endogenous variables
  - **Endogenous variables** are real quantities that one could measure
- $\text{PA}(Y)$ are the parents of variable Y in the causal DAG (described next).

Causal Directed Acyclic Graph (Causal DAG)

A simple way to describe the above data generation process is through its "execution" graph (Causal DAG):

Example of the Causal DAG from

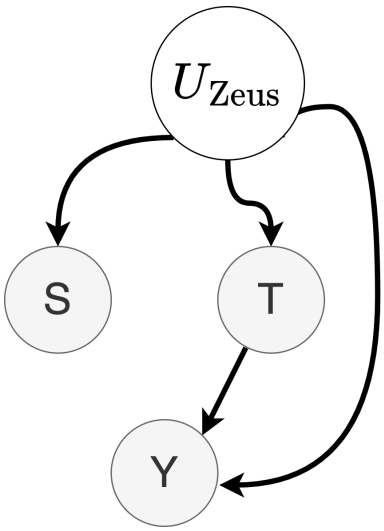
$$S = F_S^{-1}(U_{\text{Zeus}}), \tag{40}$$
$$T = F_T^{-1}(U_{\text{Zeus}}), \tag{41}$$
$$Y = F_Y^{-1}(T, U_{\text{Zeus}}). \tag{42}$$



- The solid arrows indicate a variable dependence in the SCM.
  - The solid arrows must form a Directed Acyclic Graph (DAG) over the described variables.
- The dashed arrows show how the variables are related through undescribed variables

We could also include all variables in the causal DAG:

- DAG nodes have two colors:
  - Gray means observed variables
  - White means unobserved variables



Expanding the Causal Model DAG

(Galles & Pearl (1998)) shows that any data generation process can always be represented by a DAG:

- We can ALWAYS add exoghenous variables to make the data generation process directed.

```
U = np.random.uniform(0,1)
Chiken = F_Chiken(Egg, U)
Egg = F_Egg(Chiken, U)
```

- E.g., dinosaurs already layed eggs way before chickens appeared on Earth. Can be described as

```
U_Dino = np.random.uniform(0,1)
Dinosaur = F_Dino(U_Dino)
U = np.random.uniform(0,1)
Egg = F_Egg(Dinosaur, U)
Chiken = F_Chiken(Egg, U)
```

Predicting Causal Effects

- **Goal:** We want to predict $P(Y|\text{do}(X=x))$
  - That is, we want to predict what happens to the probability distribution of Y if we force $X=x$.
- **Causal Adjustment Formula (Adjustment for Direct Causes, Theorem 3.2.2 of (Pearl 2009))**:
  - Suppose Y is any set of random variables disjoint with $(X \cup C)$, where $C = \text{PA}(X)$ includes all direct parents of X on the causal DAG.
  - Then,

$$P(Y=y|\text{do}(X=x)) = \sum_c P(Y=y|C=c, X=x)P(C=c), \qquad \forall y \in \mathbb{Y}$$

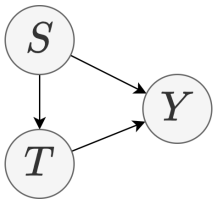
- Note the difference bettween the adjustment formula above and a *standard conditional probability* statement:

$$P(Y=y|X=x) = \sum_c P(Y=y|C=c, X=x)P(C=c|X=x).$$

	Size of kidney stone			
	Small		Large	
	Treat. A	Treat. B	Treat. A	Treat. B
Y=1	(93%) 81	(87%) 234	(73%) 192	(69%) 55
Y=0	6	36	71	25

Causal Adjustment Formula Example:

- T = Treatment
- Y = Treatment outcome
- S = Kidney stone size Let's assume the following causal DAG:



Let's compare $P(Y=1|\text{do}(T=t))$ against $P(Y=1|T=t)$, $t \in \{A, B\}$.

	Size of kidney stone			
	Small		Large	
	Treat. A	Treat. B	Treat. A	Treat. B
Y=1	(93%) 81	(87%) 234	(73%) 192	(69%) 55
Y=0	6	36	71	25

The difference between conditional and interventional distributions:

- $P(Y=1|T=A) = 0.78$
- $P(Y=1|T=B) = 0.83$

- $P(Y=1|\text{do}(T=A)) = \sum_{s\in\{\text{small,large}\}} P(Y=1|S=s,T=A)P(S=s) = 0.93 \times 0.51 + 0.73 \times 0.49 = 0.832$
- $P(Y=1|\text{do}(T=B)) = \sum_{s\in\{\text{small,large}\}} P(Y=1|S=s,T=B)P(S=s) = 0.87 \times 0.51 + 0.69 \times 0.49 = 0.781$

Example 2 (COVID in Israel, Aug 2021)

- Covid-19 hospitalizations in Israel (Aug 17. 2021)
- Data from Israeli government data dashboard
- Data collected by Jeffrey Morris

Age	Severe Cases		Score function
	Not Vax	Fully Vax	Fraction of vaccinated among severe cases
All ages	214	301	58%

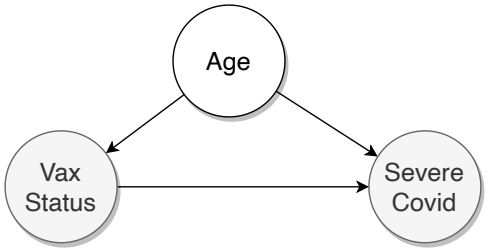
- Efficacy defined as

$$\text{Efficacy} = \frac{(\text{severe cases Fully Vax per 100k})}{(\text{All severe cases})}.$$

Age	Population %		Severe Cases		Score function
	Not Vax	Fully Vax	Not Vax per 100k	Fully Vax per 100k	Efficacy
All ages	1,302,912 18.2%	5,634,634 78.7%	16.5	5.3	67%

Covid Causal Graph

- COVID vulnerability determines disease severity
- More vulnerable individuals more likely to have the vaccine
- Older people are more vulnerable



- Data from Israeli government data dashboard
- Data collected by Jeffrey Morris
- Age-conditional efficacy defined as

$$\text{Efficacy} \mid \text{Age} = \frac{(\text{severe cases Fully Vax per 100k} \mid \text{Age})}{(\text{All severe cases} \mid \text{Age})}.$$

Age	Population %		Severe Cases		Score function
Conditional	Not Vax	Fully Vax	Not Vax per 100k	Fully Vax per 100k	Conditional Efficacy
[12,15]	62.1%	29.9%	0.3	0.0	100.0%
[16,19]	21.9%	73.5%	1.6	0.0	100.0%
[20,29]	20.5%	76.2%	1.5	0.0	100.0%
[30,39]	16.2%	80.9%	6.2	0.2	96.8%
[40,49]	13.2%	84.4%	16.5	1.0	93.9%
[50,59]	10.0%	88.0%	40.2	2.9	92.8%
[60,69]	8.8%	89.8%	76.6	8.7	88.7%
[70,79]	4.2%	94.6%	190.1	19.8	89.6%
[80,89]	5.6%	92.6%	252.3	47.9	81.1%
90+	6.1%	90.5%	510.9	38.6	92.4%

Zillow Home Purchase Case

See slides

References

- (Pearl 2009) Judea Pearl, Causality : models, reasoning, and inference, ISBN 0-521-77362-8, Cambridge University Press, 2009

In [ ]: