Towards Causal Representation Learning

Nikita Stepanov, Anton Medvedev, Victor Grishanin, Nikita Morozov

November 10, 2021

Statistical approaches

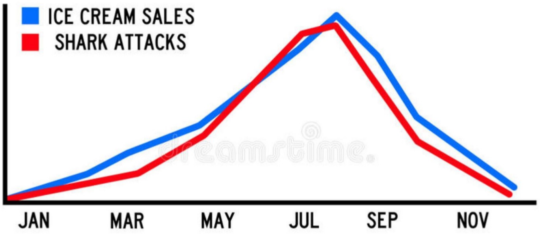
	Given	Learned
Regression	$\mathcal{D} \sim P(x, y)$	$E[Y\|X]$
Classification	$\mathcal{D} \sim P(x, y)$	$P(Y\|X)$
Generation	$\mathcal{D} \sim P(x)$	$P(X)$

Statistical approaches

	Given	Learned
Regression	$\mathcal{D} \sim P(x, y)$	$E[Y\|X]$
Classification	$\mathcal{D} \sim P(x, y)$	$P(Y\|X)$
Generation	$\mathcal{D} \sim P(x)$	$P(X)$

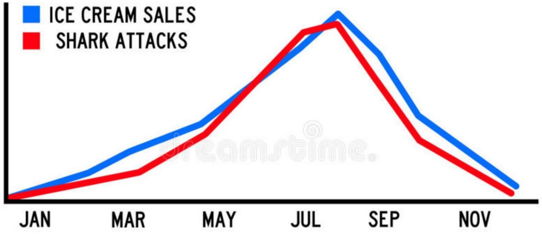
- In most cases in machine learning we use statistical approaches

Correlation does not imply causation



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

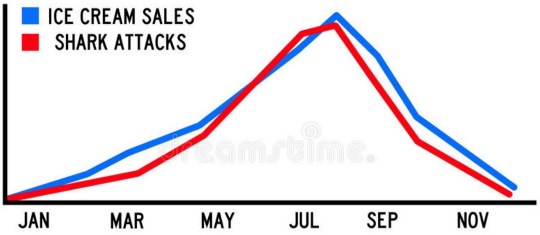
Correlation does not imply causation



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

- If we reduce ice cream sales, will the number of attacks decrease?

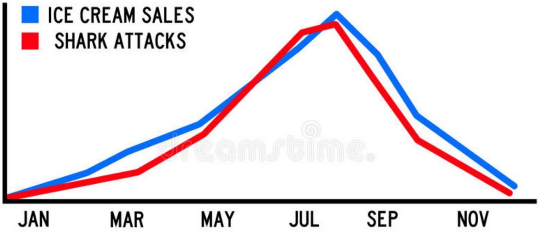
Correlation does not imply causation



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

- If we reduce ice cream sales, will the number of attacks decrease?
- $\mathcal{D} \sim \mathrm{P}(\mathrm{x}, \mathrm{y}) \to \mathrm{E}[\mathrm{Y}|\mathrm{X}]$

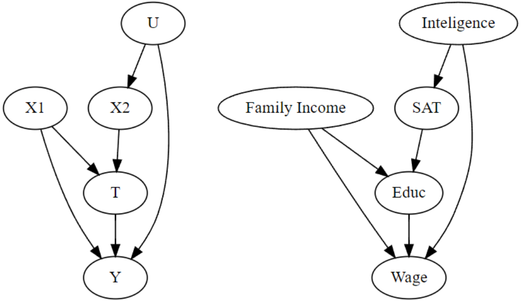
Correlation does not imply causation



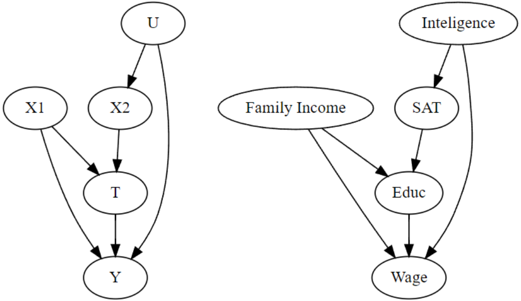
Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

- If we reduce ice cream sales, will the number of attacks decrease?
- $\mathcal{D} \sim \mathrm{P}(\mathrm{x}, \mathrm{y}) \to \mathrm{E}[\mathrm{Y}|\mathrm{X}]$
- But it's not the answer to our question because action $\neq$ observation!

Graphical causal models



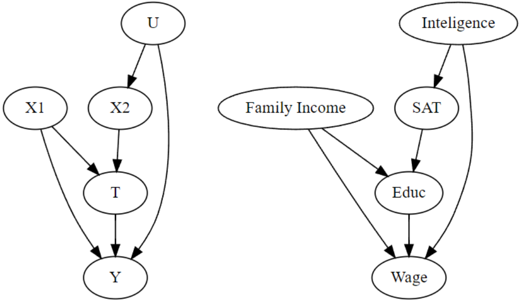
Graphical causal models



- $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_{i+1}, ..., X_n)$ — entangled representation
- $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | PA_i)$ — disentangled representation

- To define the model we need to specify $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | PA_i)$, where PA_i is a set of parents of X_i

Graphical causal models

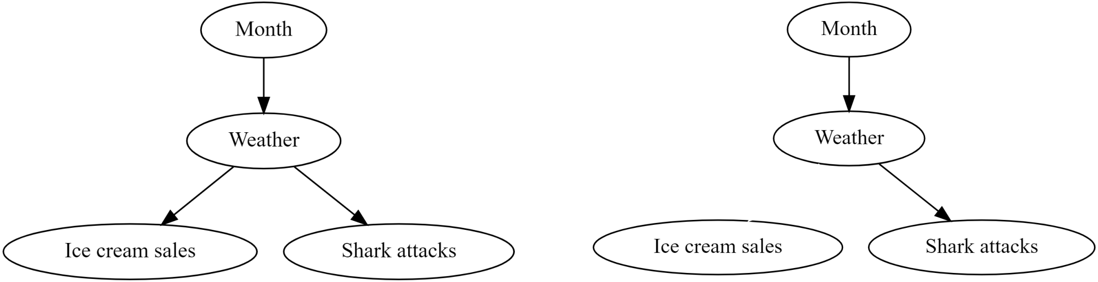


- $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | X_{i+1}, ..., X_n)$ — entangled representation
- $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | PA_i)$ — disentangled representation

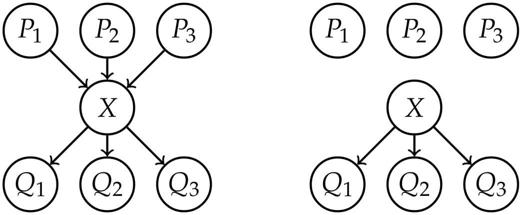
- To define the model we need to specify $P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i | PA_i)$, where PA_i is a set of parents of X_i
- In this example $P(U, X_1, X_2, T, Y) = P(U)P(X_1)P(X_2|U)P(T|X_1, X_2)P(Y|X_1, T, U)$

Interventions

- Intervention is a hypothetical action
- Intervention arbitrarily alters the model



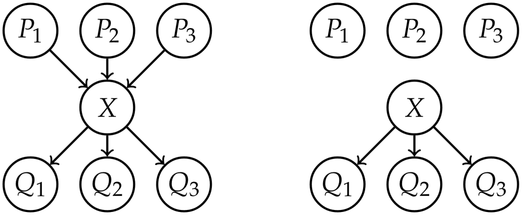
Substitution



Substitutions a.k.a do-interventions a.k.a. assignments are quite simple but also expressive

If M is the initial model, then $\mathrm{M}\left[\mathrm{do}(\mathrm{X} := \mathrm{x})\right]$ is the model after substitution

Substitution

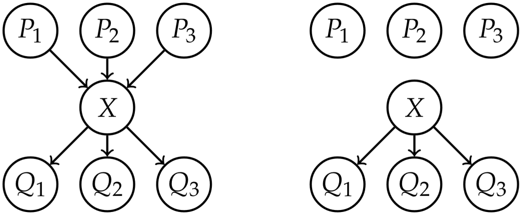


Substitutions a.k.a do-interventions a.k.a. assignments are quite simple but also expressive

If M is the initial model, then $\mathrm{M}\left[\mathrm{do}(\mathsf{X} := \mathsf{x})\right]$ is the model after substitution

$$\mathsf{P}_{\mathsf{M}[\mathrm{do}(\mathsf{X}:=\mathsf{x})]}\left[\mathsf{Y}=\mathsf{y}\right] = \sum_{\mathsf{z}} \mathsf{P}_{\mathsf{M}}\left[\mathsf{Y}=\mathsf{y}|\mathsf{X}=\mathsf{x},\mathsf{PA}=\mathsf{z}\right]\mathsf{P}_{\mathsf{M}}\left[\mathsf{PA}=\mathsf{z}\right]$$

Substitution



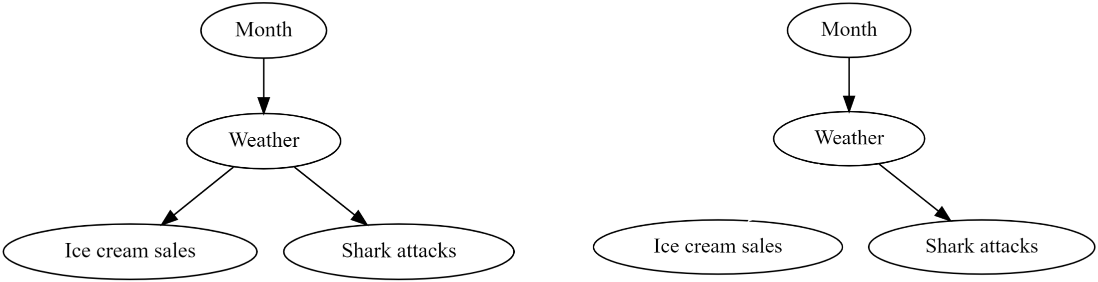
Substitutions a.k.a do-interventions a.k.a. assignments are quite simple but also expressive

If M is the initial model, then $\mathsf{M}\left[\mathrm{do}(\mathsf{X} := \mathsf{x})\right]$ is the model after substitution

$$\mathsf{P}_{\mathsf{M}[\mathrm{do}(\mathsf{X}:=\mathsf{x})]}\left[\mathsf{Y} = \mathsf{y}\right] = \sum_{\mathsf{z}} \mathsf{P}_{\mathsf{M}}\left[\mathsf{Y} = \mathsf{y} | \mathsf{X} = \mathsf{x}, \mathsf{PA} = \mathsf{z}\right] \mathsf{P}_{\mathsf{M}}\left[\mathsf{PA} = \mathsf{z}\right]$$

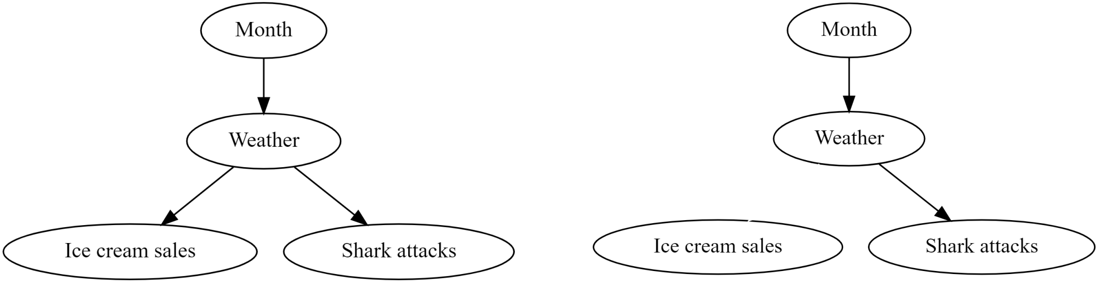
$$\mathsf{P}_{\mathsf{M}}\left[\mathsf{Y} = \mathsf{y} | \mathsf{X} = \mathsf{x}\right] = \sum_{\mathsf{z}} \mathsf{P}_{\mathsf{M}}\left[\mathsf{Y} = \mathsf{y}, \mathsf{PA} = \mathsf{z} | \mathsf{X} = \mathsf{x}\right] = \sum_{\mathsf{z}} \mathsf{P}_{\mathsf{M}}\left[\mathsf{Y} = \mathsf{y} | \mathsf{X} = \mathsf{x}, \mathsf{PA} = \mathsf{z}\right] \mathsf{P}_{\mathsf{M}}\left[\mathsf{PA} = \mathsf{z} | \mathsf{X} = \mathsf{x}\right]$$

Substitution



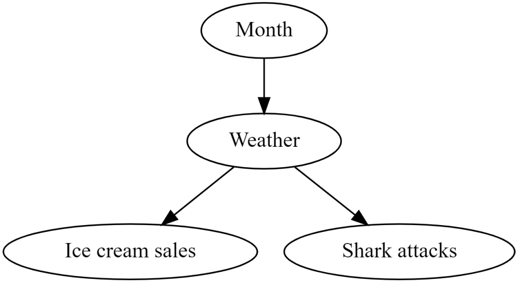
- If we reduce ice cream sales, will the number of attacks decrease?

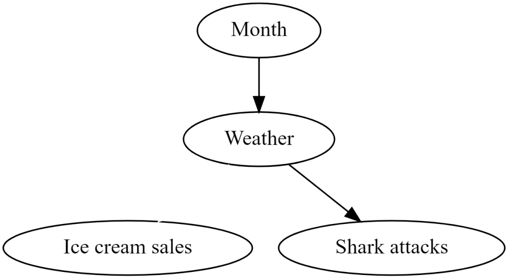
Substitution



- If we reduce ice cream sales, will the number of attacks decrease?
- Consider $M_1 = M\Big[\mathrm{do}(\text{Ice cream sales} := x_1)\Big]$ and $M_2 = M\Big[\mathrm{do}(\text{Ice cream sales} := x_2)\Big]$

Substitution

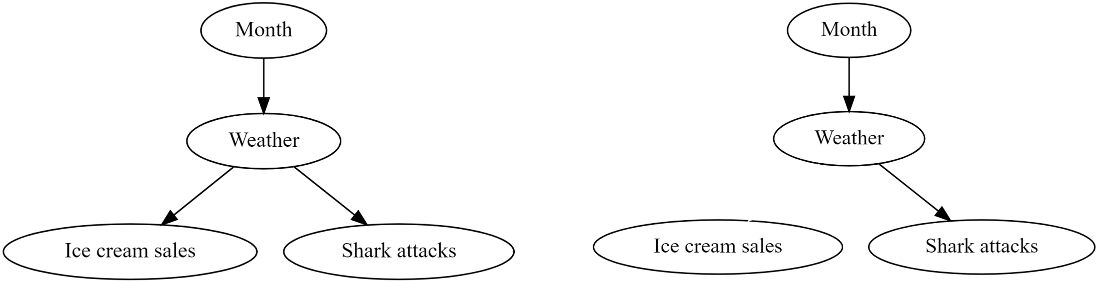




- If we reduce ice cream sales, will the number of attacks decrease?
- Consider $M_1 = M\left[\mathrm{do}(\text{Ice cream sales} := x_1)\right]$ and $M_2 = M\left[\mathrm{do}(\text{Ice cream sales} := x_2)\right]$

$$\mathbb{E}_{M_1}\left[\text{Shark attacks}\right] - \mathbb{E}_{M_2}\left[\text{Shark attacks}\right]$$

Substitution



- If we reduce ice cream sales, will the number of attacks decrease?
- Consider $\mathsf{M}_1 = \mathsf{M}\Big[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1)\Big]$ and $\mathsf{M}_2 = \mathsf{M}\Big[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_2)\Big]$

$$\mathbb{E}_{\mathsf{M}_1}\Big[\text{Shark attacks}\Big] - \mathbb{E}_{\mathsf{M}_2}\Big[\text{Shark attacks}\Big] = 0$$

Counterfactuals

- Intervention: if we reduce ice cream sales, will the number of attacks decrease?
- Counterfactual: 150 people were attacked this month. What if ice cream sales were reduced during this month?

Counterfactuals

- Intervention: if we reduce ice cream sales, will the number of attacks decrease?
- Counterfactual: 150 people were attacked this month. What if ice cream sales were reduced during this month?
- Are counterfactuals that harder than interventions?

Counterfactuals

- Intervention: if we reduce ice cream sales, will the number of attacks decrease?
- Counterfactual: 150 people were attacked this month. What if ice cream sales were reduced during this month?
- Are counterfactuals that harder than interventions?
- Intervention: $\mathsf{M}\left[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1)\right]$
- What if we consider $\mathsf{M}^* = \mathsf{M}\left[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1), \text{Shark attacks} = 150\right]$?

Counterfactuals

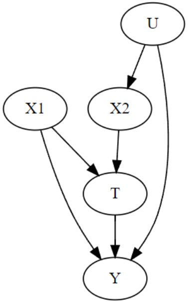
- Intervention: if we reduce ice cream sales, will the number of attacks decrease?
- Counterfactual: 150 people were attacked this month. What if ice cream sales were reduced during this month?
- Are counterfactuals that harder than interventions?
- Intervention: $\mathsf{M}\Big[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1)\Big]$
- What if we consider $\mathsf{M}^* = \mathsf{M}\Big[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1), \text{Shark attacks} = 150\Big]$?
- That's not the answer to counterfactual!
- $\mathbb{E}_{\mathsf{M}^*}\Big[\text{Shark attacks}\Big] = 150$

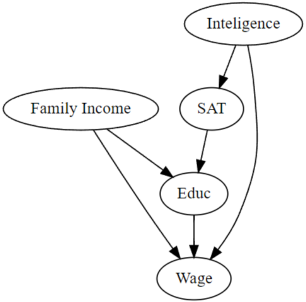
Counterfactuals

- Intervention: if we reduce ice cream sales, will the number of attacks decrease?
- Counterfactual: 150 people were attacked this month. What if ice cream sales were reduced during this month?
- Are counterfactuals that harder than interventions?
- Intervention: $\mathsf{M}\left[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1)\right]$
- What if we consider $\mathsf{M}^* = \mathsf{M}\left[\mathrm{do}(\text{Ice cream sales} := \mathsf{x}_1), \text{Shark attacks} = 150\right]$?
- That's not the answer to counterfactual!
- $\mathbb{E}_{\mathsf{M}^*}\left[\text{Shark attacks}\right] = 150$
- Graphical causal models can't handle counterfactuals!

Structured causal models

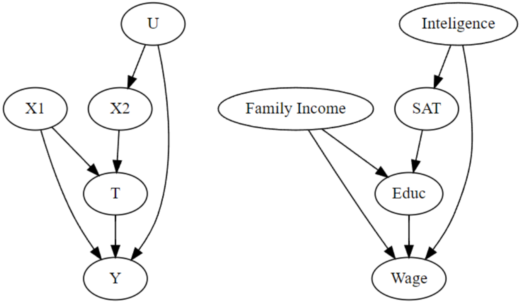
- Two groups of random variables: $X = (X_1, ..., X_n)$ are variables and $U = (U_1, ..., U_n)$ are noises
- $X_i = f_i(PA_i, U_i)$, where f_i is a deterministic function
- $P(U_1, ..., U_n) = P(U_1)...P(U_n)$ is given
- Notice that X is a function of U





Structured causal models

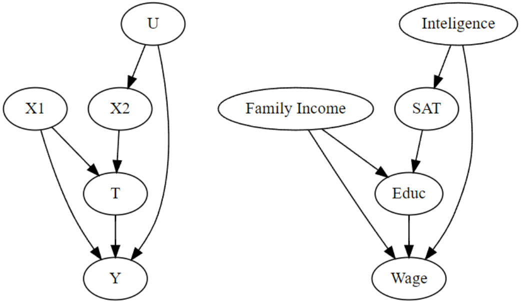
- Two groups of random variables:
  $X = (X_1, ..., X_n)$ are variables and
  $U = (U_1, ..., U_n)$ are noises
- $X_i = f_i(PA_i, U_i)$, where f_i is a deterministic function
- $P(U_1, ..., U_n) = P(U_1)...P(U_n)$ is given
- Notice that X is a function of U



- In contrast, graphical causal models define only $P(X_i|PA_i)$, so SCM is a generalization

Structured causal models

- Two groups of random variables:
  $X = (X_1, \ldots, X_n)$ are variables and
  $U = (U_1, \ldots, U_n)$ are noises
- $X_i = f_i(PA_i, U_i)$, where f_i is a deterministic function
- $P(U_1, \ldots, U_n) = P(U_1) \ldots P(U_n)$ is given
- Notice that X is a function of U



- In contrast, graphical causal models define only $P(X_i|PA_i)$, so SCM is a generalization
- 150 people were attacked this month. What if ice cream sales were reduced during this month?
- $P(U_1, \ldots, U_n) := P(U_1, \ldots, U_n | \text{Shark attacks} = 150)$
- To answer the counterfactual perform an intervention in the obtained GCM model

How to compute counterfactuals

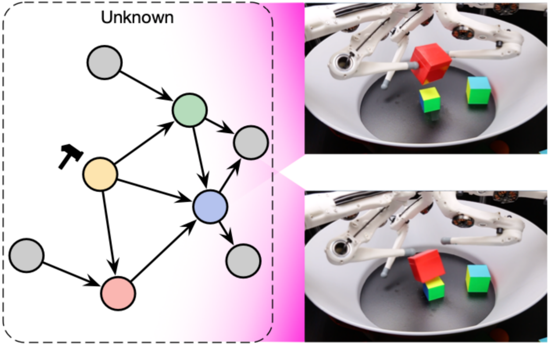
Given a structural causal model M, an observed event E, a substitution $T := t$ and target variable Y, we define the counterfactual $Y_{T:=t}(E)$ by the following three step procedure:

1. Condition the joint distribution of $U = (U_1, ..., U_n)$ on event E: $P(U') = P(U|E)$
2. Perform a substitution $T := t$ in the structural causal model M resulting in the model $M' = M\left[\mathrm{do}(T := t)\right]$
3. Compute target counterfactual $Y_{T:=t}(E)$ by using U' in M' instead of U

In general, the counterfactual $Y_{T:=t}(E)$ is a random variable that varies with U'.

SMS hypothesis

Sparse Mechanism Shift Hypothesis: small distribution changes lead to local changes in disentangled representation, i.e., they usually not affect all factors simultaneously



Disentangled representation:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|PA_i)$$

ICM Principle

Independent Causal Mechanisms Principle: the causal generative process of a system's variables is composed of autonomous modules.

ICM Principle implies that $P(X_i|PA_i)$ does not influence or give information about $P(X_j|PA_j)$ if $i \neq j$.

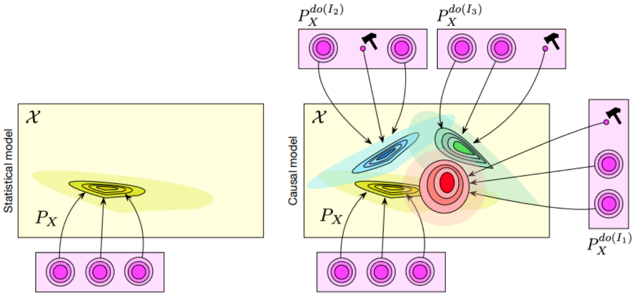
In SCMs this is achieved by joint independence of $U_1, ..., U_n$.

Comparison

	Predict in i.i.d setting	Predict under distr. shift or intervention	Answer counterfactuals	Learn from data
Statistical	yes	no	no	yes
Graphical causal	yes	yes	no	?
Structured causal	yes	yes	yes	?

Comparison

	Predict in i.i.d setting	Predict under distr. shift or intervention	Answer counterfactuals	Learn from data
Statistical	yes	no	no	yes
Graphical causal	yes	yes	no	?
Structured causal	yes	yes	yes	?



While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention

Learn SCM from data

$X = (X_1, ..., X_d)$ — an image

$S = (S_1, ..., S_n)$ — causal variables, i.e.,

$$S_i = f_i(PA_i, U_i)$$

1. Encoder $q : \mathbb{R}^d \to \mathbb{R}^n$. We expect $q(X) = U = (U_1, ..., U_n)$ representation to comprise noise variables
2. Mapping $f(U)$ which is expected to transform U into S
3. Decoder $p : \mathbb{R}^n \to \mathbb{R}^d$ which is expected to transform S into X

Thus $p \circ f \circ q$ is an autoencoder but f contains information about structural assignments $f_1, ..., f_n$

Thank you for your attention!

Sources

- https://mlstory.org/causal.html
- Towards Causal Representation Learning(arXiv: https://arxiv.org/abs/2102.11107)