Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

表格数据生成：GAN的演进与未来

硕士研究生 徐泽豪

2023年08月13日

问题回溯

- 总结反思
  - 语速过快，停顿明显
  - 知识点讲解不够深入，浮于表面

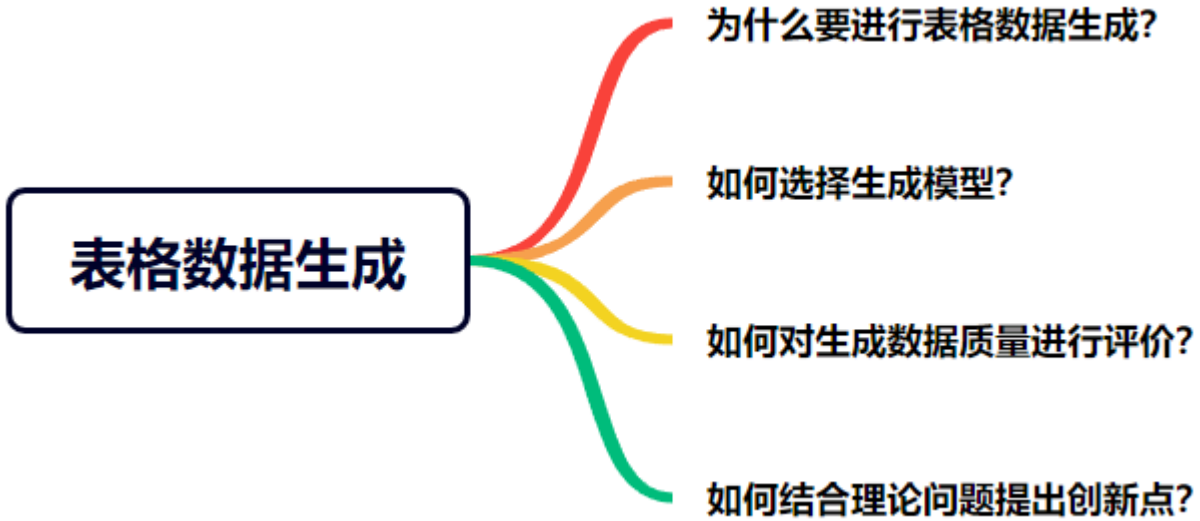
- 相关内容
  - 李班《基于GAN的表格数据生成》——2020.10.11
  - 崔成钢《表格数据的隐私保护方法》——2022.05.22
  - 万韵伟《生成扩散模型》——2023.04.16

预期收获

- 预期收获
  - 1.了解表格数据生成的目的及基本模型
  - 2.理解CTGAN的算法原理及创新思路
  - 3.理解ICDFGAN的算法原理及创新思路
  - 4.了解表格数据生成的前沿发展

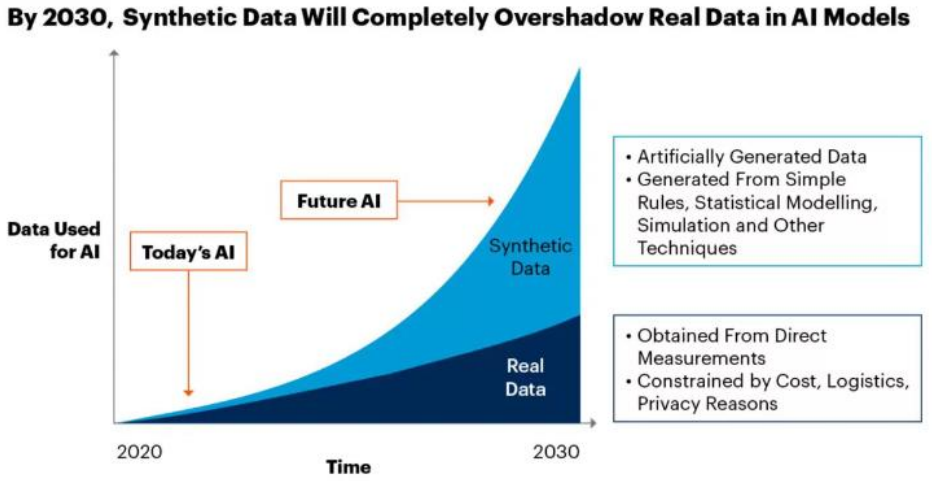
内容提要

- 背景简介
- 基础概念
  - 表格数据
  - 生成模型
  - 问题分析
  - 评价指标
- 算法原理
  - CTGAN
  - ICDFGAN
- 最新进展
- 参考文献
- 附录



研究背景与意义

- 深度学习模型的快速发展
  - 深度学习算法在部分领域逐渐成熟
  - 模型性能的提升主要依靠于数据量和数据质量
  - 合成数据可扩充不平衡的数据集，增强多样性
- 中小型企业的发展需求
  - 目前用户交互数据被大型公司垄断，数据资源匮乏
  - 数据采集、清洗和标注代价高昂，使用合成数据可降低成本
  - 深度学习应服务于各行各业
- 数据隐私与安全
  - 由于法规和隐私问题，无法直接使用真实的用户数据，需要进行数据脱敏



数据生成对深度学习在各行各业落地应用具有重要意义！

基本概念 表格数据

- 表格数据
  - 由行和列组成的结构化数据格式
  - 通常每一行代表一个数据记录
  - 每一列代表该记录的一个属性或特征
- 表格数据生成
  - 学习和模拟整个数据集的整体分布
  - 生成与实际数据集统计特性相似的新数据

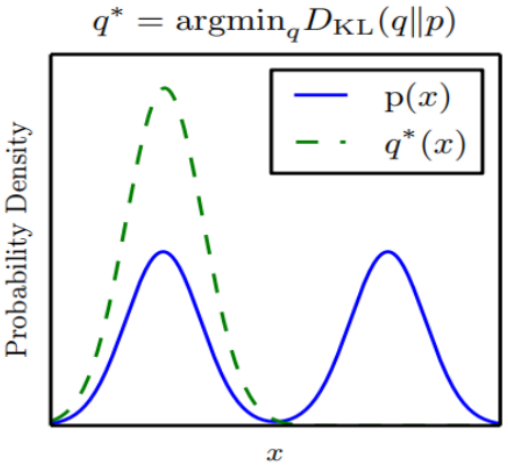
年龄	工作类别	职业	学历
39	State-gov	Adm-clerical	Bachelors
50	Self-emp-not-inc	Exec-managerial	Bachelors
38	Private	Handlers-cleaners	HS-grad
53	Private	Handlers-cleaners	11th
28	Private	Prof-specialty	Bachelors

- 生成模型与判别模型的区别
  - 生成模型关注数据的整体分布，捕获数据内在的结构和变化，考虑所有特征间的关系
  - 判别模型关注对如何对输入数据进行分类/回归，学习输入与输出之间的映射关系，不涉及对数据分布的建模

生成模型关注如何模拟数据的统计特性，判别模型关注如何使用数据来进行准确预测！

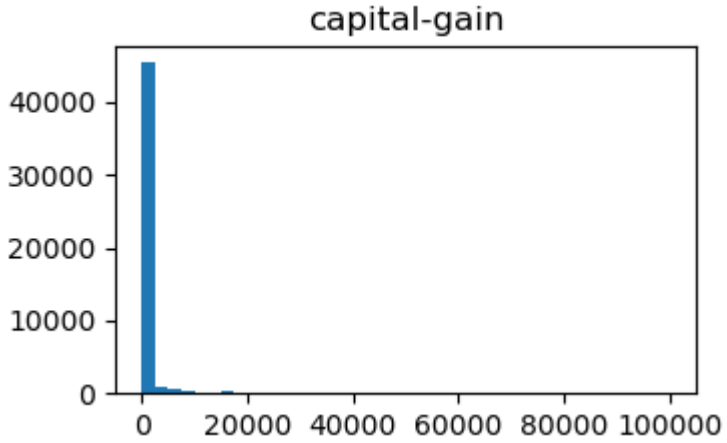
- 混合数据类型
  - 现实世界中表格数据通常由混合的离散和连续数据类型组成，而同时对离散列和连续列进行联合建模较为困难，可能无法捕捉到单个特征列的边缘分布

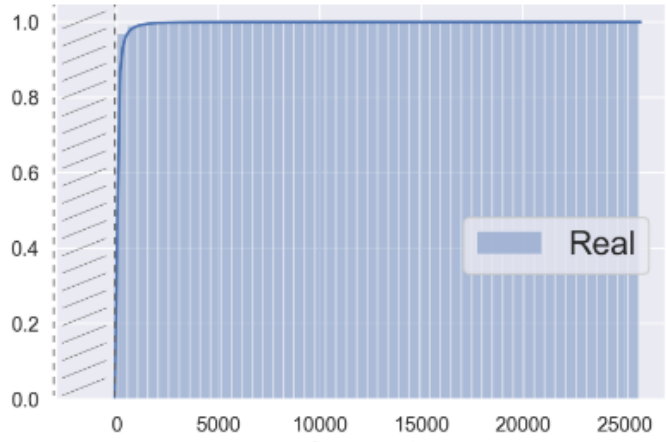
- 多峰分布（Multimode Distribution）
  - 当特征列分布峰较多时，缺少某一分布对数据列整体分布造成的影响较弱，鉴别器无法识别导致模式崩溃
  - 生成器只生成了样本空间中的一小部分，生成的样本大量重复类似，忽略了其他样本特征，不具备整个数据分布的多样性



基本概念 问题分析

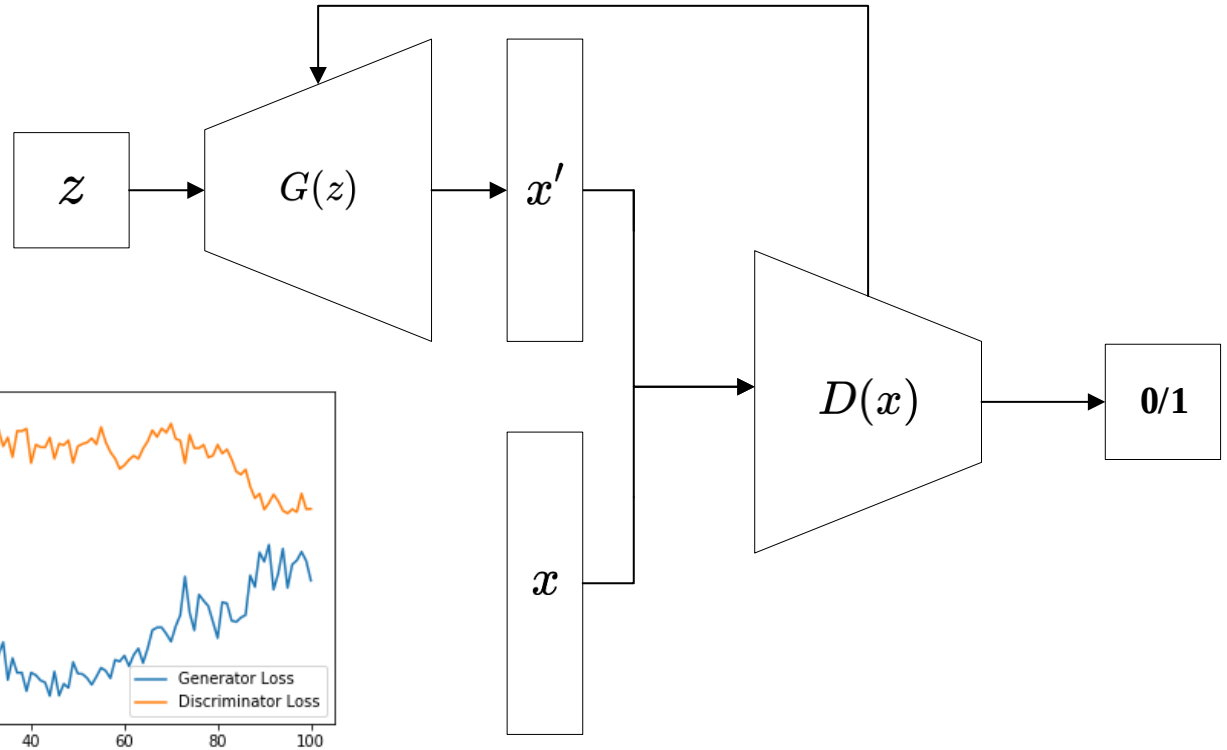
- 非高斯分布
  - 在图像中，像素的值遵循类高斯分布，可以使用Min-Max变换归一化为 [−1, 1]
  - 通常在网络的最后一层使用 tanh 函数来输出此范围内的值
  - 表格数据中的连续值通常是非高斯的，其中Min-Max变换将导致梯度消失问题
- 长尾分布
  - 大多数据点集中在分布的初始值附近，而少数罕见的数据点散布在分布的尾部





基本概念 生成模型

- GAN
  - 包含生成器和判别器两个互相对抗的网络
  - 生成器的任务是生成数据，而判别器的任务是区分生成的数据和真实数据
- 优势
  - 在不断对抗中捕获数据的高级特征、模式和分布
  - 生成高质量、高分辨率的数据
- 劣势
  - 训练可能不稳定
  - 可能出现模式崩溃

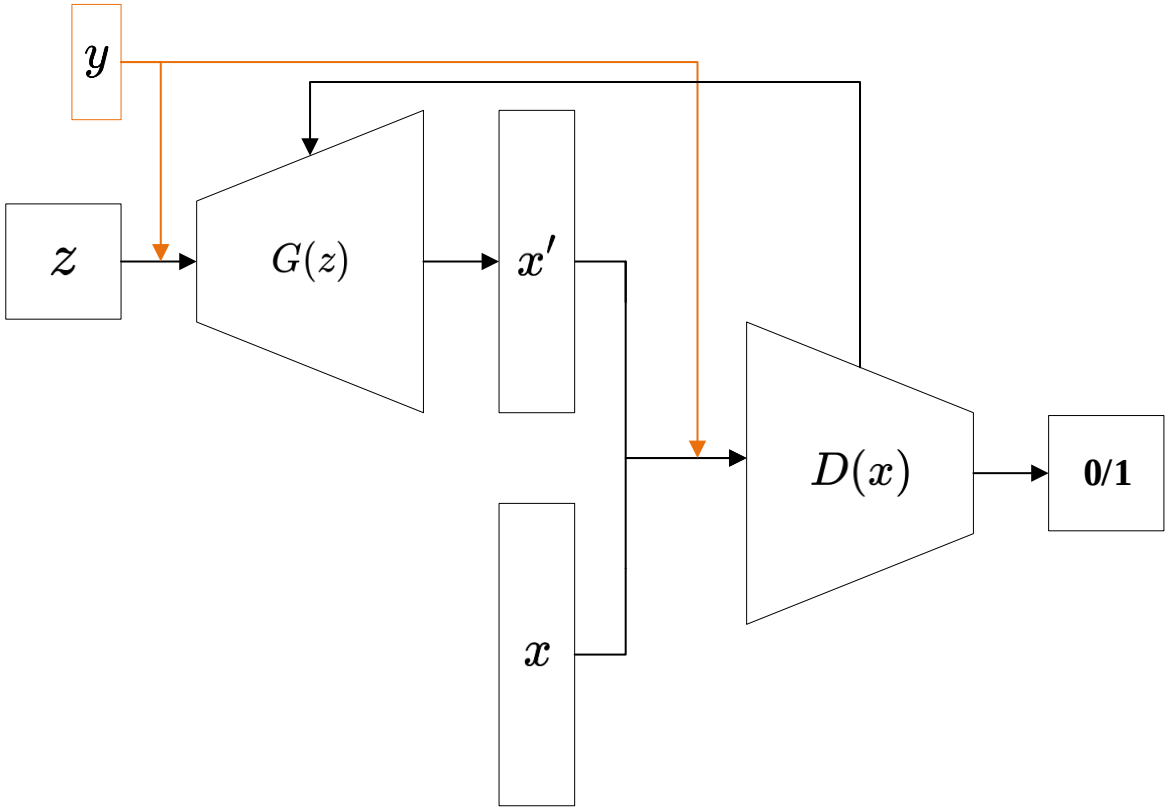


- **Conditional GAN**

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x} \mid \mathbf{y})] + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z})}[\log(1 - D(G(\mathbf{z} \mid \mathbf{y})))]$$

  - 在传统GAN的基础上添加了条件信息
  - 能够通过条件控制生成的内容类型

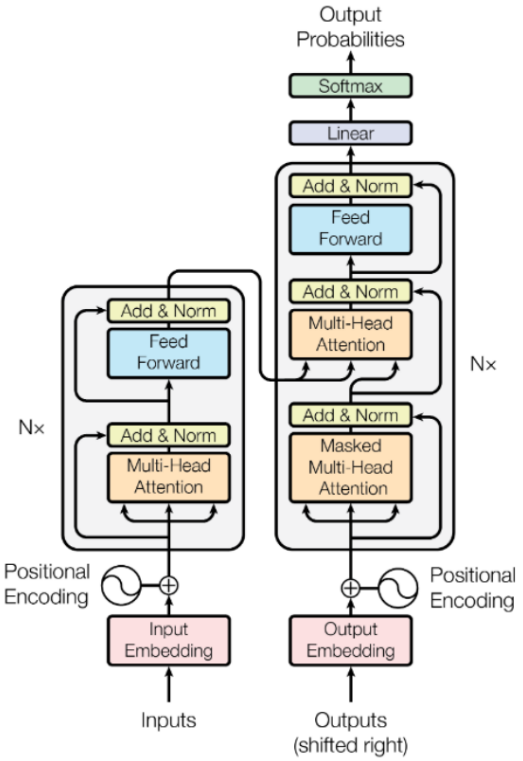


- **应用**
  - 图像领域：根据给定的类别标签生成特定的图像内容

基本概念 生成模型

- Transformer
  - 基于自注意力机制的架构，广泛应用于序列到序列的任务
- 优势
  - 自注意力能够捕获长距离的依赖关系，允许模型在序列中的每个位置上并行计算，从而提高训练和推理的效率
- 劣势
  - 不适用于处理结构化数据
  - 参数量大，需要大量的数据和计算资源
  - 原理图：
    - 李新帅《基于Transformer的时间序列分析》P10——2023.06.17

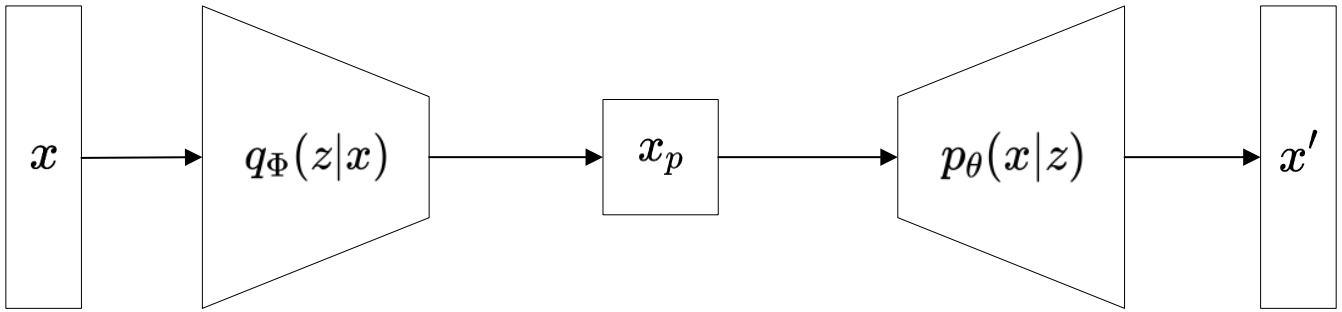
学生	语文	数学
学生A	84	90



Transformer不适用于处理表格数据等结构化数据！

基本概念 生成模型

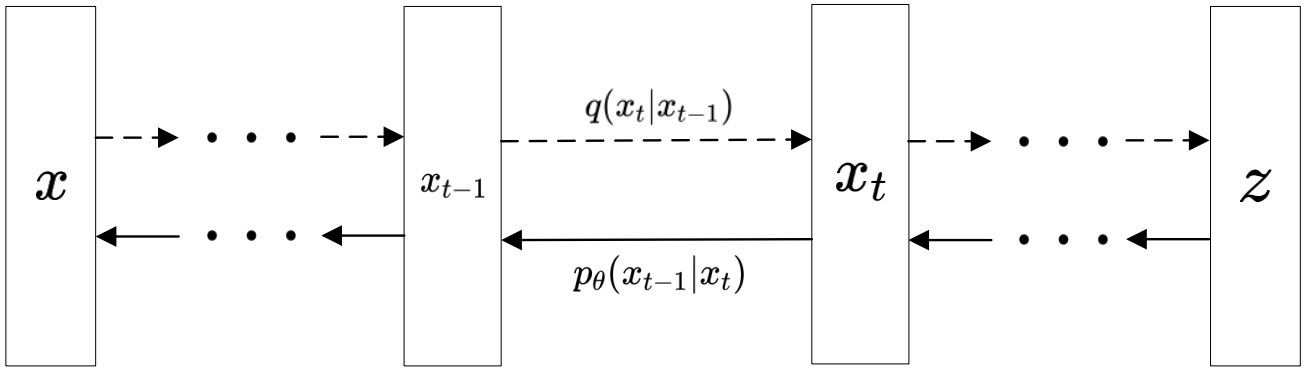
- VAE (Variational Autoencoder)
  - 基于概率的生成模型
  - 编码器将数据编码为潜在空间，而解码器从潜在空间重构数据
- 优势
  - 提供明确的潜在空间表示
  - 训练较为稳定
- 劣势
  - 其潜在空间难以应对离散表格数据
  - 通过同一个潜在编码虽能生成多个不同的样本，增加生成多样性，但同时也导致生成数据模糊，无法保证数据的保真度



VAE无法生成高保真表格数据！

基本概念 生成模型

- **Diffusion Model**
  - 通过一系列反向扩散步骤来逐渐生成数据
- 优势
  - 通过多步骤的过程模拟数据分布，能够有效地模拟复杂的表格数据分布
  - 训练过程更加稳定，是基于确定性的扩散过程，可以控制数据生成的各方面
- 劣势
  - 生成需要多个步骤，生成速度更慢
  - 模型更加复杂



ICDFGAN

【CTGAN】

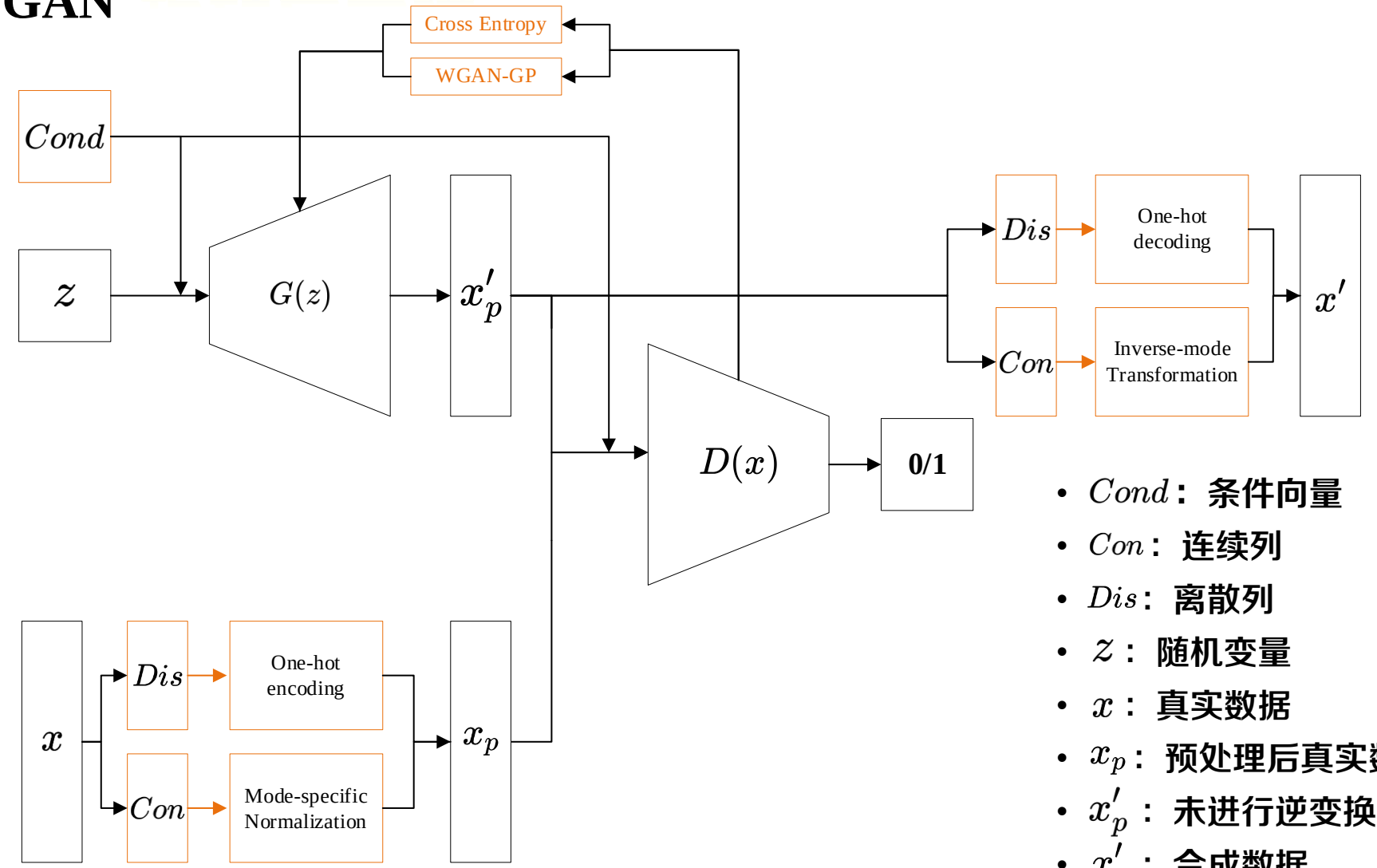
CTGAN TIPO PCDL

T	目标	构建表格数据生成模型
I	输入	真实表格数据
P	处理	1. 将真实数据分为连续、离散两部分 2. 分别进行特定于模式的归一化，one-hot编码预处理 3. 引入条件向量使生成器生成特定数据 4. 结合交叉熵和推土机距离设计损失函数，避免模式崩溃 5. 从训练好的生成模型采样出合成数据
O	输出	与真实数据同分布的合成数据

P	问题	生成和真实数据同分布的合成数据
C	条件	表格类型数据
D	难点	1. 生成具备多模式的连续数据 2. 生成 one-hot 解码后的离散数据 3. GAN的模式崩溃问题
L	水平	NIPS 2019

CTGAN 框架与原理

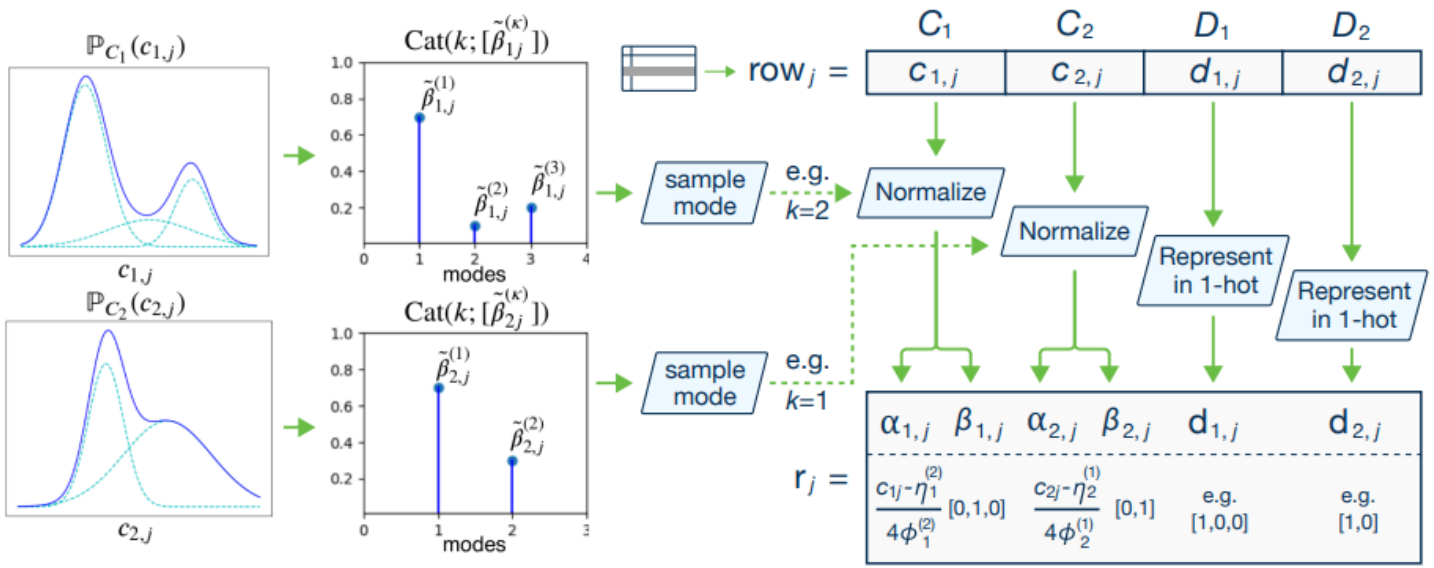
- **CTGAN**



- $Cond$：条件向量
- Con：连续列
- Dis：离散列
- z：随机变量
- x：真实数据
- x_p：预处理后真实数据
- x'_p：未进行逆变换的生成数据
- x'：合成数据

CTGAN 框架与原理

- 特定于模式的归一化
  - 模式估计
    - 使用变分高斯混合模型(VGM)估计C_i的分布中的模式数量
    - 得到一个包含m_i个模式的高斯混合概率模型$\mathbb{P}_{C_i}\left(c_{i,j}\right)$
  - 计算概率质量函数 (PMF)
    - 计算从m_i个模式中采样的值$c_{i,j}$的PMF，确定$c_{i,j}$来自每个模式的归一化概率
  - 样本采样
    - 从上一步的概率分布中采样$k_{i,j}$并将其转换为一个one-hot向量$\beta_{i,j}$
  - 归一化
    - 使用所选择模式的均值和标准偏差对$c_{i,j}$进行归一化，提高模型的训练稳定性



CTGAN 实验分析

- 数据集选择
  - 分类、多分类、图像、回归

method	adult F1	census F1	credit F1	cover. Macro	intru. Macro	mnist12/28 Acc	mnist12/28 Acc	news R^2
Identity	0.669	0.494	0.720	0.652	0.862	0.886	0.916	0.14
CLBN(3)	0.334	0.310	0.409	0.319	0.384	0.741	0.176	-6.28
PrivBN(4)	0.414	0.121	0.185	0.270	0.384	0.117	0.081	-4.49
MedGAN(6)	0.375	0.000	0.000	0.093	0.299	0.091	0.104	-8.80
VEEGAN(6)	0.235	0.094	0.000	0.082	0.261	0.194	0.136	-6.5e6
TableGAN(5)	0.492	0.358	0.182	0.000	0.000	0.100	0.000	-3.09
TVAE(1)	**0.626**	0.377	0.098	**0.433**	0.511	**0.793**	**0.794**	**-0.20**
TGAN(1)	0.601	**0.391**	**0.672**	0.324	**0.528**	0.394	0.371	-0.43

- 机器学习效果
  - 对于adult、census、credit这样的经典结构化表格数据集，其结果反映了模型处理结构化数据的能力，CTGAN更具优势
  - TVAE在类图像分布的数据中更具优势

CTGAN 优劣分析

- 优势
  - 将连续和离散数据分开处理，增强生成数据的保真度
  - 具备条件生成能力，使用户能够指定生成数据的某些属性

- 劣势
  - 对连续和离散数据均使用交叉熵损失，并非连续数据的最佳选择
  - 所选取数据集分布不复杂，转换连续数据时类别数较少，因此对生成结果影响较小
  - 对于复杂数据集，特定于模式的归一化方法可能会极大影响最终结果

【ICDFGAN】

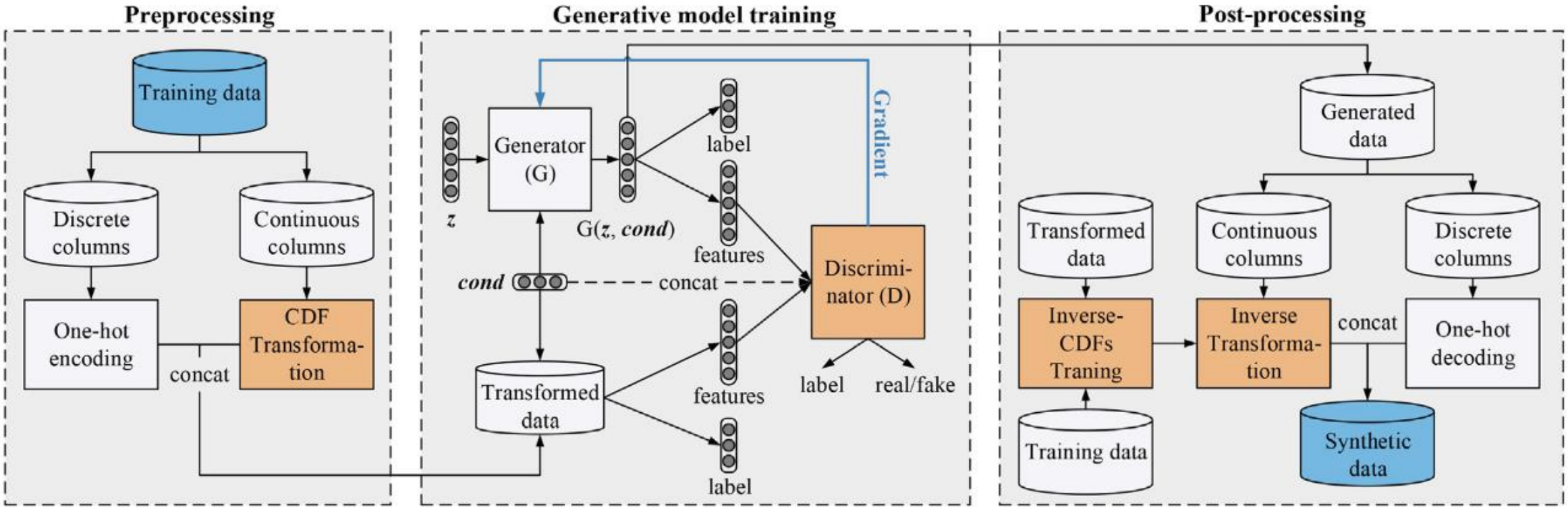
ICDFGAN TIPO PCDL

T	目标	构建表格数据生成模型
I	输入	真实表格数据
P	处理	1. 将真实数据分为连续、离散两部分 2. 分别进行累积分布转换，one-hot编码预处理 3. 引入标签重建向量约束鉴别器和生成器 4. 结合标签重建设计损失函数，使生成器能更好地提取属性间关联 5. 从训练好的生成模型采样出合成数据
O	输出	与真实数据同分布的合成数据

P	问题	边缘分布复杂时，对数据点归一化会产生梯度消失问题
C	条件	表格类型数据
D	难点	1. 引入合适的分布变换方法对训练数据分布进行非线性映射 2. 避免数据变换后GAN难以捕获到真实数据特征间关联性
L	水平	Neurocomputing 2021

ICDFGAN 框架与原理

- **ICDFGAN**
  - 将主体流程划分为预处理、生成器训练、后处理三部分
  - 分开绘制算法流程图，创新点更明显，逻辑更清楚



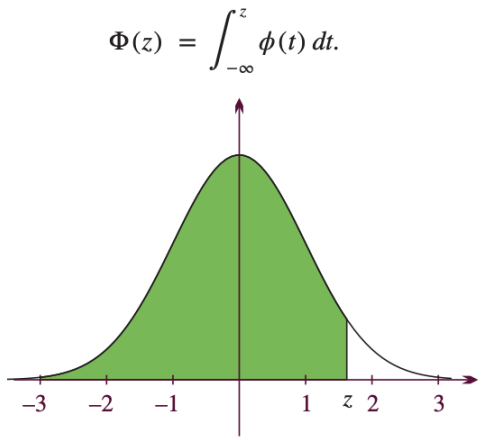
别人的算法最终还是要转化为“自己的”算法！

ICDFGAN 累积分布函数

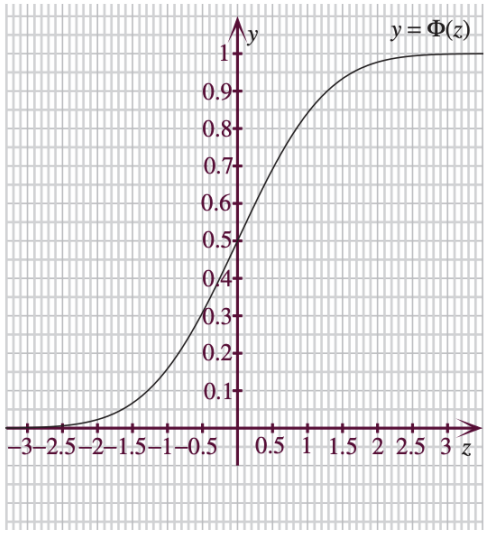
- **CDF (Cumulative Distribution Function)**
  - 累积分布函数描述了一个实数随机变量在特定值以下的概率$F_X(x) = P(X \leq x)$
  - 真实世界中的连续数据分布可能非常复杂，可能包含多模式、长尾等特性
  - 如果直接在复杂的空间中生成数据，GAN可能会产生模式崩溃等问题
  - 将真实的连续数据分布转换为均匀分布，就能通过生成均匀随机数并逆转换将其映射回原始数据的分布

为什么要转换为均匀分布而不用高斯分布?

  - 均匀分布是一个简单且已知的分布，其中所有区域的概率都是相同的



This is the graph of the standard normal probability density function $\phi(z)$.



This is the graph of the standard normal cumulative distribution function $\Phi(z)$.

ICDFGAN 鉴别器损失

- 鉴别器损失
  - WGAN损失：比较鉴别器对真实/生成数据的判断

$$\mathcal{L}_c = \frac{1}{m}\sum_{i=1}^{m} \text{critic}\left(\mathbf{x}_i^{\text{fake}}\right) - \frac{1}{m}\sum_{i=1}^{m} \text{critic}\left(\mathbf{x}_i^{\text{real}}\right)$$

  - WGAN-GP：确保鉴别器的输出在真实数据和生成数据之间是Lipschitz连续的

$$\mathcal{L}_{GP} = \frac{1}{m}\sum_{i=1}^{m}\left(\|\nabla \text{critic}(\tilde{\mathbf{x}}_i)\|_2 - 1\right)^2$$

  - 标签重建损失：确保鉴别器对真实数据标签预测准确

$$\mathcal{L}_{\text{rec}} = \frac{1}{m}\sum_{j=1}^{m} f\left(y_i^{\text{real}}, y_i^{\text{pred}}\right)$$

  - 总损失：

$$\mathcal{L}_D = \mathcal{L}_c + \lambda * \mathcal{L}_{GP} + w_{\text{rec}} * \mathcal{L}_{\text{rec}}$$

1. Initialize the paremeters of G and D;
2. **for** $e = 1 \rightarrow epochs$ **do**
3. $/^* \ i = 1, 2, \cdots, m \ ^*/$
4. Create condition vectors $\boldsymbol{cond}_i$;
5. Sample $\boldsymbol{z}_i \sim \text{MVN}(0, \mathbf{I})$;
6. $\left(\boldsymbol{x}_i^{fake}, \boldsymbol{y}_i^{fake}\right) \leftarrow G(\boldsymbol{z}_i, \boldsymbol{cond}_i)$;
7. Sample $(\boldsymbol{x}_i^{real}, \boldsymbol{y}_i^{real}) \sim \text{Uniform}(\boldsymbol{T}^{trans} | \boldsymbol{cond}_i)$;
8. $\boldsymbol{y}_i^{pred}, critic(\boldsymbol{x}_i^{real}) \leftarrow D(\boldsymbol{x}_i^{real}, \boldsymbol{cond}_{i,x})$; // $\boldsymbol{cond}_{i,x}$ is the features part of $\boldsymbol{cond}_i$.
9. $\_, critic\left(\boldsymbol{x}_i^{fake}\right) = D\left(\boldsymbol{x}_i^{fake}, \boldsymbol{cond}_{i,x}\right)$;
10. $\mathcal{L}_c \leftarrow \frac{1}{m}\sum_{i=1}^{m} critic\left(\boldsymbol{x}_i^{fake}\right) - \frac{1}{m}\sum_{i=1}^{m} critic(\boldsymbol{x}_i^{real})$;
11. Compute $\mathcal{L}_{rec}$ according to formula (9);
12. Sample $\rho_1, \ldots, \rho_m \sim \text{Uniform}(0, 1)$;
13. $\tilde{\boldsymbol{x}}_i \leftarrow \rho_i \boldsymbol{x}_i^{fake} + (1 - \rho_i)\boldsymbol{x}_i^{real}$;
14. $\_, critic(\tilde{\boldsymbol{x}}_i) \leftarrow D(\tilde{\boldsymbol{x}}_i, \boldsymbol{cond}_{i,x})$;
15. $\mathcal{L}_{GP} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(\|\nabla critic(\tilde{\boldsymbol{x}}_i)\|_2 - 1)^2$;
16. $\mathcal{L}_D \leftarrow \mathcal{L}_c + \lambda * \mathcal{L}_{GP} + w_{rec} * \mathcal{L}_{rec}$;

ICDFGAN 生成器损失

- 生成器损失
  - WGAN损失

$$\mathcal{L}_c = -\frac{1}{m}\sum_{i=1}^{m} \text{critic}\left(\mathbf{x}_i^{\text{fake}}\right)$$

  - 标签重建损失
  - 均匀损失：确保生成数据在均值和方差上接近均匀分布

$$\mathcal{L}_{\text{uniform}} = \sum_{j=1}^{N_c}\left[S^2\left(\mathbf{X}_j^{\text{real}}\right) - S^2\left(\mathbf{X}_j^{\text{fake}}\right)\right]^2 + \left[\overline{\mathbf{X}}_j^{\text{real}} - \overline{\mathbf{X}}_j^{\text{fake}}\right]^2$$

  - 条件损失：鼓励生成的离散属性接近给定的条件

$$\mathcal{L}_{\text{cond}} = \frac{1}{m}\sum_{i=1}^{m} \text{CrossEntropy}\left(\hat{\mathbf{d}}_{i,s}, \mathbf{d}_{i,s}^{(k)}\right)$$

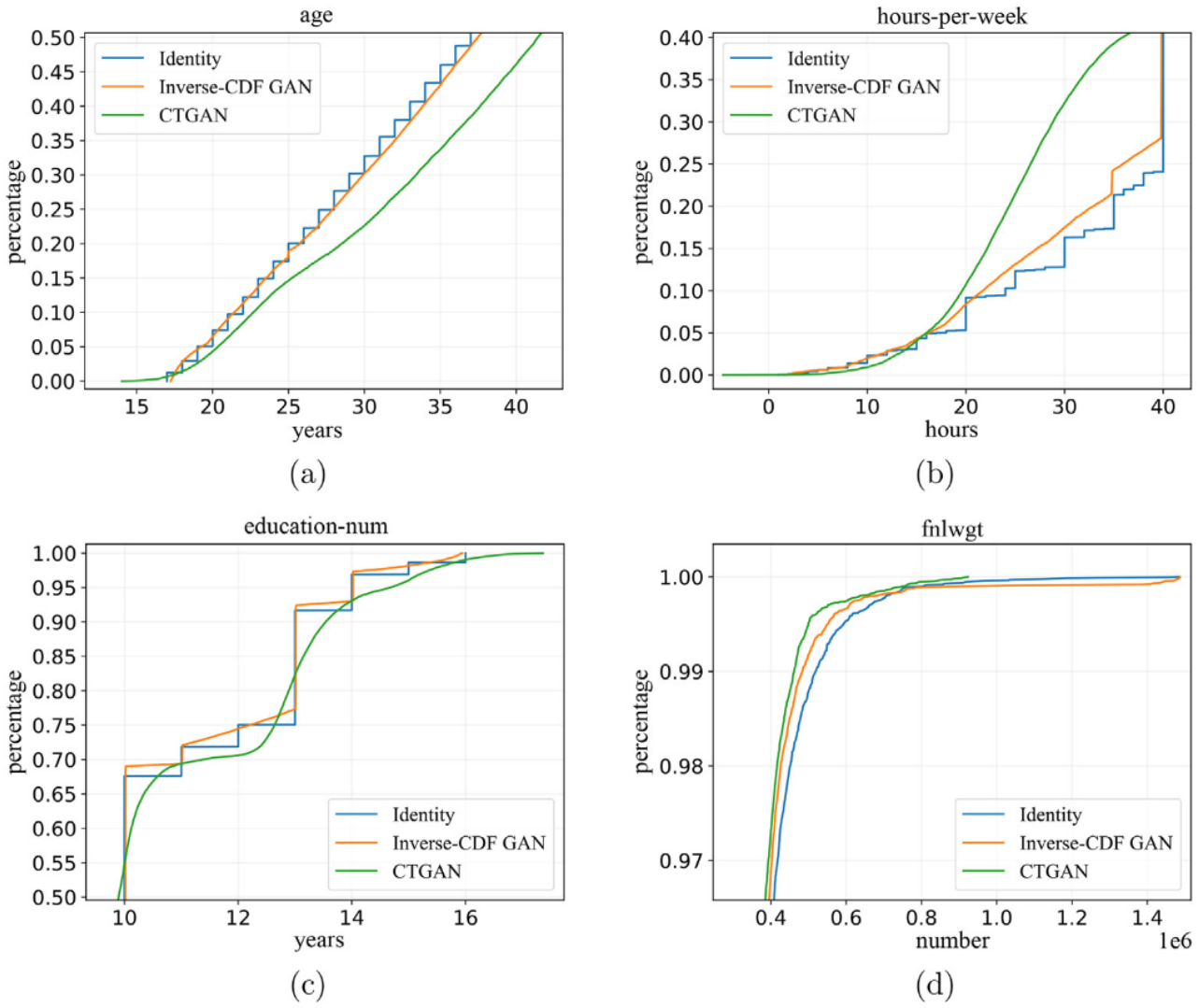
1. Initialize the paremeters of G and D;
2. **for** $e = 1 \rightarrow epochs$ **do**
3. $/^* \ i = 1, 2, \cdots, m \ ^*/$
4. Create condition vectors $\boldsymbol{cond}_i$;
5. Sample $\boldsymbol{z}_i \sim \text{MVN}(0, \mathbf{I})$;
17. Update parameters of D with $\mathscr{L}_D$;
18. Re-execute lines 3 to 5;
19. $y_i^{pred}, critic\left(\boldsymbol{x}_i^{fake}\right) = D\left(\boldsymbol{x}_i^{fake}, \boldsymbol{cond}_{i,x}\right)$;
20. $\mathscr{L}_c \leftarrow -\frac{1}{m}\sum_{i=1}^{m} critic\left(\boldsymbol{x}_i^{fake}\right)$;
21. Compute $\mathscr{L}_{uniform}$ according to formula (4);
22. Compute $\mathscr{L}_{rec}$ between y_i^{pred} and y_i^{fake} according to formula (9);
23. Compute $\mathscr{L}_{cond}$ according to formula (7);
24. $\mathscr{L}_G \leftarrow \mathscr{L}_c + w_{uniform} * \mathscr{L}_{uniform} + w_{rec} * \mathscr{L}_{rec} + w_{cond} * \mathscr{L}_{cond}$;

  - 总损失：

$$\mathcal{L}_G = \mathcal{L}_C + w_{\text{uniform}} * \mathcal{L}_{\text{uniform}} + w_{\text{rec}} * \mathcal{L}_{\text{rec}} + w_{\text{cond}} * \mathcal{L}_{\text{cond}}$$

ICDFGAN 效果分析

- 效果评估
  - 从累积分布函数角度，针对各类分布数据，相对CTGAN能够更好拟合数据分布
  - 粒度不够细，虽然累积函数更加贴近原始数据，但对于部分连续数据的“离散分布”识别效果不佳



ICDFGAN 效果分析

- 机器学习效果

Datasets	adult F1	census F1	credit F1	covtype Macro	intrusion Macro	mnist12 Acc	mnist28 Acc	news R^2
Identity	0.669	0.494	0.720	0.652	0.862	0.886	0.916	0.14
CLBN	0.334	0.310	0.409	0.319	0.384	0.741	0.176	-6.28
PrivBN	0.414	0.121	0.185	0.270	0.384	0.117	0.081	-4.49
MedGAN	0.375	0.000	0.000	0.093	0.299	0.091	0.104	-8.80
VEEGAN	0.235	0.094	0.000	0.082	0.261	0.194	0.136	-6.5e6
TableGAN	0.492	0.358	0.182	0.000	0.000	0.100	0.000	-3.09
TVAE	0.626	0.377	0.098	0.433	0.511	0.793	0.794	**-0.20**
CTGAN	0.601	0.391	**0.672**	0.324	0.528	0.394	0.371	-0.43
OVAE	0.600	0.380	0.510	**0.450**	**0.530**	0.830	0.840	-0.30
Inverse-CDF GAN	**0.635**	**0.434**	0.196	0.437	0.482	**0.832**	**0.861**	-0.22

– 在adult、census结构化数据集以及mnist12/28图像数据集上表现最佳

– 但难以应用于不平衡数据集credit

ICDFGAN 优劣分析

- 优势
  - 使用累积分布函数将原分布转化为均匀分布，能够保证合成数据多样性
  - 使用标签重构函数实现鉴别器，使生成器能更好地提取属性间关联

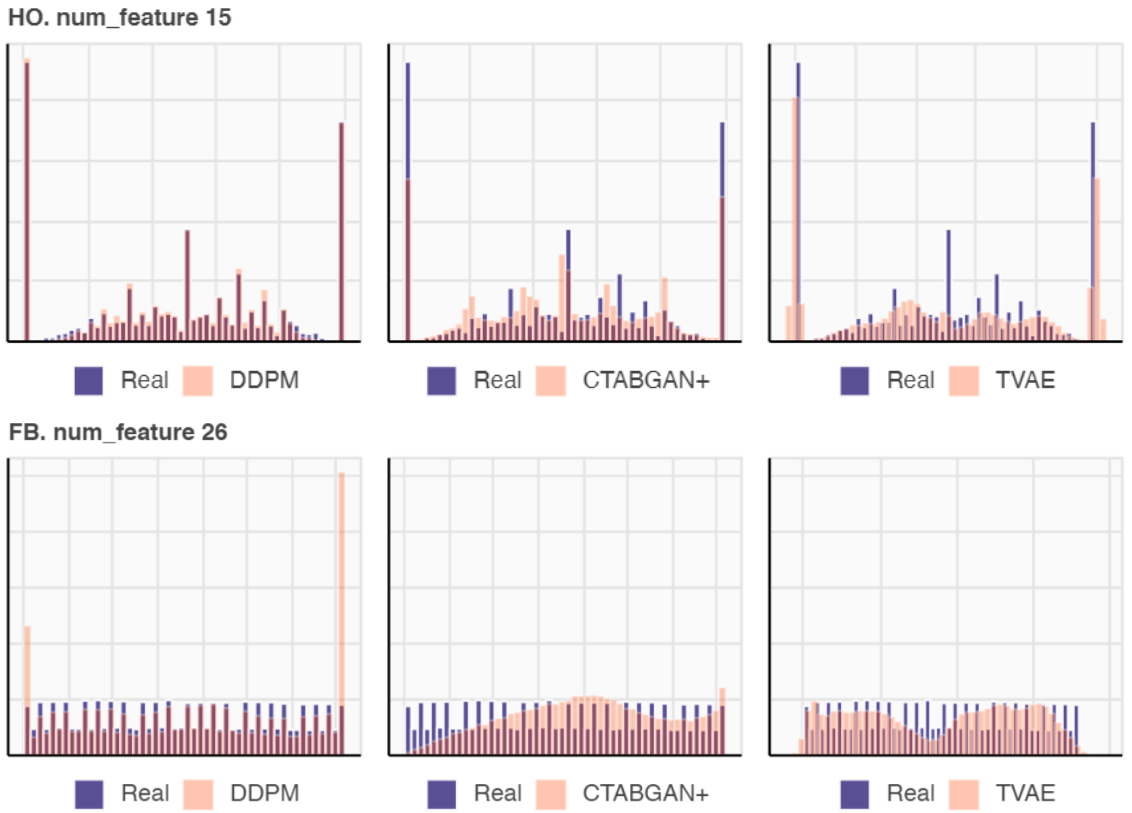
Model	Hyperparameters
Inverse-CDF GAN	Generator network structure: (64, 64), Discriminator network structure: (64, 64), Dimensions of latent vector: 32, Batch size: 500, $w_{uniform} = w_{cond} = w_{rec} = 1$;

- 劣势
  - 对于多个损失函数没有进行消融实验
  - 对于损失函数超参数选择没有进行详细分析，只是将其全设为1
  - 针对不平衡数据集credit效果较差，所提方法对于少数类采样缺乏多样性，不能很好地缓解数据不平衡问题

前沿发展

前沿发展

- **TabDDPM (ICLR 2023)**
  - 扩散模型相对GAN能够更好模拟数据列分布
  - 表格数据生成领域精度愈来愈高
  - 但数据列间存在的潜在关系仍需进一步研究



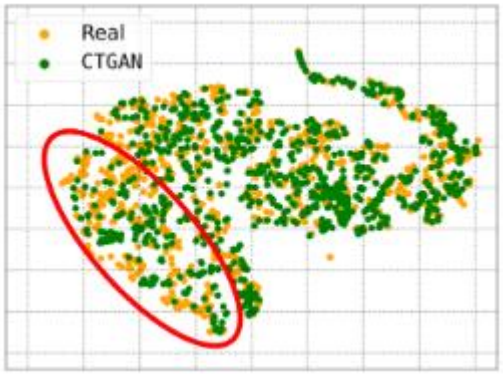
- **特点**
  - 对于离散特征使用多项式扩散
  - 对于分类特征使用高斯扩散
  - 包含多个超参数，需要进行大量的计算找到最优参数

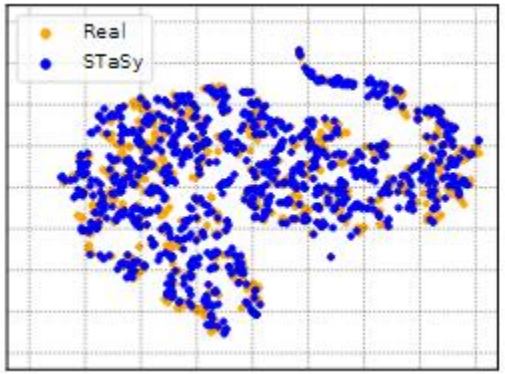
前沿发展

- **STaSy (ICML 2023)**
  - 在采样质量和多样性方面优于现有方法
  - 能够更好捕获真实数据特征

Methods	Quality ↑ (F1 & R^2)	Diversity ↑ (coverage)	Runtime ↓ (second)
VEEGAN	0.006	0.038	0.109
CTGAN	0.569	0.352	0.704
TableGAN	0.501	0.434	**0.046**
OCT-GAN	0.567	0.381	26.926
RNODE	0.490	0.328	13.392
Naïve-STaSy	0.717	0.637	8.855
STaSy	**0.733**	**0.658**	10.663

- **特点**
  - 通过直接匹配数据的分数函数和分数网络，通过数据驱动生成建模
  - 引入自适应学习方法，通过比较损失值和阈值，对数据进行排序，先学习损失值较低的数据





应用总结

应用总结

- **CTGAN**
  - 将连续与离散数据分开处理，具备条件生成能力
  - 使用多个高斯分布拟合连续数据无法应对复杂程度较高的数据集

- **ICDFGAN**
  - 使用累积分布函数和标签重构函数提高了生成质量
  - 在不平衡数据集上表现较差

- **GAN与Diffusion Model**
  - 按生成效果来看，Diffusion Model更胜一筹
  - 按生成效率来看，GAN竞争力更强，可继续挖掘其数据生成潜力

参考文献

[1]Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks[C] International conference on machine learning. PMLR, 2017: 214-223.

[2] Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional gan[J]. Advances in Neural Information Processing Systems, 2019, 32.

[3] Li B, Luo S, Qin X, et al. Improving GAN with inverse cumulative distribution function for tabular data synthesis[J]. Neurocomputing, 2021, 456: 373-383.

[4] Kotelnikov A, Baranchuk D, Rubachev I, et al. Tabddpm: Modelling tabular data with diffusion models[C]. International Conference on Machine Learning. PMLR, 2023: 17564-17579.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。

附录 时序、截面、面板数据

- 时序数据
  - 指在不同时间点上收集到的数据，反映某一事物/现象随时间的变化状态或程度
  - 通常由时间戳和与该时间戳相关联的值组成，按时序顺序记录
  - 例如股票价格、气温和人口数量变化
- 截面数据
  - 指在某一特定时间点或时间段内，多个观测对象的状态或特征
  - 不考虑时间变化，只关注某一特定时间点的情况
  - 例如不同国家在同一年份的 GDP
- 面板数据
  - 指包含多个观测对象，且同一观测对象有一系列不同时间观测点的数据
  - 例如一个公司多年来的财务报表、一个人多年来的健康状况