Causal Normalizing flow

January 08, 2024

Seoul National University

Table of Contents

1. Contribution
2. Structural causal model(SCM)
3. Autoregressive normalizing flow (ANF)
4. Causal Autoregressive normalizing flow (CAF)
5. Intervention and Counterfactual
6. Experiments

Contribution

- Introduce causal autoregressive flows(CAFs) which reflect the causal structure of data.
- Through CAFs, enable sampling an intervened sample and counterfactual sample.

Table of Contents

1. Contribution
2. Structural causal model(SCM)
3. Autoregressive normalizing flow (ANF)
4. Causal Autoregressive normalizing flow (CAF)
5. Intervention and Counterfactual
6. Experiments

Structural causal model

- A structural causal model (SCM) is a tuple $\mathcal{M} = \left(\tilde{\mathrm{f}}, P_{\mathrm{u}}\right)$ describing a data-generating process that transforms a set of d-dimensional latent variables, $\mathrm{u} \sim P_{\mathrm{u}}$, into a set of d-dimensional observed data, x, according to $\tilde{\mathrm{f}} = (\tilde{f}_1, \cdots, \tilde{f}_d) : \mathbb{R}^d \to \mathbb{R}^d$.
- Specifically, $\mathrm{x} = (\mathrm{x}_1, \cdots, \mathrm{x}_d)$ is computed as follows:

$$\mathrm{u} := (\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_d) \sim P_{\mathrm{u}}, \quad \mathrm{x}_i = \tilde{f}_i\left(\mathrm{x}_{\mathrm{pa}_i}, \mathrm{u}_i\right), \qquad \text{for } i = 1, 2, \ldots, d.$$

where $\mathrm{x}_{\mathrm{pa}_i}$ is parents of x_i which directly cause x_i.

Adjacency matrix

- Define the adjacency matrix of the causal graph as $A = (A_{ij}) \in \{0,1\}^{d \times d}$,

$$A_{ij} = \mathbb{I}\left(\frac{\partial \tilde{f}_i(\mathrm{x}_{\mathrm{pa}_i}, \mathrm{u}_i)}{\partial x_j} \neq 0\right)$$

  where $\mathbb{I}$ is the indicator function.
- We assume that A is acyclic.
- Also assume that the x are sorted according to a causal ordering.

Table of Contents

1. Contribution
2. Structural causal model(SCM)
3. Autoregressive normalizing flow (ANF)
4. Causal Autoregressive normalizing flow (CAF)
5. Intervention and Counterfactual
6. Experiments

Normalizing flow

- Normalizing flows are generative models that express the probability density of data using the change-of-variables rule.
- Given an observed data $\mathrm{x} \in \mathbb{R}^d$, a normalizing flow is a neural network with parameters θ that takes x as input, and outputs

$$T_{\boldsymbol{\theta}}(\mathrm{x}) =: \mathrm{u} \sim P_{\mathrm{u}}$$

$$\log p(\mathrm{x}) = \log p\left(T_{\boldsymbol{\theta}}(\mathrm{x})\right) + \log\left|\det\left(\nabla_{\mathrm{x}} T_{\boldsymbol{\theta}}(\mathrm{x})\right)\right|$$

  where P_{u} is a base distribution that is easy to sample.
- In normalizing flow, estimate θ by maximizing $\log p(\mathrm{x})$.

Autoregressive normalizing flow

- To sample data from u, $T_\theta^{-1}(u)$ should be exist and determinant of Jacobian matrix $\nabla_{\mathrm{x}} T_\theta(\mathrm{x})$ should be tractable.
- Autoregressive normalizing flows(ANFs) are models that satisfy the above conditions.
- ANFs are deep neural networks which is i-th output of each layer l denoted by z_i^l is computed as

$$\mathrm{z}_i^l := \tau_i^l\left(\mathrm{z}_i^{l-1}; \mathrm{h}_i^l\right), \quad \text{where} \quad \mathrm{h}_i^l := c_i^l\left(\mathrm{z}_{1:i-1}^{l-1}\right)$$

  and where τ_i and c_i are termed the transformer and the conditioner, respectively.
- Note that Jacobian matrices of ANFs are triangular matrices.

Table of Contents

1. Contribution
2. Structural causal model(SCM)
3. Autoregressive normalizing flow (ANF)
4. Causal Autoregressive normalizing flow (CAF)
5. Intervention and Counterfactual
6. Experiments

ANFs as SCM

- Want to express SCM as ANFs when A is known.
- One of the conditions that ANFs can be expressed as SCM is x_i is a function of $\mathrm{x}_{\mathrm{pa}_i}$ which depend on $\mathrm{u}_{\mathrm{pa}_i}$, and u_i.

Theorem

Theorem
If a causal NF $T_{\boldsymbol{\theta}}$ satisfies an condition in previous slide, then $\nabla_{\mathrm{x}} T_{\boldsymbol{\theta}}(\mathrm{x}) \equiv \boldsymbol{I} - A$ and $\nabla_{\mathrm{u}} T_{\boldsymbol{\theta}}^{-1}(\mathrm{u}) \equiv \boldsymbol{I} + \sum_{n=1}^{\mathrm{diam}(A)} A^n$, where A is the causal adjacency matrix of $\mathcal{M}$.

where $\mathrm{diam}(A) = \min\{k : A^k = \tilde{0}\}$ when $\tilde{0}$ is zero matrix.

Model & Objective function

- (Abductive model) Causal NF from x to u

$$\mathrm{z}_i^l = \tau_i\left(\mathrm{z}_i^{l-1}; \mathrm{h}_i^l\right), \quad \text{where} \quad \mathrm{h}_i^l = c_i\left(\mathrm{z}_{\mathrm{pa}_i}^{l-1}\right)$$

- To penalize spurious correlations from x to $T_{\boldsymbol{\theta}}(\mathrm{x})$, add penalized term :

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \mathbb{E}_{\mathrm{x}}\left[-\log p\left(T_{\boldsymbol{\theta}}(\mathrm{x})\right) + \left\|\nabla_{\mathrm{x}} T_{\boldsymbol{\theta}}(\mathrm{x}) \odot (\tilde{1} - \boldsymbol{A})\right\|_2\right]$$

where $\tilde{1}$ is a matrix of ones.

Table of Contents

1. Contribution
2. Structural causal model(SCM)
3. Autoregressive normalizing flow (ANF)
4. Causal Autoregressive normalizing flow (CAF)
5. Intervention and Counterfactual
6. Experiments

Intervention

- The do-operator, denoted as $do\,(x_i = \alpha)$, is a mathematical operator that fixes the observational value $x_i = \alpha$, and thus removes any causal dependency on x_i.

Algorithm 1 Algorithm to sample from the interventional distribution, $P(\mathbf{x} \mid do(\mathrm{x}_i = \alpha))$.

function SAMPLEINTERVENEDDIST(i, α)
  $\mathbf{u} \sim P_{\mathbf{u}}$
  $\mathbf{x} \leftarrow T_{\boldsymbol{\theta}}^{-1}(\mathbf{u})$ ▷ Sample a value from the observational distribution.
  $\mathrm{x}_i \leftarrow \alpha$ ▷ Set x_i to the intervened value α.
  $\mathrm{u}_i \leftarrow T_{\boldsymbol{\theta}}(\mathbf{x})_i$ ▷ Change the i-th value of $\mathbf{u}$.
  $\mathbf{x} \leftarrow T_{\boldsymbol{\theta}}^{-1}(\mathbf{u})$
  return $\mathbf{x}$ ▷ Return the intervened sample.
end function

Figure 1: Do-operation

Counterfactual sample

Algorithm 2 Algorithm to sample from the counterfactual distribution, $P(\mathbf{x}^{\mathrm{cf}} \mid do(\mathrm{x}_i = \alpha), \mathbf{x}^{\mathrm{f}})$.

function GETCOUNTERFACTUAL($\mathbf{x}^{\mathrm{f}}$, i, α)
  $\mathbf{u} \leftarrow T_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{f}})$ ▷ Get $\mathbf{u}$ from the factual sample.
  $\mathrm{x}_i^{\mathrm{f}} \leftarrow \alpha$ ▷ Set x_i to the intervened value α.
  $\mathrm{u}_i \leftarrow T_{\boldsymbol{\theta}}(\mathbf{x}^{\mathrm{f}})_i$ ▷ Change the i-th value of $\mathbf{u}$.
  $\mathbf{x}^{\mathrm{cf}} \leftarrow T_{\boldsymbol{\theta}}^{-1}(\mathbf{u})$
  return $\mathbf{x}^{\mathrm{cf}}$ ▷ Return the counterfactual value.
end function

Figure 2: Counterfactual sample

Table of Contents

1. Contribution
2. Structural causal model(SCM)
3. Autoregressive normalizing flow (ANF)
4. Causal Autoregressive normalizing flow (CAF)
5. Intervention and Counterfactual
6. Experiments

3-CHAIN$_{\text{LIN}}$:

$$\begin{aligned}\tilde{f}_1(u_1) &= u_1 \\ \tilde{f}_2(x_1, u_2) &= 10 \cdot x_1 - u_2 \\ \tilde{f}_3(x_2, u_3) &= 0.25 \cdot x_2 + 2 \cdot u_3\end{aligned}$$

3-CHAIN$_{\text{NLIN}}$:

$$\begin{aligned}\tilde{f}_1(u_1) &= u_1 \\ \tilde{f}_2(x_1, u_2) &= e^{x_1/2} + u_2/4 \\ \tilde{f}_3(x_2, u_3) &= \frac{(x_2 - 5)^3}{15} + u_3\end{aligned}$$

4-CHAIN$_{\text{LIN}}$:

$$\begin{aligned}\tilde{f}_1(u_1) &= u_1 \\ \tilde{f}_2(x_1, u_2) &= 5 \cdot x_1 - u_2 \\ \tilde{f}_3(x_2, u_3) &= -0.5 \cdot x_2 - 1.5 \cdot u_3 \\ \tilde{f}_4(x_3, u_4) &= x_3 + u_4\end{aligned}$$

Figure 3: Counterfactual sample

Experiments

		Performance			Time Evaluation (μs)		
Dataset	Model	KL	ATE_{RMSE}	CF_{RMSE}	Training	Evaluation	Sampling
3-CHAIN LIN [36]	Causal NF	$\mathbf{0.00_{0.00}}$	$\mathbf{0.05_{0.01}}$	$\mathbf{0.04_{0.01}}$	$\mathbf{0.41_{0.06}}$	$\mathbf{0.48_{0.10}}$	$\mathbf{0.76_{0.06}}$
	CAREFL†	$\mathbf{0.00_{0.00}}$	$0.20_{0.13}$	$0.20_{0.09}$	$0.68_{0.24}$	$0.97_{0.33}$	$1.94_{0.77}$
	VACA	$4.44_{1.03}$	$5.76_{0.07}$	$4.98_{0.10}$	$36.19_{1.54}$	$28.33_{0.72}$	$75.34_{4.58}$
3-CHAIN NLIN [36]	Causal NF	$\mathbf{0.00_{0.00}}$	$\mathbf{0.03_{0.01}}$	$\mathbf{0.02_{0.01}}$	$\mathbf{0.52_{0.06}}$	$\mathbf{0.56_{0.03}}$	$\mathbf{1.02_{0.05}}$
	CAREFL†	$\mathbf{0.00_{0.00}}$	$\mathbf{0.05_{0.02}}$	$0.04_{0.02}$	$\mathbf{0.60_{0.22}}$	$0.84_{0.22}$	$1.66_{0.41}$
	VACA	$12.82_{1.00}$	$1.54_{0.03}$	$1.32_{0.02}$	$39.45_{4.12}$	$30.93_{2.30}$	$84.36_{9.60}$
4-CHAIN LIN	Causal NF	$\mathbf{0.00_{0.00}}$	$\mathbf{0.07_{0.02}}$	$\mathbf{0.04_{0.01}}$	$\mathbf{0.56_{0.08}}$	$\mathbf{0.62_{0.15}}$	$\mathbf{1.54_{0.40}}$
	CAREFL†	$\mathbf{0.00_{0.00}}$	$0.16_{0.07}$	$0.14_{0.04}$	$\mathbf{0.70_{0.28}}$	$0.99_{0.20}$	$2.85_{0.54}$
	VACA	$13.14_{0.73}$	$3.82_{0.01}$	$3.72_{0.05}$	$61.85_{5.06}$	$49.31_{4.11}$	$92.06_{7.93}$

Figure 4: Result 1

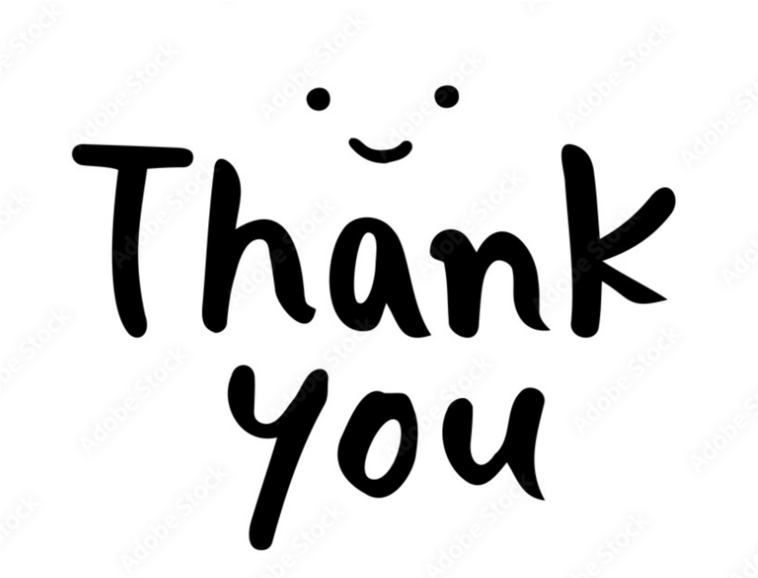
Experiments - Fairness

German data set when a sensitive variable is sex.

Table 3: Accuracy, F1-score, and counterfactual unfairness of the audited classifiers. Causal NFs enable both fair classifiers and accurate unfairness metrics. Results are averaged on five runs.

	Logistic classifier				SVM classifier			
	full	unaware	fair $\mathbf{x}$	fair $\mathbf{u}$	full	unaware	fair $\mathbf{x}$	fair $\mathbf{u}$
f1	$72.28_{6.16}$	$72.37_{4.90}$	$59.66_{8.57}$	$73.08_{4.38}$	$76.04_{2.86}$	$76.80_{5.82}$	$68.28_{5.74}$	$77.39_{1.52}$
accuracy	$67.00_{3.83}$	$66.75_{2.63}$	$54.75_{5.91}$	$66.50_{3.70}$	$69.50_{3.11}$	$71.00_{3.83}$	$59.25_{2.99}$	$69.75_{1.26}$
unfairness	$5.84_{2.93}$	$2.81_{0.72}$	$0.00_{0.00}$	$0.00_{0.00}$	$6.65_{2.45}$	$2.78_{0.40}$	$0.00_{0.00}$	$0.00_{0.00}$

Figure 5: Result 2


Thank
you