Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

归一化流在表格数据生成中的应用

硕士研究生 徐泽豪

2024年04月06日

问题回溯

- 总结反思
  - PPT部分内容制作不严谨，未介绍前沿算法
  - 详略不得当，基础知识部分占比过多
  - 算法讲解部分未标注符号含义

- 相关内容
  - 2023.04.16 万韵伟：《扩散模型加速采样方法与应用》
  - 2023.08.14 徐泽豪：《表格数据生成：GAN的演进与未来》
  - 2023.11.06 吴肖龙：《智能模型的不确定性估计》
  - 2024.01.07 段学明：《DNN中的理论可解释性》

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - DP-Hflow
  - CeFlow
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 1. 理解归一化流模型的基本概念
  - 2. 理解归一化流模型在表格数据生成相关任务中的基本应用
  - 3. 了解归一化流模型的前沿发展

目标内涵

- 研究目标
  - 以表格数据为研究对象，面向隐私保护/反事实解释任务
  - 结合归一化流、变分去量化、条件流高斯混合模型技术
  - 探讨归一化流如何提高表格数据概率建模相关应用的准确性
- 内涵解析
  - 表格数据：以行和列的形式存储的结构化数据，每列代表一维属性，每行代表一条数据样本
  - 归一化流：利用一系列可逆变换将简单分布映射为复杂数据分布的生成模型，具有精确计算概率密度的能力
  - 概率建模：精确捕获和表达数据的复杂分布特性，为后续的数据生成、分析和预测等下游任务建立基础

背景意义

- 研究背景
  - 表格数据在商业智能、医疗健康、金融分析等领域广泛应用
  - 这些领域的复杂数据分布特性对概率建模提出了挑战
- 研究意义
  - 通过精准映射简单分布至复杂分布，对于理解和模拟真实数据复杂性，改进数据生成质量、增强模型解释性具有重要作用
  - 直接使用真实表格数据进行研究或商业分析面临隐私泄露风险，需要在不泄露个人隐私的前提下利用敏感数据
  - 现有的机器学习模型缺乏足够的透明度和解释性，为表格数据生成反事实样本有助于揭示模型决策背后的因果关系，具有重要的实用价值

研究历史

2014 Rezende等人通过将简单的初始密度通过一系列可逆变换转换成更复杂的密度，提出了可扩展的近似后验分布的归一化流方法

2016 Dinh等人通过引入真实值非体积保持（Real NVP）变换，扩展了可学习概率模型的空间，使归一化流具备精确对数似然计算、维持可解释潜在空间的能力

2018 Grover等人提出了Flow-GAN，结合最大似然和对抗进行混合训练，弥补了GAN回避显式密度表征，难以进行定量评估的不足

2019 Durkan等人提出基于单调有理二次样条的全微分模块，证明了神经样条流在密度估计、变分推断和生成模型中的有效性

2020 Izmailov等人提出半监督学习方法FlowGMM，融合归一化流和高斯混合模型对标记和未标记数据进行统一处理，增强了模型的可解释性

2021 Caterini等人通过学习嵌入在高维空间中的低维流形上的分布来解决建模不匹配的问题

2022 Lee等人通过结合差分隐私和归一化流模型提出了一种表格数据生成方法，能够在同时保护数据隐私的同时维持数据的可用性

2023 Duong等人利用归一化流生成与给定实例相近但在某些特征上不同的数据点，以提供直观的模型决策解释，不仅提高了反事实解释的质量，还保证了模型的稳健性

归一化流在表格数据中的应用
- 隐私保护
  - 差分隐私技术
  - 自回归样条变换
- 反事实解释
  - 模型决策解释
  - 条件流高斯混合

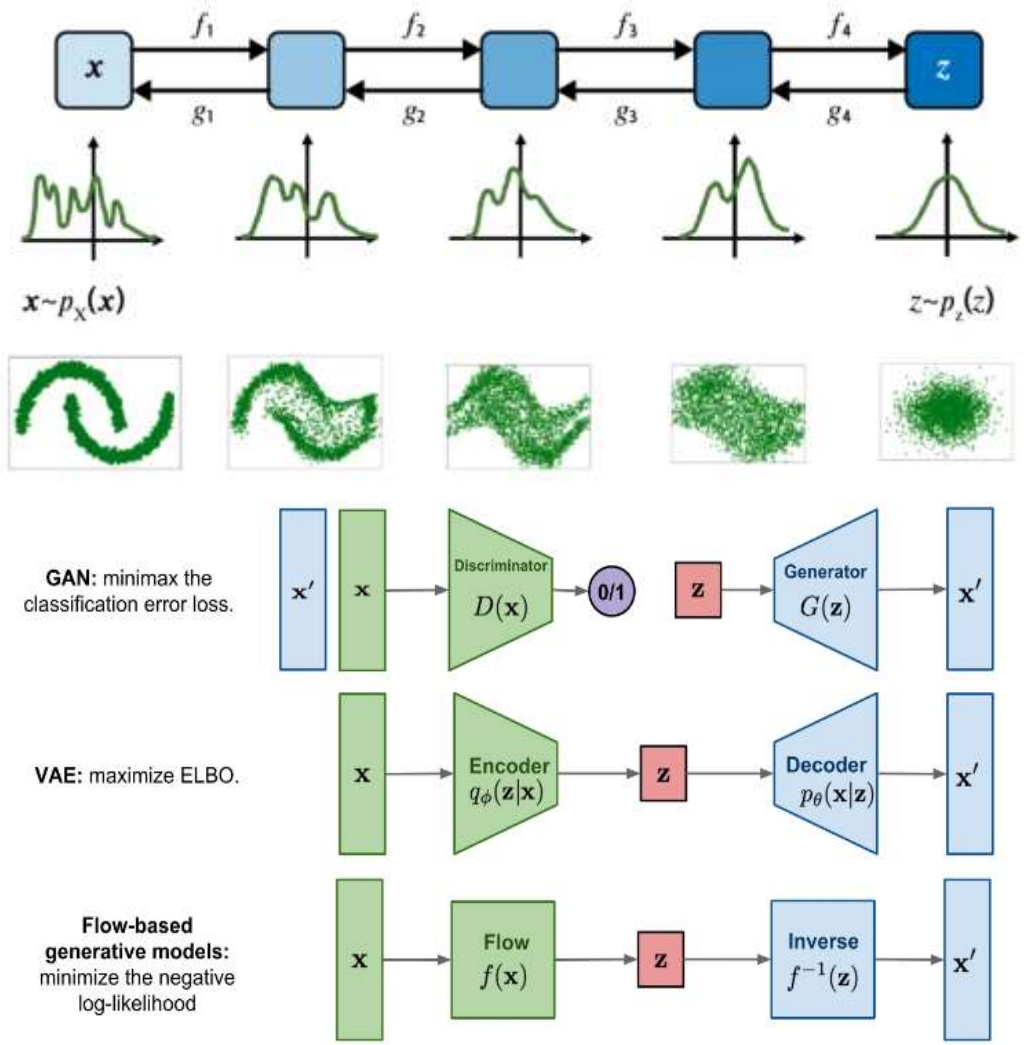
研究现状

- 基于隐私保护的表格数据生成
  - 平衡隐私保护和数据可用性：现有方法难以在保护个人隐私的同时，确保合成数据保持足够的统计特性
- 表格数据的反事实解释生成
  - 反事实样本的稀疏性：现有方法难以通过在特定特征上略微改变，生成可行的反事实样本，稳定性不足，不能确保只改变少数几个特征

			Private				Non-Private		
		Real	DP-CGAN	GS-WGAN	DP-MERF	**DP-HFlow**	CTGAN	DP-MERF	**DP-HFlow**
Macro-F1	Adult	0.79±0.02	0.46±0.07	0.42±0.09	0.37±0.15	**0.56±0.07**	0.74±0.02	0.41±0.16	**0.75±0.01**
	Census	0.75±0.01	0.45±0.09	0.44±0.13	0.48±0.14	**0.52±0.03**	0.67±0.04	0.50±0.14	**0.70±0.04**
	Covertype	0.77±0.16	0.15±0.03	0.11±0.03	**0.31±0.05**	0.22±0.03	0.22±0.04	0.29±0.05	**0.49±0.04**
	Intrusion	0.86±0.07	0.19±0.09	0.13±0.08	0.36±0.05	**0.40±0.03**	**0.54±0.05**	0.38±0.06	0.46±0.06

知识基础 归一化流

- 归一化流
  - 利用可逆变换将简单概率分布转换为复杂概率分布，也可逆向变换
  - 能够通过可逆变换和雅可比行列式，直接计算变换后的数据点的对数似然
- 与其他生成模型的区别
  - GAN和VAE两种方法都不能对生成数据点的概率密度进行精确评估
  - Diffusion通过随机变换实现分布变换，归一化流更注重变换的可解释性，每一步变换都是确定的

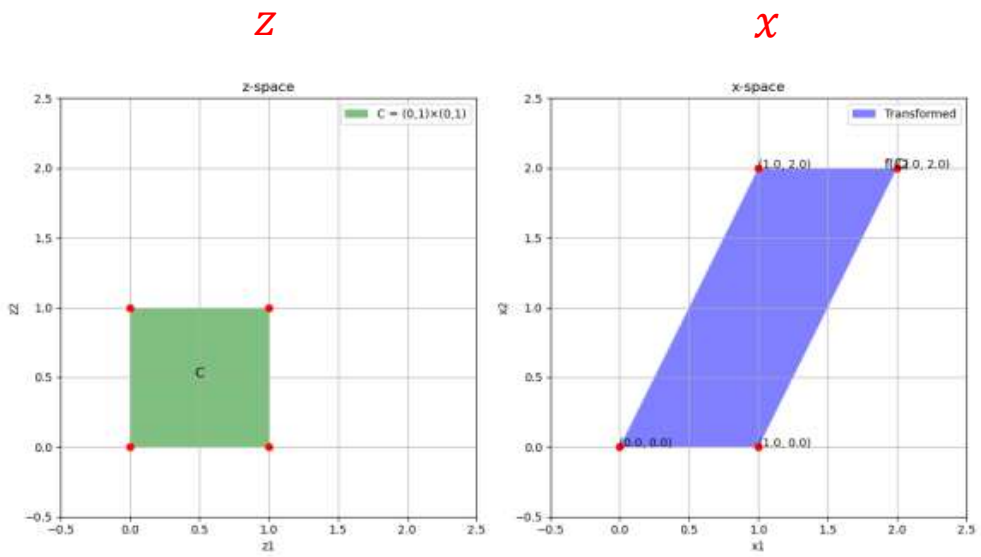


- 雅可比矩阵的性质
  - 设有一个变换 f 由 z 映射到 x，$x = f(z)$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = f(z) = \begin{bmatrix} z_1 + z_2 \\ 2z_1 \end{bmatrix}$$

  - 那么变换 f 的雅可比矩阵 J_f 为

$$J_f = \begin{bmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix}$$



  - 变换 f 的逆变换 f^{-1}，$z = f^{-1}(x)$，f^{-1} 的雅可比矩阵 $J_{f^{-1}}$

$$z = \begin{bmatrix} \frac{1}{2}x_2 \\ x_1 - \frac{1}{2}x_2 \end{bmatrix} \qquad J_{f^{-1}} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} \\ 1 & -\frac{1}{2} \end{bmatrix}$$

$$\Rightarrow \quad |\det J_f| = |\det J_{f^{-1}}|^{-1}$$

变换f 与其逆变换的雅可比行列式互为倒数，确保概率密度转换时保持总概率不变

- 变量变换定理
  - 已知变量 z 到变量 x 的可逆变换 f，其概率密度函数分别是$p_z(z)$，$p_x(x)$

$$\int_x p_x(x)\,dx = 1 = \int_z p_z(z)\,dz$$

  - z 与 x 一一对应，单位区域内积分相等

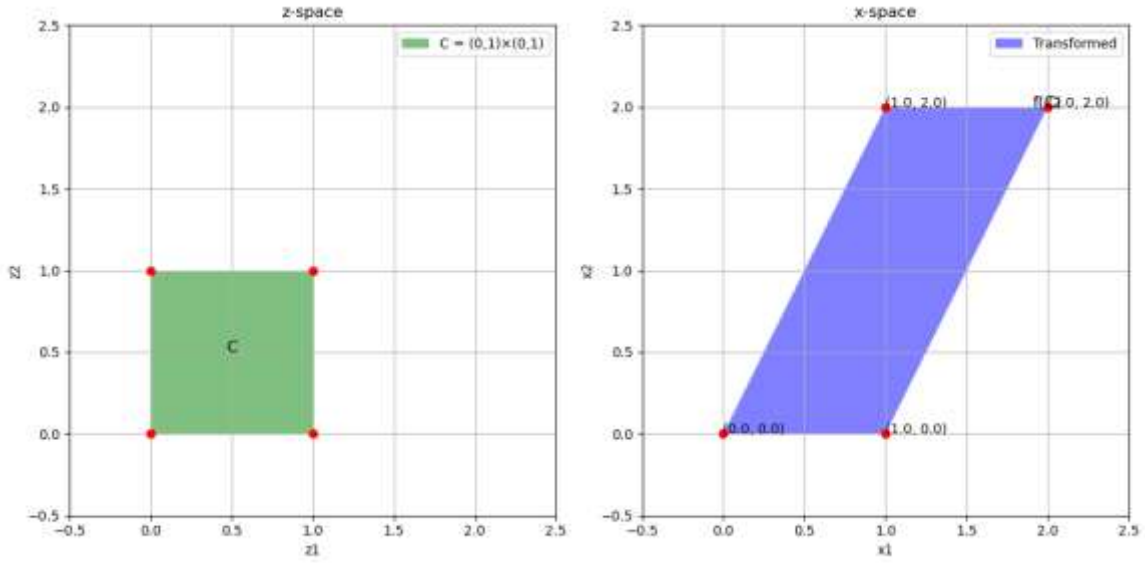
$$|p_x(x)\,dx| = |p_z(z)\,dz|$$

$$p_x(x) = p_z(z)\frac{dz}{dx}$$

$$p_x(x) = p_z(z)\left|\frac{\partial f^{-1}(x)}{\partial x}\right| \qquad x = f(z)$$

$$p_x(x) = p_z(f^{-1}(x))\left|\det J_{f^{-1}}(x)\right| \qquad \text{概率密度非负}$$

$$|\det J_{f^{-1}}(x)| = |\frac{1}{2}\cdot(-1) - 0\cdot 1| = \frac{1}{2}$$



$$p_x(x) = p_z\left(\begin{bmatrix}\frac{1}{2}x_2\\ x_1 - \frac{1}{2}x_2\end{bmatrix}\right)\cdot\frac{1}{2} = \frac{1}{2}$$

- 归一化流的训练
  - 已知简单分布

$$p_x(x) = p_z(z)\left|\frac{\partial f^{-1}(x)}{\partial x}\right| \qquad \text{变换确定}$$

$$p_x(x) = p_z(f^{-1}(x))\left|\det J_{f^{-1}}(x)\right|$$

  - 推广到一般

$$p_X(x) = p_Z(f_\theta(x)) \cdot \left|\det\left(\frac{\partial f_\theta}{\partial x}\right)\right| \qquad \text{变换未知，参数为}\theta$$

$$\propto \log(p_Z(f_\theta(x))) + \log\left(\left|\det\left(\frac{\partial f_\theta}{\partial x}\right)\right|\right)$$

  - 有N个训练数据点$D = \{x_n\}_{n=1}^N$，模型的参数θ可以通过最大化对数似然来训练

$$\theta = \arg\max_\theta \sum_{n=1}^{N}\left(\log(p_Z(f_\theta(x_n))) + \log\left(\left|\det\left(\frac{\partial f_\theta}{\partial x_n}\right)\right|\right)\right) \qquad x_n\text{为向量}$$

Differentially Private Normalizing Flows for Synthetic Tabular Data Generation

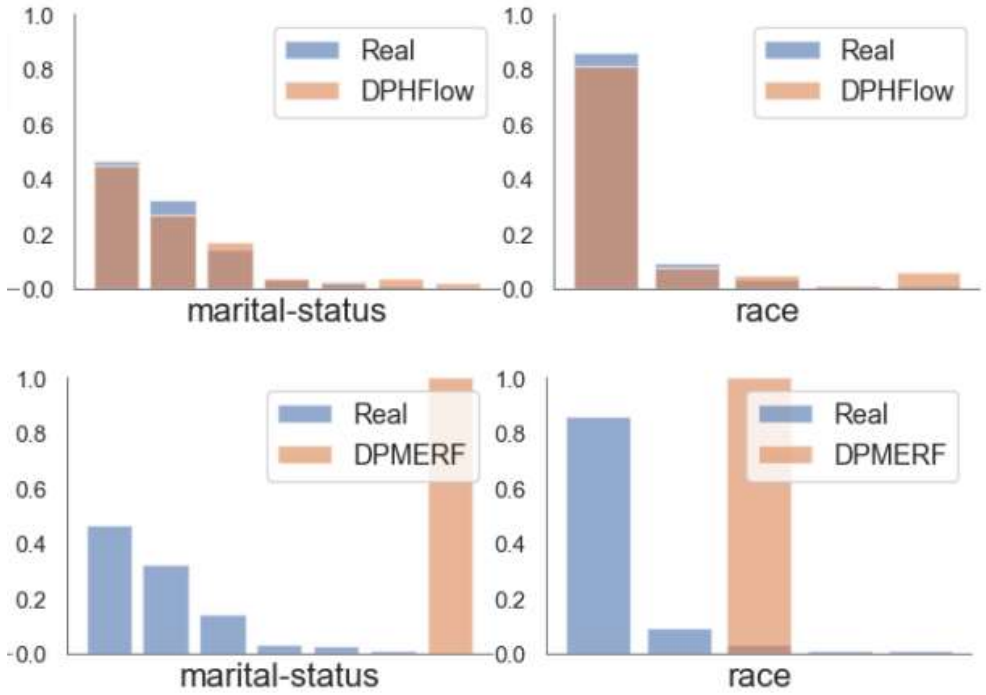
算法原理 TIPO

T	目标	在差分隐私下使用归一化流模型模拟含有连续和离散变量的表格数据
I	输入	Adult（14维，二分类）、Census（40维，二分类）、Covertype（54维，多分类）、Intrusion（40维，多分类/不平衡）分类数据集
P	处理	1.使用变分去量化层将离散的分类变量转换为连续变量 2.引入条件样条流（Conditional Spline Flow）模拟复杂的多模式密度 3.通过细粒度梯度剪辑技术来控制训练过程中的隐私损失
O	输出	接近真实数据集分布且维持数据隐私的合成数据

P	问题	其他生成方法隐式地对概率密度建模或对其近似，无法对概率密度进行准确评估；归一化流难以对离散数据进行建模
C	条件	严格的差分隐私保证
D	难点	如何使用归一化流处理混合类型（同时包含连续/离散变量）的数据集
L	水平	AAAI 2022

算法原理　变分去量化

- 连续模型的局限性
  - 归一化流依赖于变量的连续变化来建模数据分布，而离散数据没有连续的PDF
  - 使用最大似然估计进行训练时，会使模型聚焦于数据集中的少数几个点，从而给这些点分配异常高的概率
  - 导致离散变量被模拟为“退化”分布
- “退化”分布的影响
  - 模型被过度吸引到少数“高概率”点上，无法准确捕捉数据集的整体特征
  - 导致模型泛化能力下降，生成的数据缺乏多样性

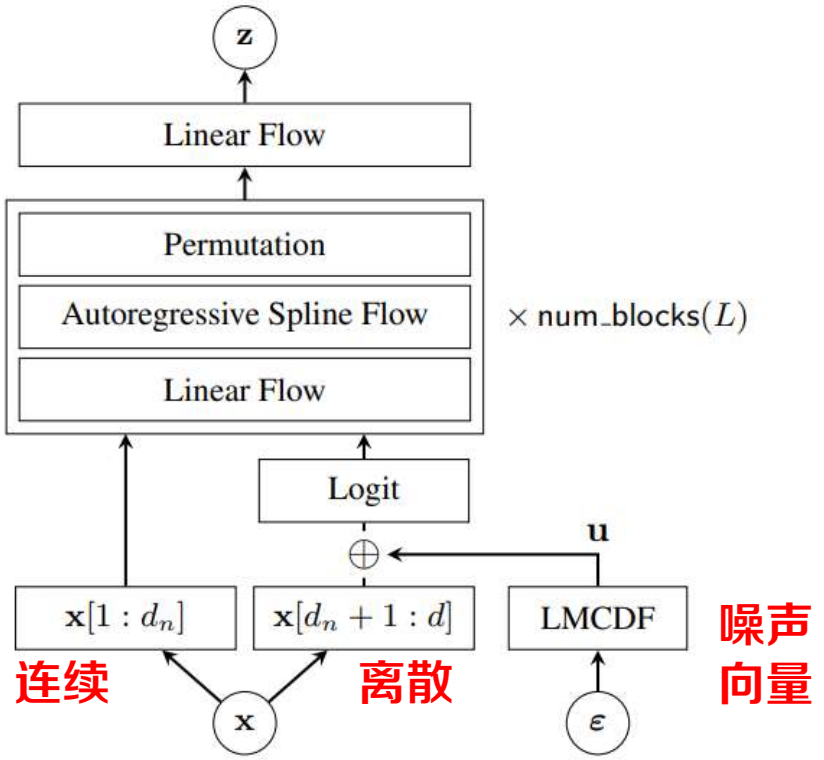


算法原理　变分去量化

- 变分去量化（Variational Dequantization）
  - 对于离散变量的每个值，通过添加连续噪声将其“扩散”到连续的范围内
  - 与均匀去量化引入来源于均匀分布的噪声不同
- 逻辑混合累积分布函数
  - 用于生成噪声向量，其参数从数据中学习

$$F_{\text{LMCDF}}(x;\pi,\mu,s)=\sum_{i=1}^{M}\pi_i\,\sigma((x-\mu_i)\cdot\exp(-s_i))$$

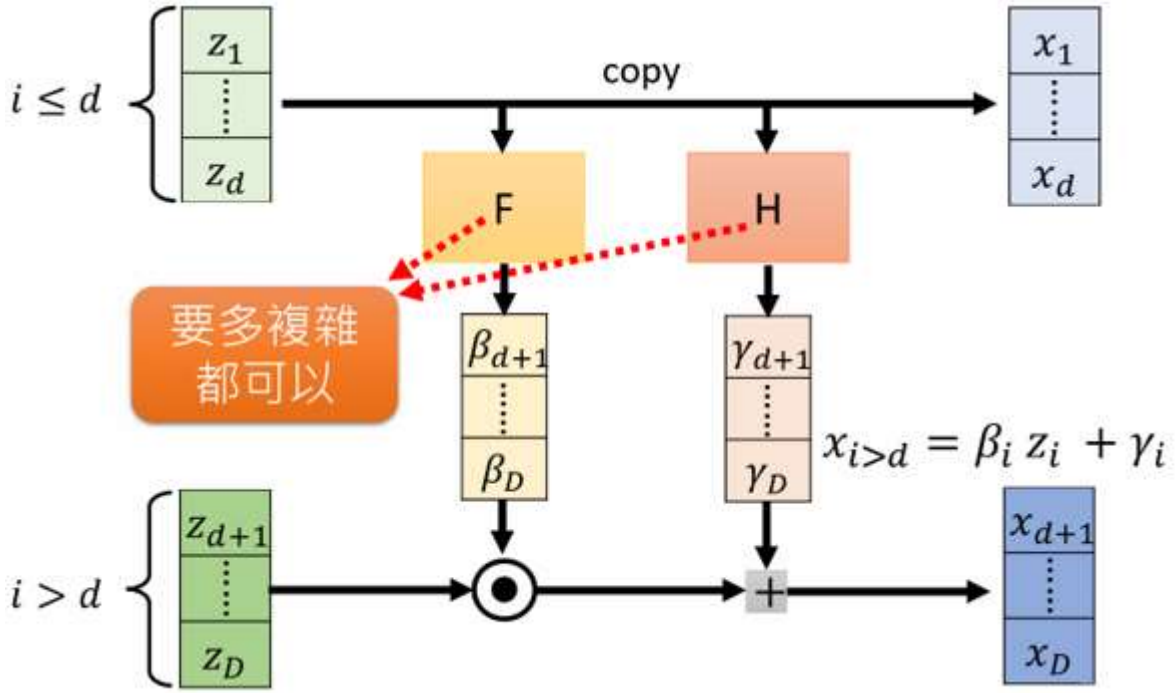
x：输入变量，分布函数的特定点
π_i：第i个部分的混合权重
σ：逻辑斯蒂（Sigmoid）函数
μ_i：第i个逻辑斯蒂分布的位置参数，分布的平均值
s_i：第i个逻辑斯蒂分布的尺度参数，影响分布的宽度或标准差
M：混合模型中逻辑斯蒂分布的总数

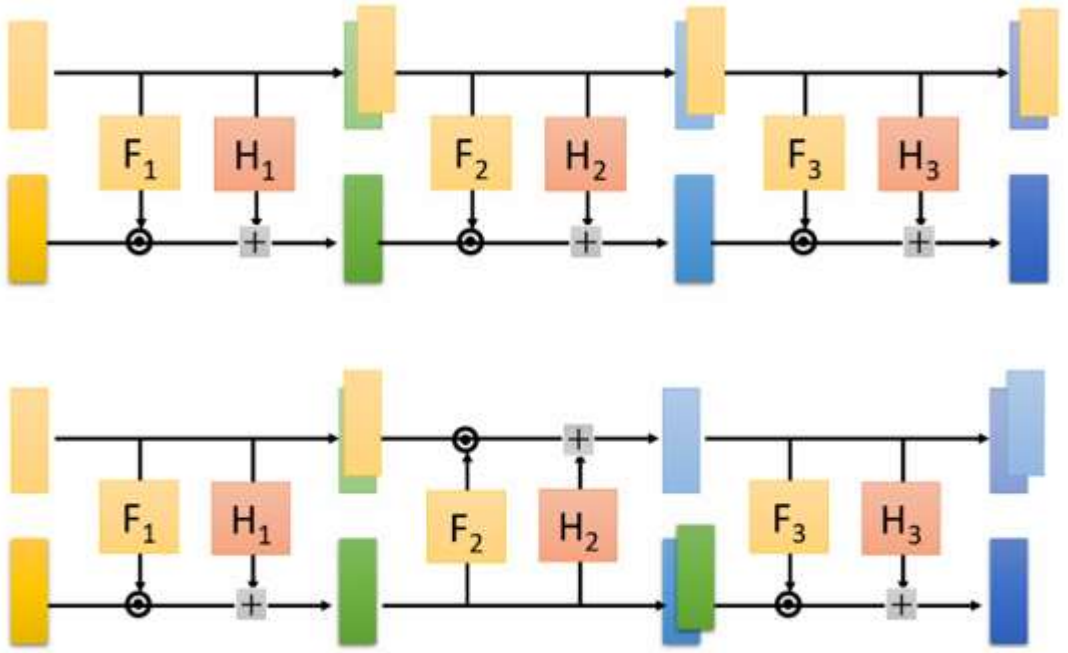


Logit模块有什么作用?

算法原理　自回归样条变换

- 归一化流具有更高的参数复杂度
  - 为了精确地建模和捕捉复杂数据分布的细节特征，需要使用多个复杂的变换层
  - 每个子流的设计需要足够复杂以确保模型能够捕捉变量之间的依赖关系





算法原理　自回归样条变换

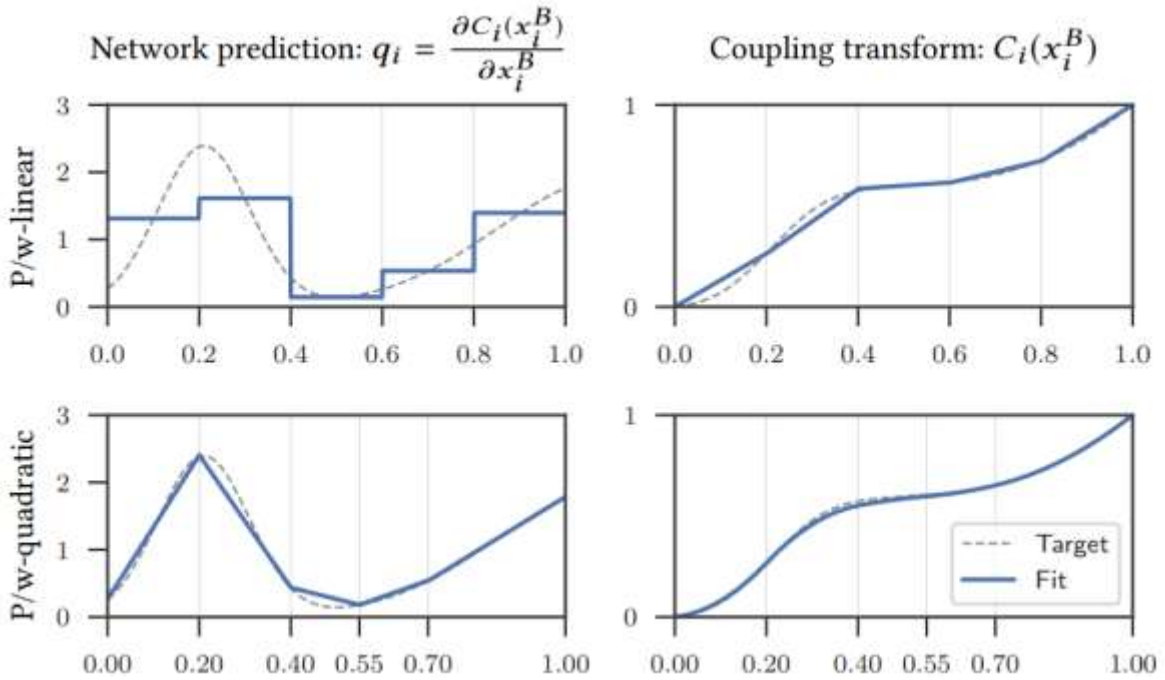
- 自回归样条变换（Autoregressive Spline Transformation）
  - 自回归在归一化流中用于建模变量间的复杂非线性关系
  - 通过使用样条函数，可以创建平滑且可微的映射，适应数据中的非线性结构

- 算法步骤
  - 确定节点

$x(k-1) < x(k), y(k-1) < y(k) \quad k = 1, \ldots, K$

$x(0) = y(0) = -B, x(K) = y(K) = B \quad B > 0$

  - 定义有理二次样条函数
  - 构建自回归模型

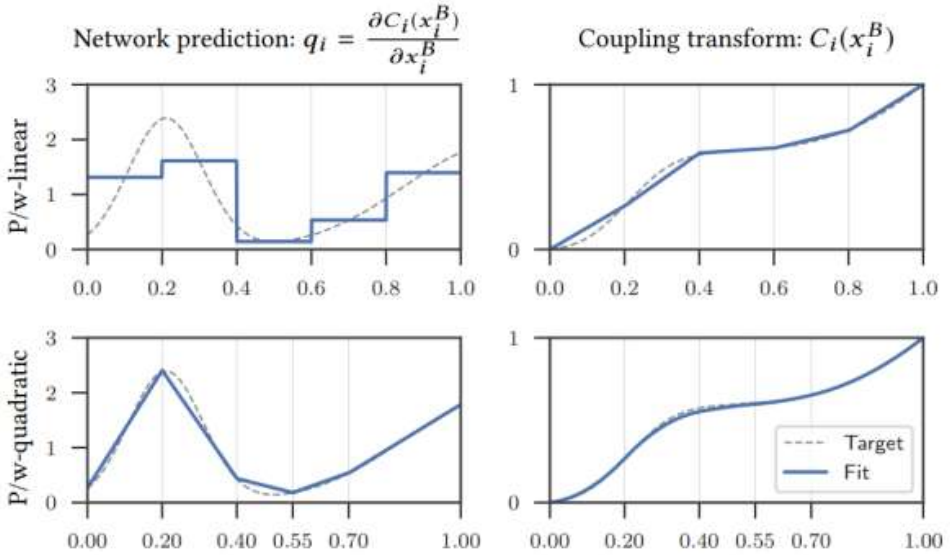


分段拟合效果

有理二次样条函数通过保持单调性，确保分布变换的可逆性

算法原理　自回归样条变换

- 局部坐标映射
  - 对于给定的输入x，首先确定它所在的区间
  - 计算ξ，也就是x在该区间内的相对位置
  - 用节点$y^{(k+1)}$和$y^{(k)}$以及导数$\delta^{(k+1)}$和$\delta^{(k)}$计算样条函数$S_k(x)$在x处的值
- 区间k上的单调有理二次样条函数



$$S_k(x) = y^{(k)} + \frac{\left(y^{(k+1)} - y^{(k)}\right)\left[\Delta^{(k)}\xi^2 + \delta^{(k)}\xi(1-\xi)\right]}{\Delta^{(k)} + \left[\delta^{(k+1)} + \delta^{(k)} - 2\Delta^{(k)}\right]\xi(1-\xi)}$$

$$\Delta^{(k)} = \frac{y^{(k+1)} - y^{(k)}}{x^{(k+1)} - x^{(k)}}$$ 斜率

$$\xi(x) = \frac{x - x(k)}{x^{(k+1)} - x^{(k)}}$$ 坐标

这种变换使得模型能够学习任何一维概率分布，并且可通过自回归的方式扩展到多维

算法原理　细粒度梯度剪辑

- 单元剪辑
  - 为每个神经元**单独设定**剪辑阈值，用于限制单个数据点对总梯度的贡献，以避免敏感信息的泄露

$$C_\ell[i] = C_\ell \sqrt{\frac{\| G[i,:] \|_1}{\sum_{j=1}^{m} \| G[j,:] \|_1}}, \text{ for } i = 1, \ldots, m$$

C_ℓ：全局剪辑阈值，一个预设常数
G：梯度矩阵
$G[i,:]$：第i个参数的梯度向量
$\| G[i,:] \|_1$：第i个参数梯度向量的1-范数

- 随机稀疏化
  - 随机地将梯度向量的一些元素设置为零
  - 选择一个梯度的**稀疏级别**，计算梯度向量中第k大的绝对值作为**阈值**

$$T_{\tau,\gamma}(x) = \begin{cases} x & \text{if } |x| > \tau \\ \text{sign}(x) \cdot \tau & \text{if } \tau \cdot \gamma \le |x| \le \tau \\ 0 & \text{if } |x| < \tau \cdot \gamma \end{cases}$$

τ：全局梯度阈值
γ：比例因子，进一步区分梯度值

为绝对值较大的梯度分配较大的裁剪阈值，从而使重要的学习信号得到较少裁剪

实验设计

- 数据集：
  - Adult
    - 成人收入数据集，预测任务是确定一个人年薪是否超过50K，数据来源于1994年美国人口普查局的数据库，包含如年龄、工作类别、教育水平、婚姻状况、种族等特征
  - Census
    - 人口普查数据集，包含从1990年美国人口普查中抽取的人口统计和就业信息，具有41个与人口和就业相关的变量，预测任务是个人年收入分类
  - Covertype
    - 植被覆盖类型数据集，该数据集中包含一系列的地理和环境特征，如土壤类型、地形阴影、坡度、高度等，被用来预测每个样本点的森林覆盖类型，共有7种不同的森林覆盖类型标签
  - Intrusion
    - 该数据库包含一组要审计的标准数据，其中包括在军事网络环境中模拟的各种入侵，具备正常记录及4种入侵记录

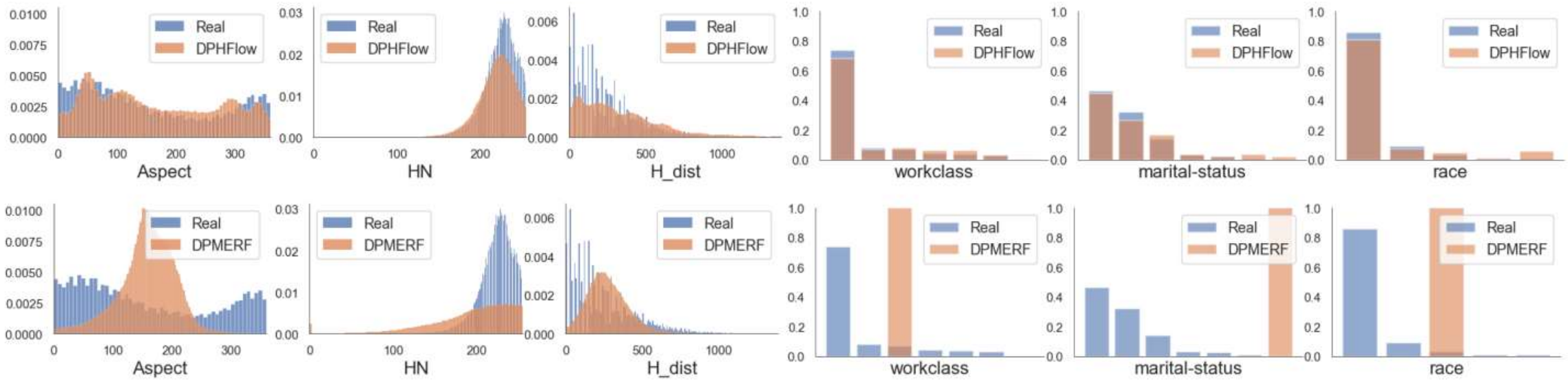
实验设计　对比实验

- 对比方法
  - DP-CGAN（CVPR, 2019）
  - GRU-D（NeurIPS, 2020）
  - DP-MERF（AISTATS, 2021）

- 评价方法
  - 边缘分布对比
    - 评估每个属性列与原始数据列的分布差异
  - 依赖结构比较
    - 评估模型捕获变量之间依赖关系的程度
  - 分类性能评估
    - 通过使用生成的数据来训练分类模型，评估分类模型在真实测试数据上的性能，从而度量生成数据的效用

实验设计　边缘分布对比

- 实验结果
  - DP-HFlow 能够学习具有多个模式的分布，而 DP-MERF 则在平均值上模式崩溃
  - DP-HFlow 能够捕捉到复杂的边缘分布（Covertype、Adult）



实验设计　依赖结构比较

- 实验结果
  - 在 Covertype 数据集上，DP-HFlow 在保留依赖结构方面优于其他方法，具有最小的肯德尔系数矩阵间的均方误差（RMSE）和平均绝对误差（MAE）

	DPCGAN	GSWGAN	DPMERF	**DPHFlow**
RMSE	0.4434	0.2058	0.1863	**0.0717**
MAE	0.3549	0.1396	0.1342	**0.0482**

- Kendall（肯德尔）系数：
  - 同序对（concordant pairs）和异序对（discordant pairs）之差与总对数（n(n-1)/2）的比值定义为Kendall系数

$$\tau = \frac{\text{一致对的数量} - \text{不一致对的数量}}{\frac{n(n-1)}{2}}$$

实验设计　分类性能评估

- 实验结果
  - 使用不同方法生成的合成数据来训练分类器时，DP-HFlow 在 Adult、Census 和 Intrusion 数据集上的性能优于其他基线
  - DP-MERF 在 Covertype 数据集上的表现超过了其他方法，虽然其生成数据的边缘分布与原始数据集有很大的偏差

			Private				Non-Private		
		Real	DP-CGAN	GS-WGAN	DP-MERF	**DP-HFlow**	CTGAN	DP-MERF	**DP-HFlow**
Macro-F1	Adult	0.79±0.02	0.46±0.07	0.42±0.09	0.37±0.15	**0.56±0.07**	0.74±0.02	0.41±0.16	**0.75±0.01**
	Census	0.75±0.01	0.45±0.09	0.44±0.13	0.48±0.14	**0.52±0.03**	0.67±0.04	0.50±0.14	**0.70±0.04**
	Covertype	0.77±0.16	0.15±0.03	0.11±0.03	**0.31±0.05**	0.22±0.03	0.22±0.04	0.29±0.05	**0.49±0.04**
	Intrusion	0.86±0.07	0.19±0.09	0.13±0.08	0.36±0.05	**0.40±0.03**	**0.54±0.05**	0.38±0.06	0.46±0.06

			Private				Non-Private		
		Real	DP-CGAN	GS-WGAN	DP-MERF	**DP-HFlow**	CTGAN	DP-MERF	**DP-HFlow**
AUROC	Adult	0.90±0.02	0.53±0.14	0.48±0.11	0.63±0.09	**0.75±0.05**	0.86 ±0.02	0.65±0.10	**0.87±0.02**
	Census	0.93±0.02	0.50±0.16	0.56±0.20	0.68±0.13	**0.78±0.06**	0.89±0.04	0.69±0.11	**0.91±0.04**
APC	Adult	0.77±0.04	0.28±0.09	0.25±0.05	0.34±0.08	**0.50±0.07**	0.66±0.04	0.39±0.10	**0.70±0.05**
	Census	0.59±0.06	0.07±0.03	0.10±0.06	0.15±0.07	**0.17±0.05**	0.43±0.07	0.15±0.06	**0.49±0.08**

CeFlow: A Robust and Efficient Counterfactual Explanation Framework for Tabular Data Using Normalizing Flows

算法原理 TIPO

T	目标	克服VAE采样过程中的随机性而产生的不稳定结果，为基于表格数据的机器学习模型生成反事实解释
I	输入	Law（16维特征，二分类）、Campus（14维特征，二分类）、Adult（14维特征，二分类）分类数据集
P	处理	1.使用预训练的归一化流模型建模 2.添加扰动向量并生成反事实样本 3.基于优化目标函数反复迭代
O	输出	反事实解释样本

P	问题	基于VAE的方法样本生成缓慢且结果缺乏稳定性
C	条件	反事实解释须保证接近度和稀疏性
D	难点	如何构造目标函数来学习反事实标签的条件分布
L	水平	PAKDD 2023

算法原理　反事实解释

- 反事实解释
  - 一种可解释的机器学习形式，它通过对输入样本进行微小的变化来生成新的样本，新样本在模型的预测结果上与原始样本有所不同
  - 这种方法的目的是为了解释和理解模型的决策过程
  - 反事实解释通过提供“如果...会怎样”（What if）的情景分析，帮助用户理解需要如何改变输入特征才能得到不同的预测结果
- 反事实举例
  - 假设小徐想要申请一项奖学金，奖学金的评审系统基于机器学习模型，考虑了GPA、项目论文等多个因素，小徐提交了申请后系统做出不予发放奖学金的决定
  - 反事实解释就是指，如果小徐希望改变这一决定结果，需要做出哪些具体改变：例如系统可能会告诉他：如果发表一篇SCI 1区，将有资格获得这项奖学金

算法原理　反事实解释

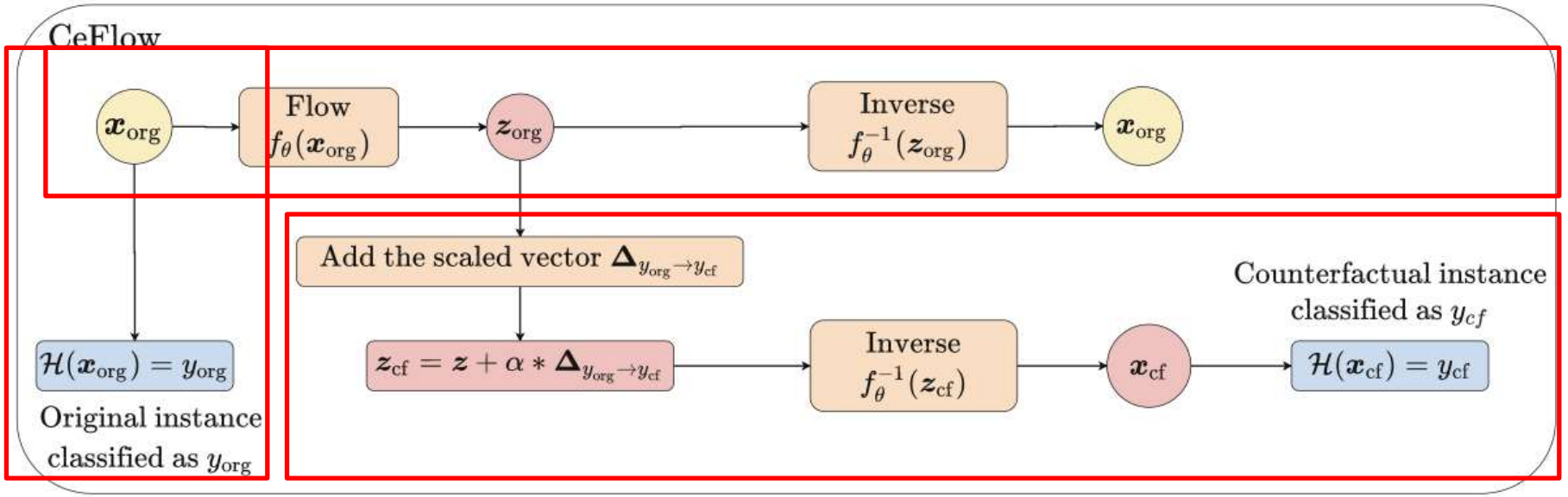
- 反事实解释
  - 准确性（Accuracy）
    - 反事实能够改变模型的预测结果至目标类别，模型的输出与期望的目标结果一致
  - 可行性（Feasibility）
    - 生成的反事实不应该违反任何已知的数据约束或现实世界的逻辑（如年龄不为负）
  - 接近度（Proximity）
    - 反事实与原始实例在特征空间中的距离尽可能小，相较于原始实例，应该只对少数特征进行微小的修改，以保持可信性
  - 稀疏性（Sparsity）
    - 优良的反事实解释应该只涉及少数几个特征的改变，而不是对特征进行大规模调整
  - 多样性（Diversity）
    - 对于同一个实例，可以生成多个有效的反事实解释，以提供用户不同的改变选项

算法原理　反事实解释

- VAE的局限性
  - 不稳定的潜在表示
    - 从编码器模型中采样的潜在表示会随着不同的采样次数而改变，导致不稳定的反事实样本，在重复实验中无法保持一致性
  - 缓慢的扰动转换过程
    - VAE的优化过程涉及对潜在空间进行采样，并对采样得到的潜在向量添加随机扰动，这个过程需要多次迭代才能达到理想的输出，难以在实时响应的场景中应用
  - 反事实样本与密度区域脱节
    - VAE生成的反事实样本可能不会落在接近决策边界的高密度区域，使得生成的解释不可操作，不能指导用户进行实际的改变

为黑箱模型提供反事实样本能够促进人机交互，从而促进ML模型在多个领域的应用

算法原理　算法原理图



- 构建原始分类器
  - 实现从原始特征到标签的映射
- 构建条件流高斯混合模型
  - 将分类特征转换为连续表示后，通过归一化流建模原始样本的分布
- 反事实样本生成
  - 在潜在空间中加入缩放后的扰动向量改变样本分类，通过逆映射生成反事实样本

算法原理　目标函数

- 反事实解释的优化目标
  - 反事实解释的目的是找到最接近的反事实样本x_{cf}，从而将x_{cf}的分类器的结果更改为期望的输出类y_{cf}

$$x_{cf} = \arg\min_{x_{cf} \in \mathcal{X}} d(x_{cf}, x_{org}) \qquad \mathcal{H}(x_{cf}) = y_{cf}$$

  - 基于 $x_{cf} = f_\theta^{-1}(z_{org} + \delta_z)$ 可将目标函数改写为：

$$\begin{cases} \delta_z & = \arg\min_{\delta_z, z \in Z} d(z_{org} + \delta_z, z_{org}) \\ \mathcal{H}(x_{cf}) & = y_{cf} \end{cases}$$

$\mathcal{H}: \mathcal{X} \to \mathcal{Y}$：分类器，输入特征$\mathcal{X}$，输出分类$\mathcal{Y} = \{1 \ldots C\}$
f_θ^{-1}：归一化流模型的逆函数，参数θ
x_{org}：原始样本
δ_z：反事实样本扰动

反事实解释应考虑原始样本特征的最小改变，以使解释尽可能地可行

算法原理　条件流高斯混合模型

- 条件流高斯混合模型构建
  - 使用高斯去量化将分类特征转换为连续表示
  - 分类特征 x^{cat} 高斯去量化转换为连续表示，然后与x^{con}合并得到 x^{full}

$$p_z(z^{full}|y=k)=\mathcal{N}(z^{full}|\mu_k,\Sigma_k)$$

  - 变换后随机变量 $x^{full}=f_\theta^{-1}(z^{full})$ 的密度为：

$$p_x(x^{full})=\log(p_z(f_\theta(x^{full})))+\log\left(\left|\det\left(\frac{\partial f_\theta}{\partial x^{full}}\right)\right|\right)$$

$$p_x(x^{full}|y=k)=\log(\mathcal{N}(f_\theta(x^{full})|\mu_k,\Sigma_k))+\log\left(\left|\det\left(\frac{\partial f_\theta}{\partial x^{full}}\right)\right|\right)$$

$\mathcal{N}$：高斯分布，均值为 μ_k ，协方差为 Σ_k

  - 最大化连续/离散分类特征联合似然

$$\theta^*,\theta_{cat}^*,\theta_{con}^*=\arg\max_{\theta,\theta_{cat},\theta_{con}}\left(\prod_{n=1}^{N}p_x\left(x_n^{con}|x_n^{con}\right)p_x(x_n^{cat}|x_n^{cat})\right)$$

$$=\arg\max_{\theta,\theta_{cat},\theta_{con}}\sum_{n=1}^{N}\left(\log(\mathcal{N}(f_\theta(x_n^{full})|\mu_k,\Sigma_k))+\log\left(\left|\det\left(\frac{\partial f_\theta}{\partial x_n^{full}}\right)\right|\right)\right)$$

算法原理 反事实样本生成

- 反事实样本生成
  - 训练可逆函数
    - 最大化对数似然
  - 计算平均潜在向量

$$G_k = \{(x_m, y_m)\}_{m=1}^{M}$$ 相同预测类别

$$\mu_k = \frac{1}{M} \sum_{x_m \in G_k} f_\theta\,(x_m)$$

  - 计算扰动向量

$$\Delta_{y_{org} \to y_{cf}} = \left| \mu_{y_{org}} - \mu_{y_{cf}} \right|$$

  - 生成反事实样本

$$x_{cf} = f_\theta^{-1}\left(f_\theta(x_{org}) + \alpha \Delta_{y_{org} \to y_{cf}}\right)$$

Algorithm 1. Counterfactual explanation flow (CeFlow)

Input: An original sample $\boldsymbol{x}_{\text{org}}$ with its prediction y_{org}, desired class y_{cf}, a provided machine learning classifier $\mathcal{H}$ and encoder model Q_ϕ.

1: Train the invertible function f_θ by maximizing the log-likelihood:

$$\theta^*, \phi^*_{\text{cat}}, \theta^*_{\text{cat}} = \arg\max_{\theta, \phi_{\text{cat}}, \theta_{\text{cat}}} \prod_{n=1}^{N} \left(\prod_{x_n^{\text{con}} \in \mathcal{X}^{\text{con}}} p_{\mathcal{X}}\left(\boldsymbol{x}_n^{\text{con}}\right) \prod_{x_n^{\text{cat}} \in \mathcal{X}^{\text{cat}}} p_{\mathcal{X}}\left(\boldsymbol{x}_n^{\text{cat}}\right) \right)$$

$$= \arg\max_{\theta, \phi_{\text{cat}}, \theta_{\text{cat}}} \prod_{n=1}^{N} \left(\log\left(\mathcal{N}\left(f_\theta(\boldsymbol{x}_n^{\text{full}}) \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right) \right) + \log\left(\left| \det\left(\frac{\partial f_\theta}{\partial \boldsymbol{x}_n^{\text{full}}} \right) \right| \right) \right)$$

2: Compute mean latent vector $\boldsymbol{\mu}_k$ for each class k by $\boldsymbol{\mu}_k = \frac{1}{M} \sum_{\boldsymbol{x}_m \in \mathcal{G}_k} f(\boldsymbol{x}_m)$.

3: Compute the scaled vector $\boldsymbol{\Delta}_{y_{\text{org}} \to y_{\text{cf}}} = \left| \boldsymbol{\mu}_{y_{\text{org}}} - \boldsymbol{\mu}_{y_{\text{cf}}} \right|$.

4: Find the optimal hyperparameter α by searching a range of values.

5: Compute $\boldsymbol{x}_{\text{cf}} = f_\theta^{-1}(f_\theta(\boldsymbol{x}_{\text{org}}) + \alpha \boldsymbol{\Delta}_{y_{\text{org}} \to y_{\text{cf}}})$.

Output: $\boldsymbol{x}_{\text{cf}}$.

实验设计

- 数据集
  - Law
    - 法学院招生委员会LSAC的数据集，包含学生的入学考试成绩等信息，共39维特征，23维离散特征，16维连续特征，预测目标是律师考试是否通过
  - Compas
    - 由2013年至2014年在佛罗里达州Broward县接受COMPAS筛查的所有刑事被告组成，用于研究刑事量刑中的机器偏见，包含6167名被告的14维特征信息，包括种族，先前犯罪记录等相关属性，预测目标是被告是否会再次犯罪
  - Adult
    - 成人收入数据集，来源于UCI机器学习库，包含48842条记录共14维特征，如年龄、工作类别、教育水平、婚姻状况等，预测目标是个体收入是否超过50K美元/年

实验设计　对比实验

- 对比方法
  - Actionable Recourse (AR)（ACM FAT， 2019）线性模型
  - Growing Sphere (GS)（Arxiv，2017）球状搜索
  - FACE（AAAI，2020）真实数据
  - CERTIFAI（AAAI，2020）公平性、鲁棒性和解释性
  - DiCE（ACM FAT， 2020）专攻多样性
  - C-CHVAE（WWW，2020）VAE，特定约束条件
- 评价指标
  - success rate：衡量生成反事实的成功次数比例
  - L1-norm：评估所需改变的最小幅度
  - mean log-density：评估反事实的可信度或出现的可能性

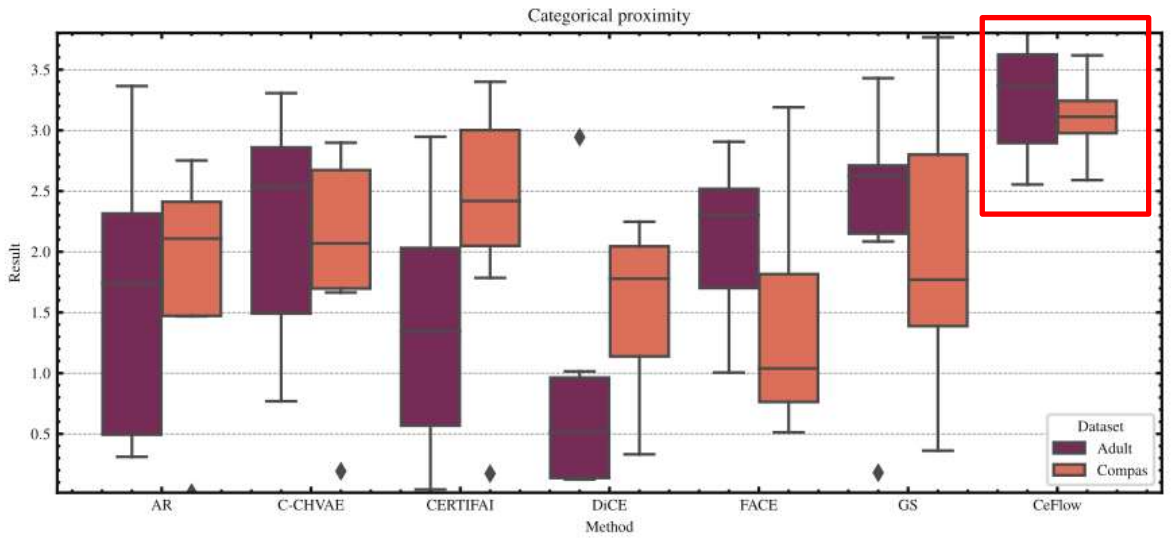
实验设计　对比实验

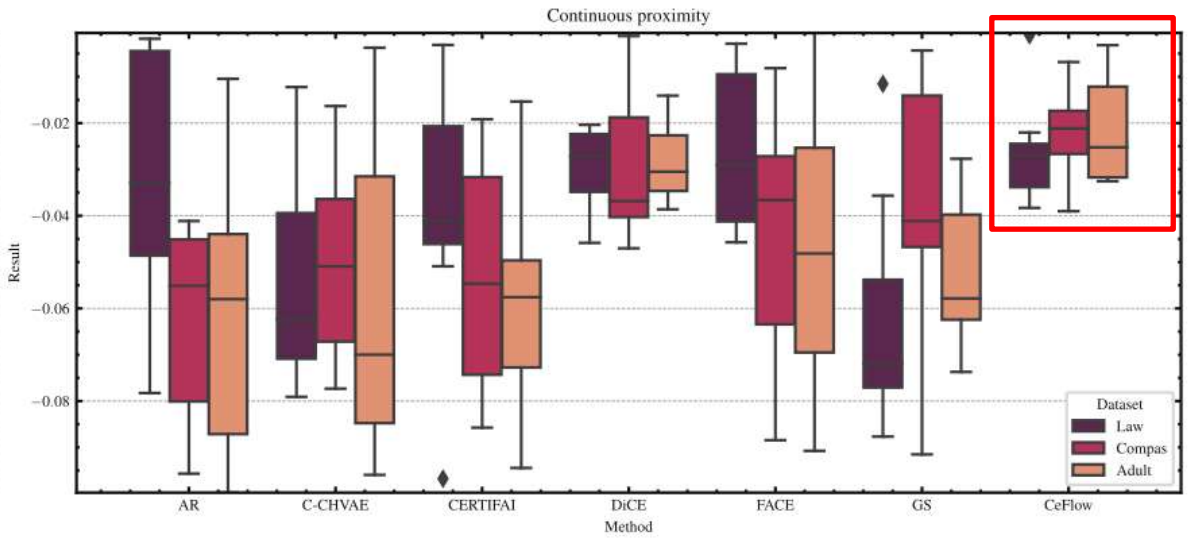
- 实验结果
  - success rate：CeFlow在所有数据集上都实现了100%的成功率
  - L1-norm：显示出算法在生成稳定且接近原始实例的反事实方面的鲁棒性
  - mean log-density：表明生成的反事实在数据分布中的可能性较高，与真实数据分布更为一致

Dataset	Method	Performance				p-value		
		success	l_1-mean	l_1-var	log-density	success	l_1	log-density
Law	AR	98.00	3.518	2.0e-03	-0.730	0.041	0.020	0.022
	GS	100.00	3.600	2.6e-03	-0.716	0.025	0.048	0.016
	FACE	100.00	3.435	2.0e-03	-0.701	0.029	0.010	0.017
	CERTIFAI	100.00	3.541	2.0e-03	-0.689	0.029	0.017	0.036
	DiCE	94.00	**3.111**	2.0e-03	-0.721	0.018	0.035	0.048
	C-CHVAE	100.00	3.461	1.0e-03	-0.730	0.040	0.037	0.016
	CeFlow	**100.00**	3.228	**1.0e-05**	**-0.679**	-	-	-
Compas	AR	97.50	1.799	2.4e-03	-14.92	0.038	0.034	0.046
	GS	100.00	1.914	3.2e-03	-14.87	0.019	0.043	0.040
	FACE	98.50	1.800	4.8e-03	-15.59	0.036	0.024	0.035
	CERTIFAI	100.00	1.811	2.4e-03	-15.65	0.040	0.048	0.038
	DiCE	95.50	1.853	2.9e-03	-14.68	0.030	0.029	0.018
	C-CHVAE	100.00	1.878	1.1e-03	-13.97	0.026	0.015	0.027
	CeFlow	**100.00**	**1.787**	**1.8e-05**	**-13.62**	-	-	-
Adult	AR	100.00	3.101	7.8e-03	-25.68	0.044	0.037	0.018
	GS	100.00	3.021	2.4e-03	-26.55	0.026	0.049	0.028
	FACE	100.00	2.991	6.6e-03	-23.57	0.027	0.015	0.028
	CERTIFAI	93.00	3.001	4.1e-03	-25.55	0.028	0.022	0.016
	DiCE	96.00	2.999	9.1e-03	-24.33	0.046	0.045	0.045
	C-CHVAE	100.00	3.001	8.7e-03	-24.45	0.026	0.043	0.019
	CeFlow	**100.00**	**2.964**	**1.5e-05**	**-23.46**	-	-	-

实验设计　对比实验

- 实验结果
  - 分类近似性：CeFlow能够保持生成反事实的分类特征与原始数据接近
  - 连续近似性：表现仅次于DiCE

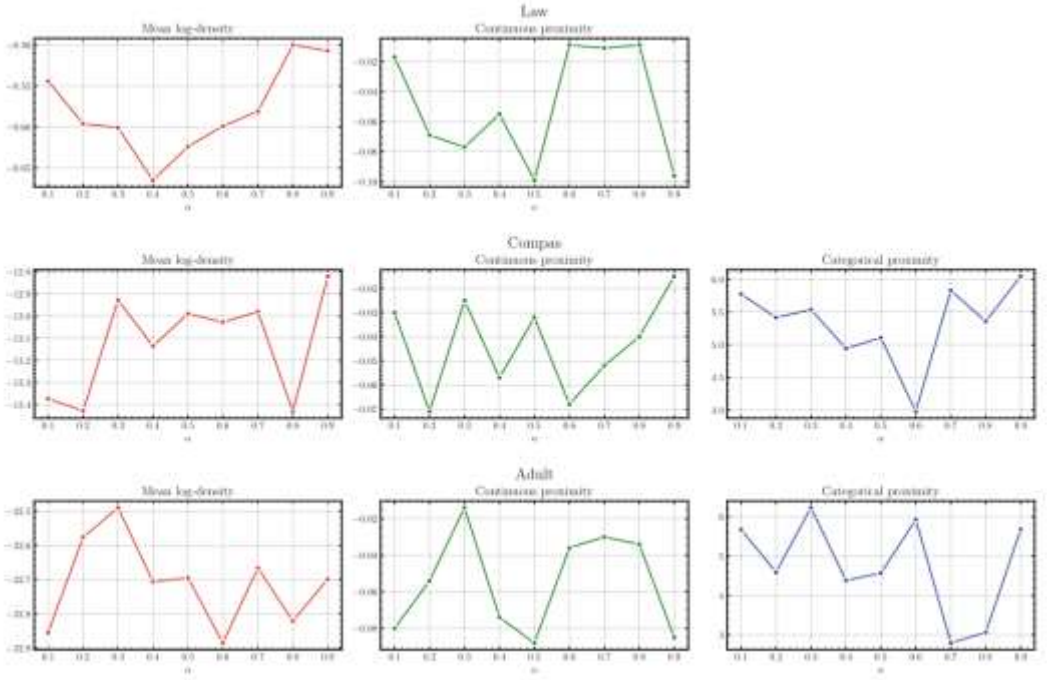




反事实解释：接近度（Proximity）

实验设计　对比实验

- 超参数实验
  - Law、Compas、Adult的超参数分别对应于0.8、0.9、0.3
  - 使用平均对数密度、分类近似性、连续近似性进行评估
- 运行时间对比
  - CeFLow极大加速了扰动转换过程
  - 加快了反事实样本的生成速度



Dataset	AR	GS	FACE	CERTIFAI	DiCE	C-CHVAE	CeFlow
Law	3.030 ± 0.105	7.126 ± 0.153	6.213 ± 0.007	6.522 ± 0.088	8.022 ± 0.014	9.022 ± 0.066	**0.850 ± 0.055**
Compas	5.125 ± 0.097	8.048 ± 0.176	7.688 ± 0.131	13.426 ± 0.158	7.810 ± 0.076	6.879 ± 0.044	**0.809 ± 0.162**
Adult	7.046 ± 0.151	6.472 ± 0.021	13.851 ± 0.001	7.943 ± 0.046	11.821 ± 0.162	12.132 ± 0.024	**0.837 ± 0.026**

特点总结与未来展望

特点总结

- 归一化流
  - 由一系列可逆变换组成，可构建从简单到复杂分布的精确映射，也能反向进行
  - 通过变换雅可比行列式，可直接计算复杂数据的概率密度，不需要近似
- DP-Hflow
  - 使用变分去量化将离散的分类变量转换为连续变量
  - 利用自回归样条变换建模变量的复杂分布
- CeFlow
  - 利用归一化流优化了扰动转换过程
  - 增强了反事实解释生成的稳定性
- 工作展望
  - 设计更精准的梯度裁剪标准
  - 设计用于机器学习中反事实公平性的归一化流架构

预期收获　回顾分析

- 预期收获
  - 1. 理解归一化流模型的基本概念
    - 归一化流与其他生成模型的区别
    - 归一化流分布转换推导
  - 2. 理解归一化流模型在表格数据生成相关任务中的基本应用
    - 差分隐私下的数据生成，对概率密度进行准确建模
    - 反事实解释生成速率优化，确保稳定生成
  - 3. 了解归一化流模型的前沿发展
    - 异常检测
    - 数据去噪
    - 语音合成与转换
    - 图像风格迁移

参考文献

[1] Lee J, Kim M, Jeong Y, et al. Differentially private normalizing flows for synthetic tabular data generation[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022, 36(7): 7345-7353.

[2] Duong T D, Li Q, Xu G. CeFlow: A Robust and Efficient Counterfactual Explanation Framework for Tabular Data Using Normalizing Flows[C]. Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer Nature Switzerland, 2023: 133-144.

[3] Müller T, McWilliams B, Rousselle F, et al. Neural importance sampling[J]. ACM Transactions on Graphics (ToG), 2019, 38(5): 1-19.

[4] https://blog.csdn.net/m0_56942491/article/details/136345430（归一化流）

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！