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

```python
import numpy as np

random_exp_values = np.random.exponential(my_lambda)
random_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 $x = F_X^{-1}(r) = -\frac{1}{\lambda} \ln(1 - r)$:

![Diagram of Inversion Transform Sampling Method](image)

- 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

\( F_Y^{-1} \) 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.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=1</td>
<td>273 (78%)</td>
<td>289 (83%)</td>
</tr>
<tr>
<td>Y=0</td>
<td>77</td>
<td>61</td>
</tr>
</tbody>
</table>
• 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

<table>
<thead>
<tr>
<th>Size of kidney stone</th>
<th>Small</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=1</td>
<td>(93%) 81</td>
<td>(87%) 234</td>
</tr>
<tr>
<td>Y=0</td>
<td>6</td>
<td>36</td>
</tr>
</tbody>
</table>

• 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: \( D = \{(y_i,t_i,s_i)\}_{i=1}^n \) 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 $P(Y|T, S)P(T|S)P(S)$:

\[
S = F_S^{-1}(U_S),  \\
T = F_T^{-1}(S, U_T),  \\
Y = F_Y^{-1}(T, S, U_Y),
\]

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

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

\[
T = F_T^{-1}(U_T),  \\
S = F_S^{-1}(T, U_S),  \\
Y = F_Y^{-1}(T, S, U_Y),
\]

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

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

\[
T = F_T^{-1}(U_T),  \\
S = F_S^{-1}(U_S),  \\
Y = F_Y^{-1}(T, S, U_Y),
\]

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

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

\[
P(Y, S|do(T = A))
\]

- \textbf{do()} notation: The \textit{do} notation asks what would happen if we forced a variable to have a certain value. This is the notation developed by Judea Pearl (Turing Award Winner 2011).

\textbf{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

\[
S = F_S^{-1}(U_S),  \\
T = F_T^{-1}(S, U_T),  \\
Y = F_Y^{-1}(T, S, U_Y),
\]

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

\[
S = F_S^{-1}(U_S),  \\
T = A,  \\
Y = F_Y^{-1}(T, S, U_Y).
\]

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

\[
T = A,  \\
S = F_S^{-1}(T, U_S),  \\
Y = F_Y^{-1}(T, S, U_Y).
\]

1. In yet another data generation process, the "do(T=A)" operation gets the following data
\[ T = A, \]
\[ S = F_S^{-1}(U_S), \]
\[ Y = F_Y^{-1}(T, S, U_Y). \]  

(19)  
(20)  
(21)

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

<table>
<thead>
<tr>
<th>Size of kidney stone</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y = 1</td>
<td>(93%) 81</td>
</tr>
<tr>
<td>Y = 0</td>
<td>6</td>
</tr>
</tbody>
</table>

The Dangers of Data-driven Machine Learning

<table>
<thead>
<tr>
<th>Treatment</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y = 1</td>
<td>(78%) 273</td>
</tr>
<tr>
<td>Y = 0</td>
<td>77</td>
</tr>
</tbody>
</table>

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

\[ S = F_S^{-1}(U_{stone\ size}), \]  
(22)  
\[ T = F_T^{-1}(S, U_{treatment}), \]  
(23)  
\[ Y = F_Y^{-1}(T, S, U_{outcome}), \]  
(24)

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

The above data generation can be described by an execution graph, called the *causal Directed Acyclic Graph (DAG)*:
Confounding 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 \(P(Y, T, S)\):

\[
S = F_S^{-1}(U_{Zeus}),
\]

\[
T = F_T^{-1}(U_{Zeus}),
\]

\[
Y = F_Y^{-1}(U_{Zeus}),
\]

where \(U_{Zeus} \sim \text{Uniform}(0, 1)\) 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:

- **Solid arrow** means that variable \(A\) is an input to the function that generates \(B\).
- **Dashed line** means that variables \(A\) and \(B\) are generated by some other variable not described in the diagram.
- **Hammer** generally means a do\((C=c)\) operation.

**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 exogenous variables, $m \geq n$.
- Exogenous 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 exogenous variables $U_1, \ldots, U_m$ as vertices.
- Semantics: Parents = direct causes.
- $PA(H)$ are the parents of variable $H$ in the causal DAG (described next).
- We define a vertex $X_i$ as

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

where $U_i$ are denoted as noise variable (or just exogenous 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 exogeneous variables $U_1, \ldots, U_m$.
  2. The exogeneous 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:

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

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

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

$$S = F_S^{-1}(U_{Zeus})$$ (31)
$$T = B$$ (32)
$$Y = F_Y^{-1}(U_{Zeus})$$ (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:
\[ S = F_S^{-1}(U_{Z\text{ras}}), \quad T = F_T^{-1}(U_{Z\text{ras}}), \quad Y = F_Y^{-1}(T, U_{Z\text{ras}}). \]  

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

\[ S = F_S^{-1}(U_{Z\text{ras}}), \quad T = B, \quad Y = F_Y^{-1}(T, U_{Z\text{ras}}). \]

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 \( U_{\text{Zras}}, U_Y, U_T, U_S \) are called exogenous variables
  - Exogenous 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

\[ S = F_S^{-1}(U_{Z\text{ras}}), \quad T = F_T^{-1}(U_{Z\text{ras}}), \quad Y = F_Y^{-1}(T, U_{Z\text{ras}}). \]  

- 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 exogenous variables to make the data generation process directed.

\[ U = \text{np.random.uniform}(0, 1) \]
\[ \text{Chiken} = \text{F}_\text{Chiken}(\text{Egg}, U) \]
\[ \text{Egg} = \text{F}_\text{Egg}(\text{Chiken}, U) \]

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

\[ \text{U_Dino} = \text{np.random.uniform}(0, 1) \]
\[ \text{Dinosaur} = \text{F}_\text{Dino}(\text{U_Dino}) \]
\[ U = \text{np.random.uniform}(0, 1) \]
\[ \text{Egg} = \text{F}_\text{Egg}(\text{Dinosaur}, U) \]
\[ \text{Chiken} = \text{F}_\text{Chiken}(\text{Egg}, U) \]

Predicting Causal Effects

- **Goal:** We want to predict \( P(Y \mid \text{do}(X = x)) \)
  - 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,

\[
P(Y = y \mid \text{do}(X = x)) = \sum_c P(Y = y \mid C = c, X = x) P(C = c), \quad \forall y \in \mathbb{Y}
\]

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

\[
P(Y = y \mid X = x) = \sum_c P(Y = y \mid C = c, X = x) P(C = c \mid X = x).
\]
<table>
<thead>
<tr>
<th>Size of kidney stone</th>
<th>Small</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=1</td>
<td>(93%) 81</td>
<td>(73%) 192</td>
</tr>
<tr>
<td>Y=0</td>
<td>6</td>
<td>36</td>
</tr>
</tbody>
</table>

Causal Adjustment Formula Example:

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

Let's assume the following causal DAG:

```
S -> T -> Y
```

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

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

<table>
<thead>
<tr>
<th>Age</th>
<th>Severe Cases</th>
<th>Score function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Vax</td>
<td>Fully Vax</td>
</tr>
<tr>
<td>All ages</td>
<td>214</td>
<td>301</td>
</tr>
</tbody>
</table>

- Efficacy defined as

\[
\text{Efficacy} = \frac{(\text{severe cases Fully Vax per 100k})}{(\text{All severe cases})}
\]
<table>
<thead>
<tr>
<th>Age</th>
<th>Population %</th>
<th>Severe Cases</th>
<th>Score function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Efficacy</td>
</tr>
<tr>
<td></td>
<td>Not Vax</td>
<td>Fully Vax</td>
<td>Not Vax per 100k</td>
</tr>
<tr>
<td>All ages</td>
<td>1,302,912</td>
<td>5,634,634</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td>18.2%</td>
<td>78.7%</td>
<td></td>
</tr>
</tbody>
</table>

Covid Causal Graph

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

![Causal Graph Diagram]

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

\[
Efficacy | Age = \frac{(\text{severe cases Fully Vax per 100k | Age})}{(\text{All severe cases | Age})}.
\]

<table>
<thead>
<tr>
<th>Age</th>
<th>Population %</th>
<th>Severe Cases</th>
<th>Score function</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Efficacy</td>
</tr>
<tr>
<td></td>
<td>Conditional</td>
<td>Not Vax</td>
<td>Fully Vax</td>
</tr>
<tr>
<td>[12,15]</td>
<td>62.1%</td>
<td>29.9%</td>
<td>0.3</td>
</tr>
<tr>
<td>[16,19]</td>
<td>21.9%</td>
<td>73.5%</td>
<td>1.6</td>
</tr>
<tr>
<td>[20,29]</td>
<td>20.5%</td>
<td>76.2%</td>
<td>1.5</td>
</tr>
<tr>
<td>[30,39]</td>
<td>16.2%</td>
<td>80.9%</td>
<td>6.2</td>
</tr>
<tr>
<td>[40,49]</td>
<td>13.2%</td>
<td>84.4%</td>
<td>16.5</td>
</tr>
<tr>
<td>[50,59]</td>
<td>10.0%</td>
<td>88.0%</td>
<td>40.2</td>
</tr>
<tr>
<td>[60,69]</td>
<td>8.8%</td>
<td>89.8%</td>
<td>76.6</td>
</tr>
<tr>
<td>[70,79]</td>
<td>4.2%</td>
<td>94.6%</td>
<td>190.1</td>
</tr>
<tr>
<td>[80,89]</td>
<td>5.6%</td>
<td>92.6%</td>
<td>252.3</td>
</tr>
<tr>
<td>90+</td>
<td>6.1%</td>
<td>90.5%</td>
<td>510.9</td>
</tr>
</tbody>
</table>

Zillow Home Purchase Case

See slides

References
