Diffusion Models for Tabular Data

2024. 10. 18

Data Mining & Quality Analytics Lab.

윤지현

Introduction

발표자 소개

❖ **윤지현 (Jihyun Yun)**

- 고려대학교 산업경영공학과 석사 과정 (2023.09 ~ Present)
- Data Mining & Quality Analytics Lab

❖ **Research Interest**

- Application of Diffusion Models

❖ **Contact**

- whle56@korea.ac.kr

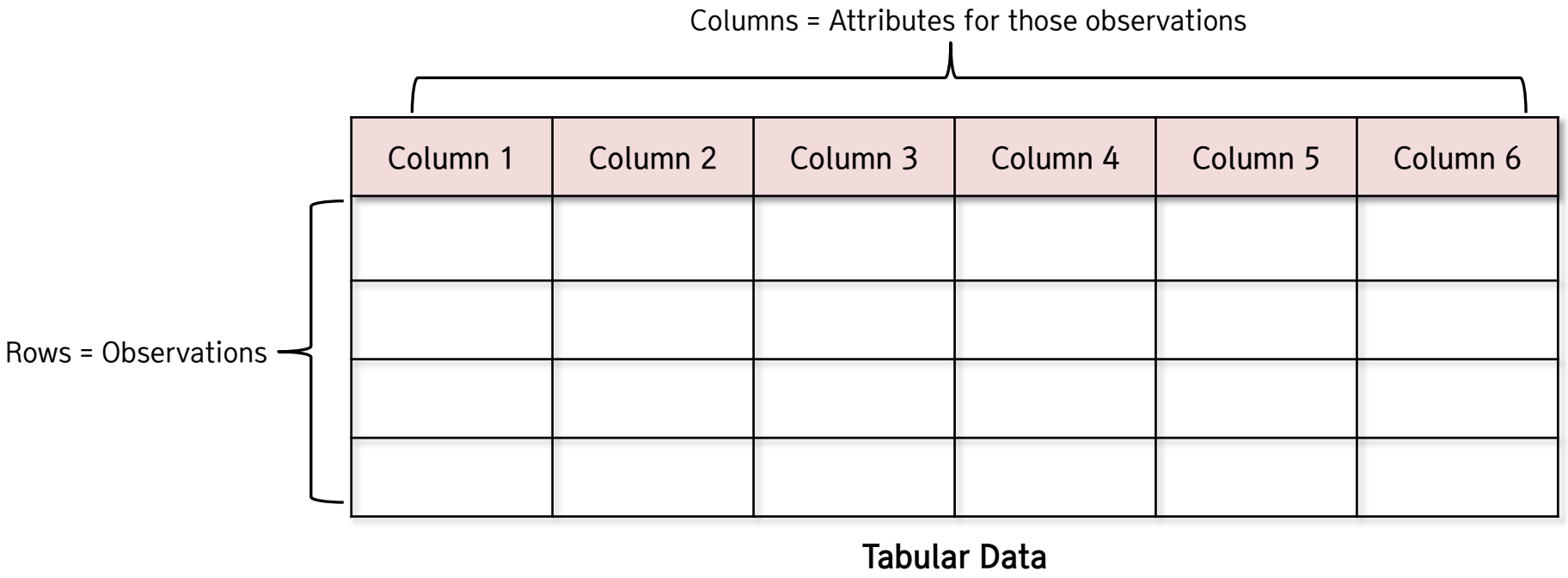
Contents

Introduction

Understanding Tabular Data

❖ Understanding Tabular Data

- 정형 데이터의 분석은 주로 트리 기반 모델들의 앙상블 모델들이 사용됨
- 비정형 데이터에서 높은 성능을 보이는 딥러닝이 정형 데이터에서는 여전히 전통적 통계 기반의 방법론들이 우세



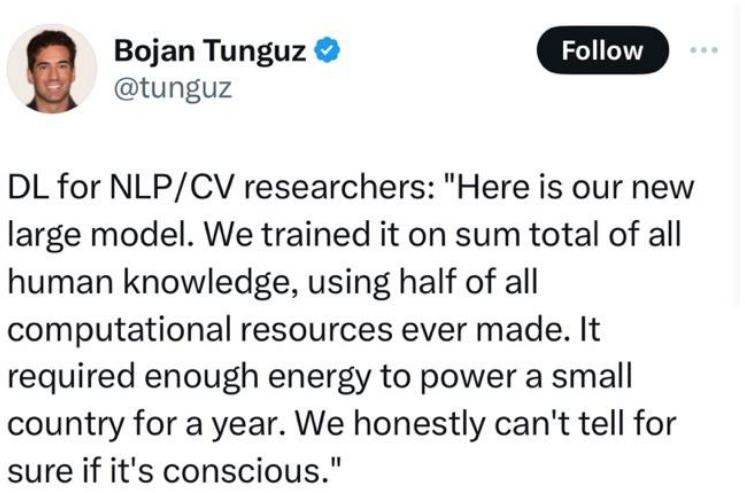
Tabular Data

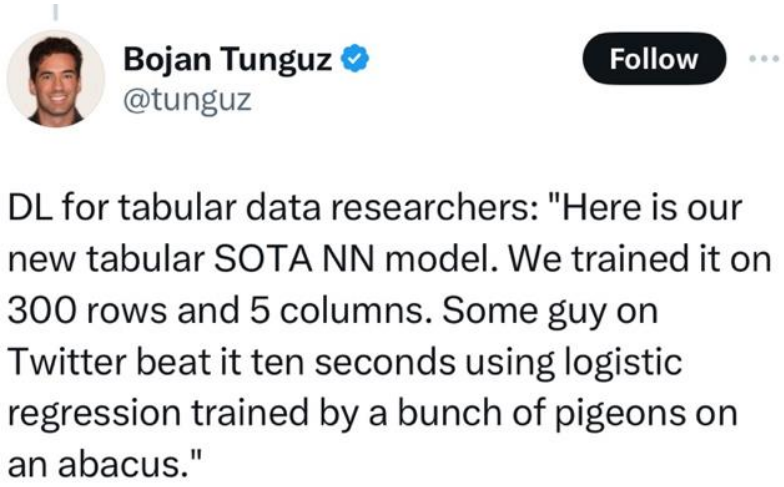
Introduction

Understanding Tabular Data

❖ **Why is it hard to model deep learning methods to tabular data?**

- 개별 변수 간 **형태적 다양성** (heterogeneity of individual features)
- 상대적으로 학습시키기 작은 사이즈의 데이터 유형
- 최근 TabNet과 같은 정형 데이터 딥러닝 모델 연구가 발표되고 있음





Introduction

Understanding Tabular Data

❖ **Why should we generate the tabular data?**

- 데이터 크기 증가 가능
- Regulation 문제를 우회해 더 범용적인 데이터의 사용이 가능해짐
- Privacy의 관점에서 사용하지 못했던 데이터를 사용할 수 있게 됨

Introduction

Understanding Tabular Data

❖ How should we train model for tabular data?

- 모델이 잘 학습할 수 있도록 데이터를 표현하는 것이 가장 중요함
- **범주형 데이터**의 처리가 관건!

1) Discrete한 데이터를 어떻게 Continuous한 공간으로 처리하는지가 중요한 문제
2) 범주형 데이터를 인코딩해서 숫자화하여 연속형 변수와 같이 딥러닝 네트워크에 적용

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
		Female	A		
		Male	O		
		Female	B		
		Female	AB		

Introduction

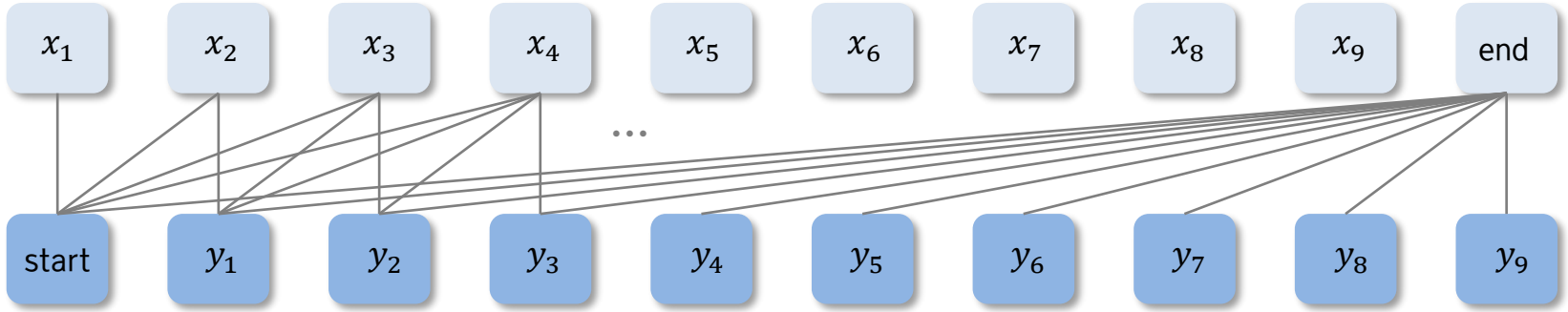
Understanding Generative Models for Discrete Data

❖ **Existing Discrete Generative Models**

- VAE : autoencoder with gaussian prior
- Flow based model : reversible transformation
- GAN : distribution matching with a discriminator

❖ **Existing best model for generating discrete data**

- Autoregressive models with Transformers (e.g., GPT and other language models)



Introduction

Understanding Generative Models for Discrete Data

❖ **What about diffusion models?**

- Computation and memory
  - 샘플의 각 timestep을 병렬적으로 처리 가능하여 training 및 inference가 상대적으로 간결함
  - 입력 차원에 비례하지 않는 학습 방식이므로 고차원의 입력에 적절한 생성 모델

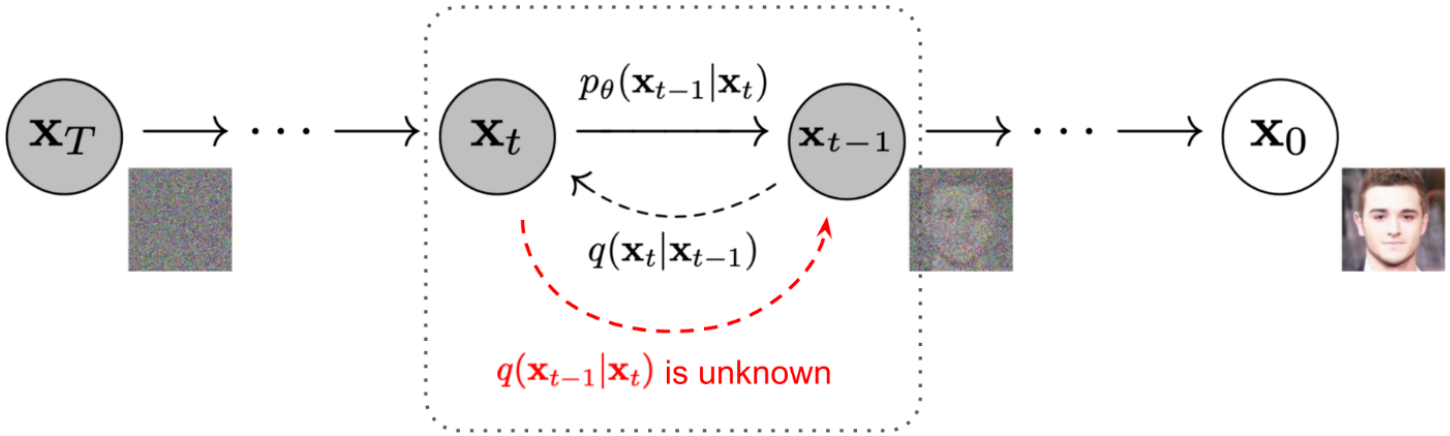


Fig. 2. The Markov chain of forward (reverse) diffusion process of generating a sample by slowly adding (removing) noise. (Image source: Ho et al. 2020 with a few additional annotations)

Introduction

Understanding Generative Models for Discrete Data

❖ **Can we model discrete data with diffusion models?**

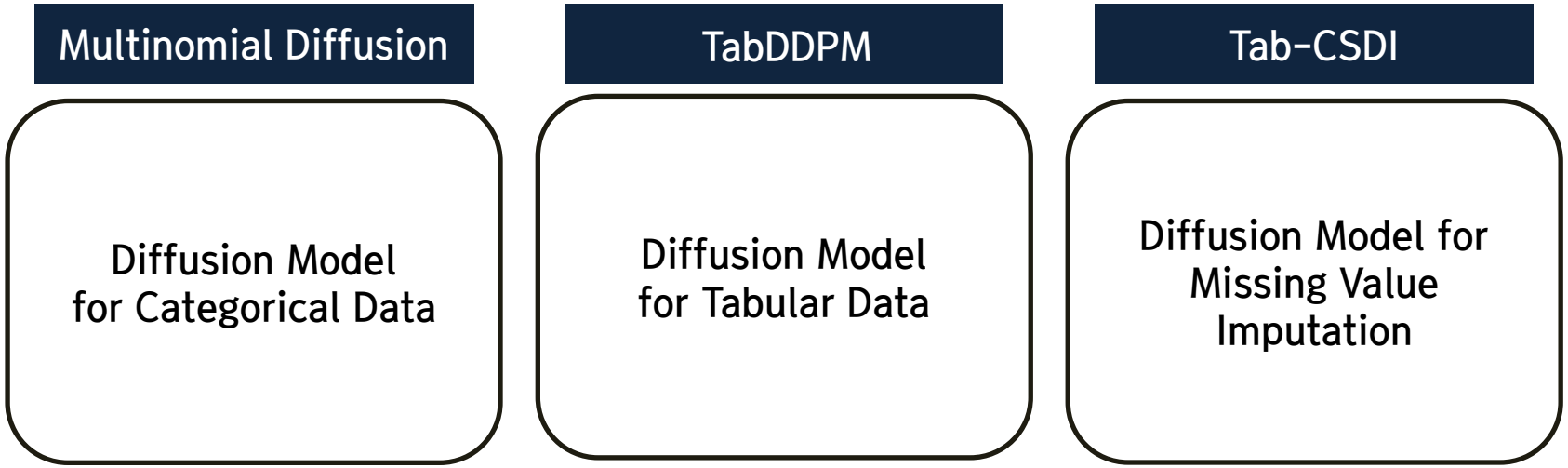
- Diffusion model : Continuous한 데이터를 처리하고 모델링하는데 적합한 모델
  - 기존 diffusion algorithm을 바로 적용할 수 없음
  - Discrete한 변수를 다루는 것이 가장 중요한 문제
  - tabular data 에서 discrete 데이터는 범주형 변수를 의미함

	Survived	Pclass	Sex	Parch	Fare	Embarked
0	0	3	Male	0	7.25	S
1	1	1	Female	0	71.28	C
2	1	3	Male	0	7.92	S
…	…	…	…	…	…	…
98	1	2	Female	1	23.10	S
99	0	2	male	0	16.00	S

	Survived	Pclass	Sex	Parch	Fare	Embarked
0	0	3	Male	0	7.25	S
1	1	1	Female	0	71.28	C
2	1	3	Male	0	7.92	S
…	…	…	…	…	…	…
998	1	3	Female	1	46.10	C
999	1	1	Male	0	24.23	S

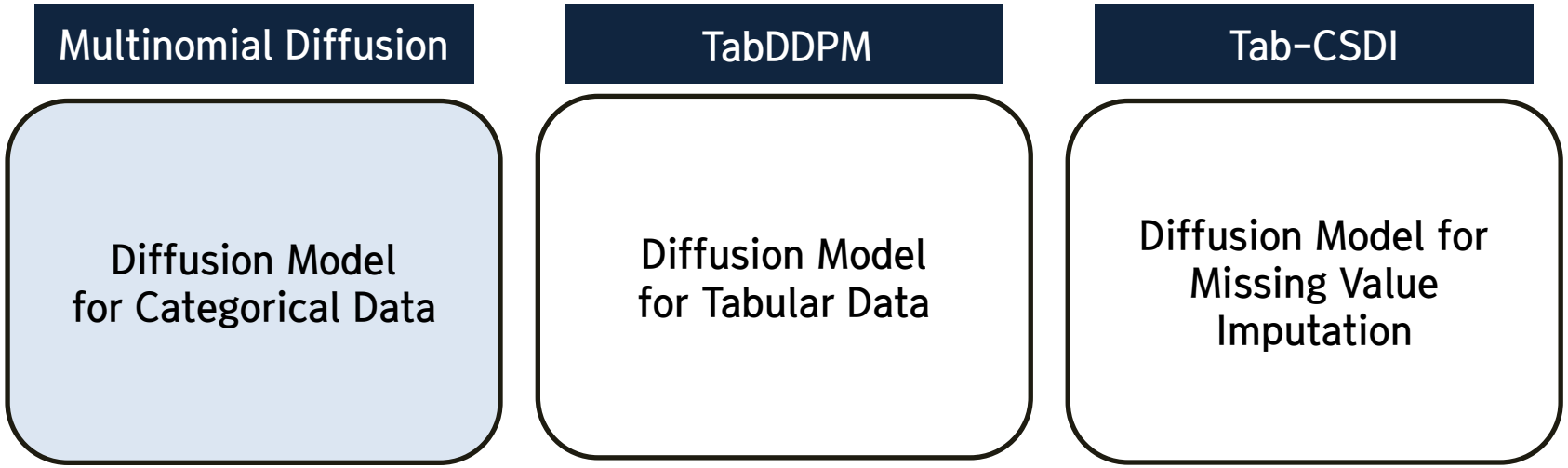
Diffusion Models

Diffusion Models for Tabular Data



Diffusion Models

Diffusion Models for Tabular Data



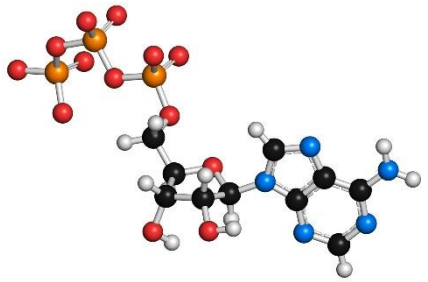
Multinomial Diffusion

Learning Categorical Distributions

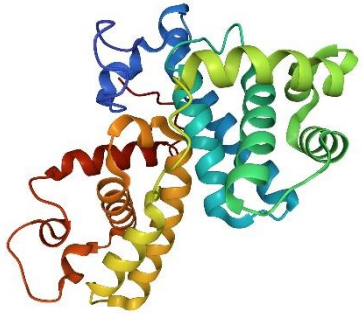
❖ **Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)**

- Motivation : categorical data를 적절하게 다루는 생성모델을 만들어보자
- 기존 범주형 변수는 autoregressive model을 사용해 다룸
  - 병렬화가 어려워 학습과 추론이 느릴 수 있음
  - 고차원의 입력에 적절하지 않은 방법론

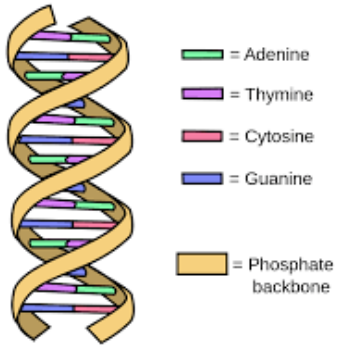
text



molecule



protein



DNA

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

Learning Categorical Distributions

❖ Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)

Categorical data를 다루는 다른 모델링 방법론은 없을까?

Discrete

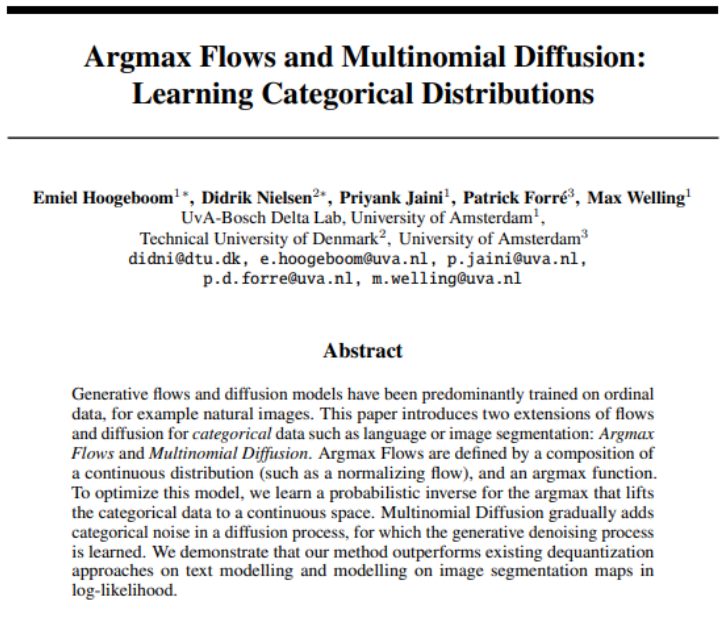
Our focus

Ordinal

- Ordered Categories
- Images, audio, etc

Categorical

- Unordered Categories
- Text, molecules, etc



Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions

Emiel Hoogeboom[1 *], **Didrik Nielsen**[2 *], **Priyank Jaini**[1], **Patrick Forré**[3], **Max Welling**[1]
UvA-Bosch Delta Lab, University of Amsterdam[1],
Technical University of Denmark[2], University of Amsterdam[3]
didni@dtu.dk, e.hoogeboom@uva.nl, p.jaini@uva.nl,
p.d.forre@uva.nl, m.welling@uva.nl

Abstract

Generative flows and diffusion models have been predominantly trained on ordinal data, for example natural images. This paper introduces two extensions of flows and diffusion for *categorical* data such as language or image segmentation: *Argmax Flows* and *Multinomial Diffusion*. Argmax Flows are defined by a composition of a continuous distribution (such as a normalizing flow), and an argmax function. To optimize this model, we learn a probabilistic inverse for the argmax that lifts the categorical data to a continuous space. Multinomial Diffusion gradually adds categorical noise in a diffusion process, for which the generative denoising process is learned. We demonstrate that our method outperforms existing dequantization approaches on text modelling and modelling on image segmentation maps in log-likelihood.

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

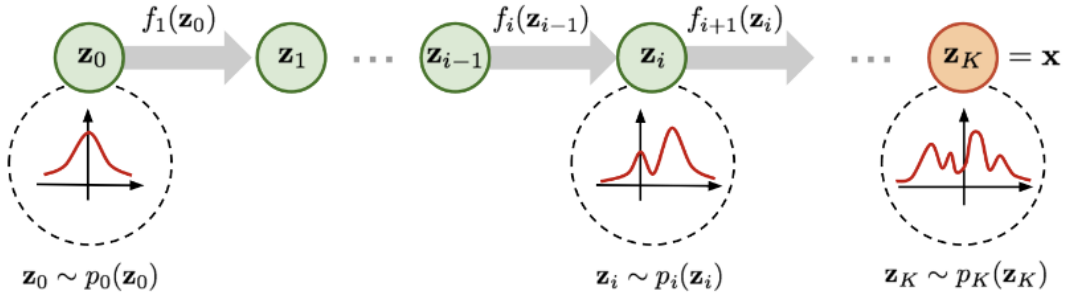
Learning Categorical Distributions

❖ Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)

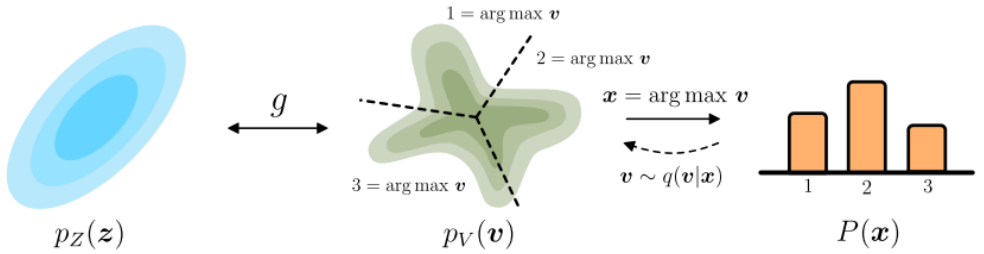
- Argmax Flow

✓ Flow-based generative models : 연속적인 역변환을 통해 데이터의 분포를 학습 / 생성하는 방식

✓ Normal Flow



✓ Argmax Flow



Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

Learning Categorical Distributions

❖ **Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)**

- Gaussian Diffusion

Fixed noising process : $q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}\,,\, \beta_t \mathrm{I}\right)$

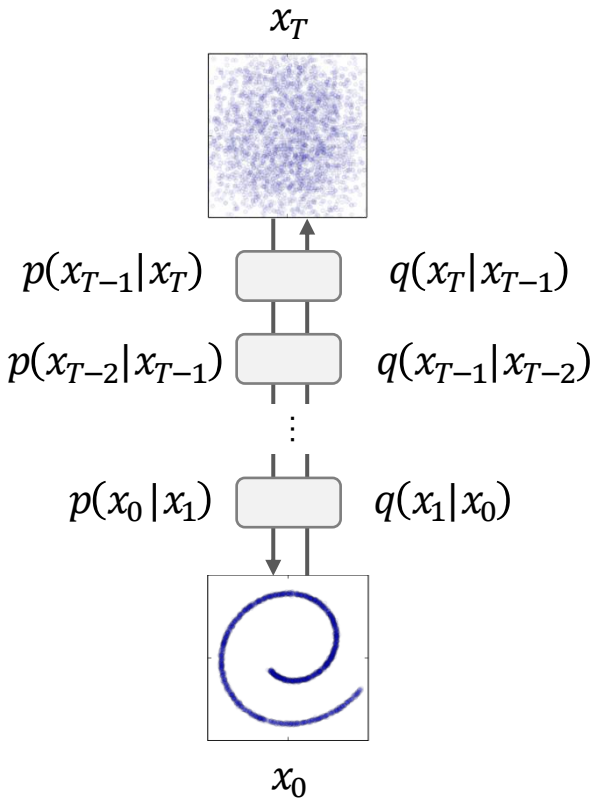
Learnable denoising process: $p(x_{t-1}|x_t) = N(x_{t-1};\, \mu_\theta(x_t)\,,\, \Sigma_\theta(x_t))$

Simple objective function :

$$L_{simple}(\theta) := \mathbb{E}_{t,x_0,\varepsilon}\left[\left|\left|\varepsilon - \varepsilon_\theta\left(\sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\varepsilon, t\right)\right|\right|^2\right]$$

Efficient training by sampling, using that:

$$\left.\begin{matrix} q(x_t|x_0) \\ q(x_{t-1}|x_t, x_0) \end{matrix}\right\} \text{ Closed-form Gaussian}$$



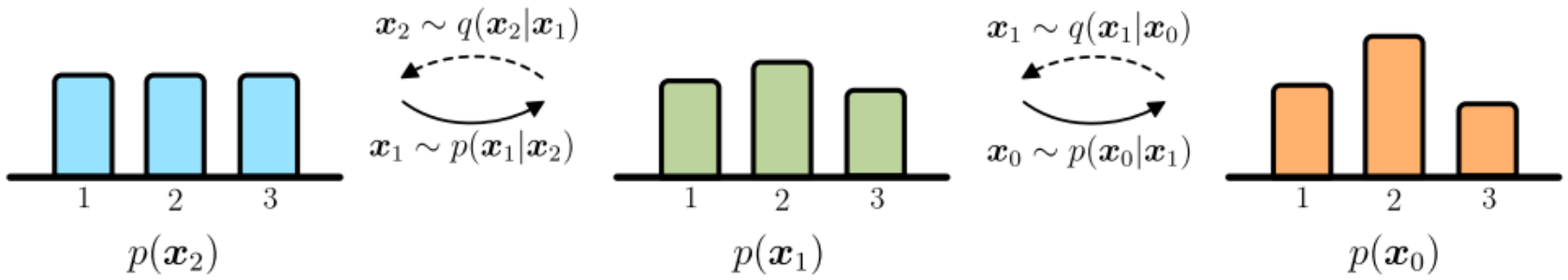
Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

Learning Categorical Distributions

❖ Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)

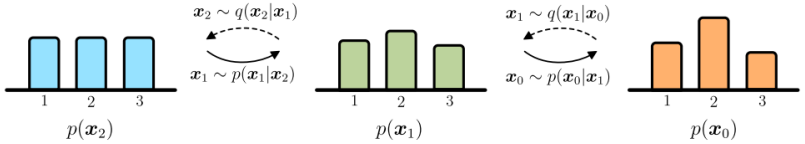
- Multinomial Diffusion
  - ✓ Idea : Discrete한 데이터를 diffusion model에 직접 적용
  - ✓ Language나 Image segmentation과 같은 범주형 데이터를 위한 diffusion 모델 개발



Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

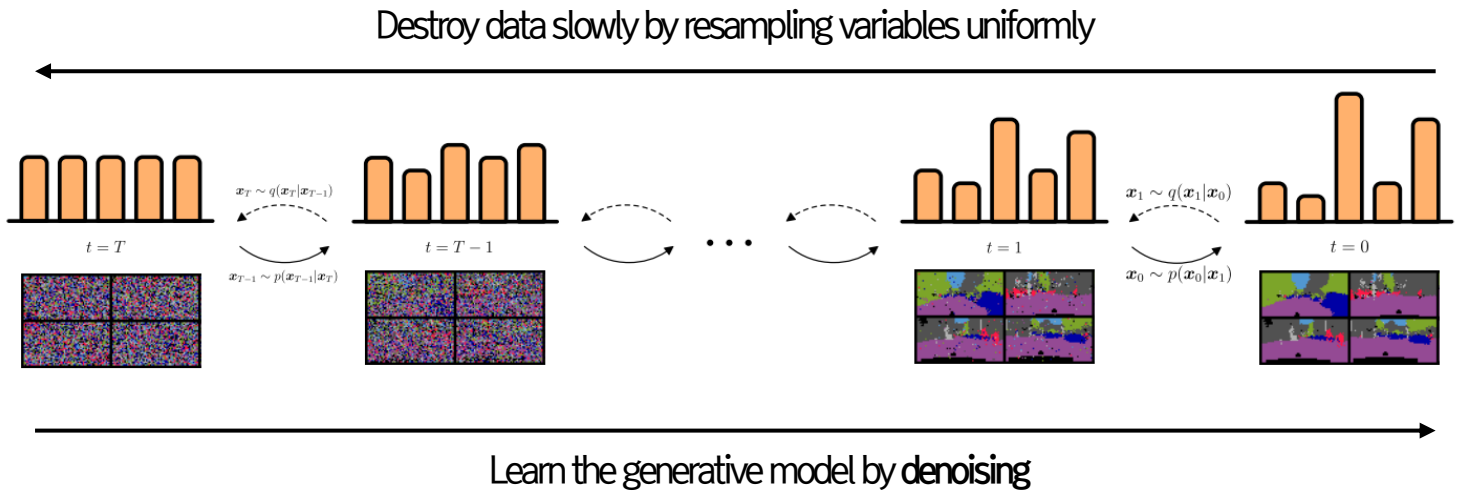
Multinomial Diffusion

Learning Categorical Distributions



❖ Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)

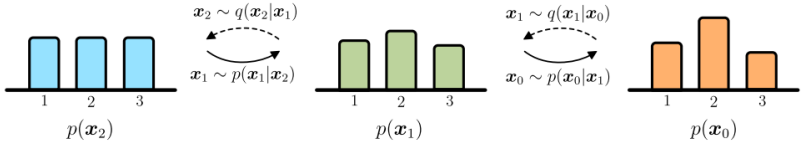
- Multinomial Diffusion
  - ✓ Gaussian noise를 주는 것처럼 Categorical noise를 점점 더해 줌
  - ✓ Categorical noise : 각 단계마다 새로운 카테고리에 속할 확률을 랜덤하게 넣어준다! 라고 생각
  - ✓ Pixel value가 가지는 각각의 categorical 한 확률을 점점 uniform 분포로 만드는 방식으로 노이즈를 더함



Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

Learning Categorical Distributions



❖ Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)

- Multinomial Diffusion

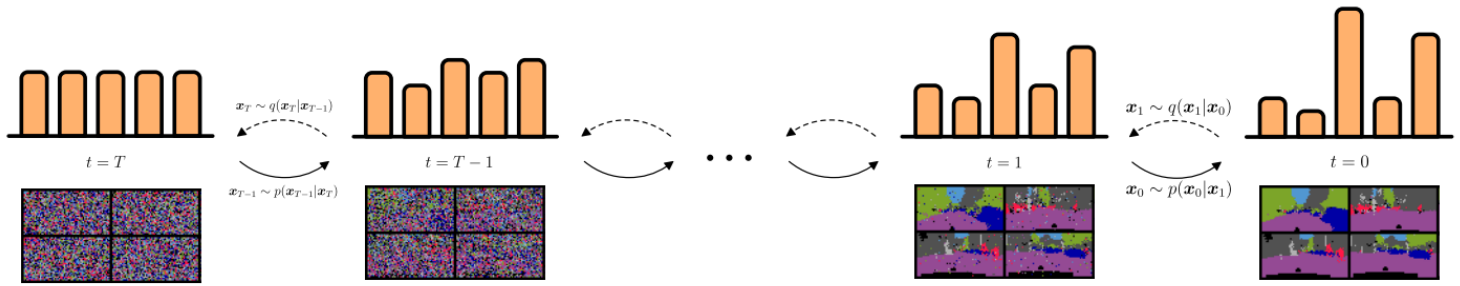
✓ Noising / Denoising process

β_t : chance of resampling a category uniformly

Cat : a categorical distribution with probability parameters after |

K : number of classes

$$\text{Destroy}: q(x_t|x_{t-1}) = Cat\left(x_t\,|(1-\beta_t)\,x_{t-1} + \frac{\beta_t}{K}\right)$$

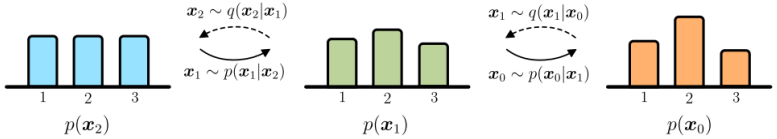


$$\text{Learn to denoise}: p(x_{t-1}|x_t) = Cat\left(x_{t-1}|\theta_{post}(x_t, \widehat{x_0})\right) \text{ where } \widehat{x_0} = \mu(x_t, t)$$

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

Learning Categorical Distributions



❖ **Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)**

- Multinomial Diffusion

Fixed noising process : $q(x_t|x_{t-1}) = Cat\left(x_t\,|(1-\beta_t)\,x_{t-1} + \beta_t\frac{1}{K}\right)$

Learnable denoising process: $p(x_{t-1}|x_t) = Cat(x_{t-1}|\,\pi_\theta(x_t,\hat{x_0})\,)$, *where* $\hat{x_0} = \mu(x_t,t)$

Objective function :

$$KL([q(x_{t-1}|x_t,x_0)|p(x_{t-1}|x_t)) = KL(Cat(\,\pi_\theta(x_t,x_0))|Cat(\,\pi_\theta(x_t,\hat{x_0}))$$

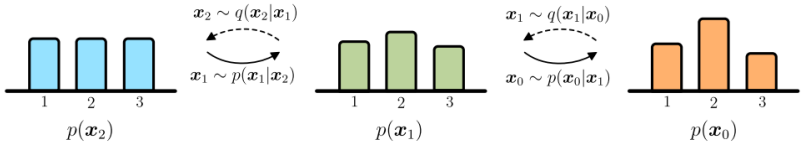
Efficient training by sampling, using that:

$$\left.\begin{matrix} q(x_t|x_0) \\ q(x_{t-1}|x_t,x_0) \end{matrix}\right\}$$ Closed-form Categoricals

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Multinomial Diffusion

Learning Categorical Distributions



❖ Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions(NeurIPS, 2021)

- Results
  - ✓ Categorical 분포를 학습하는 방법론 제안
  - ✓ Standard ARM, continuous autoregressive models, non autoregressive model 모두와 비교해도 우수한 성능을 가지는 것을 확인

Table 3: Comparison of different methods on `text8` and `enwik8`. Results are reported in negative log-likelihood with units bits per character (bpc) for `text8` and bits per raw byte (bpb) for `enwik8`.

Model type	Model	text8 (bpc)	enwik8 (bpb)
ARM	64 Layer Transformer (Al-Rfou et al., 2019)	1.13	1.06
	TransformerXL (Dai et al., 2019)	1.08	0.99
VAE	AF/AF* (AR) (Ziegler and Rush, 2019)	1.62	1.72
	IAF / SCF* (Ziegler and Rush, 2019)	1.88	2.03
	CategoricalNF (AR) (Lippe and Gavves, 2020)	1.45	-
Generative Flow	Argmax Flow, AR (ours)	1.39	1.42
	Argmax Coupling Flow (ours)	1.82	1.93
Diffusion	Multinomial Text Diffusion (ours)	1.72	1.75

```
mexico city the aztec stadium estadio azteca home of club america is on
e of the world s largest stadiums with capacity to seat approximately o
ne one zero zero zero zero fans mexico hosted the football world cup in
 one nine seven zero and one nine eight six
```

(a) Ground truth sequence from `text8`.

```
mexico citi the aztec stadium estadio azteca home of clup amerika is on
e of the world s largest stadioms with capakity to seat approsimately o
ne one zeto zero zero zero fans mexico hosted the footpall wolld cup in
 one nine zeven zero and one nyne eiggt six
```

(b) Corrupted sentence.

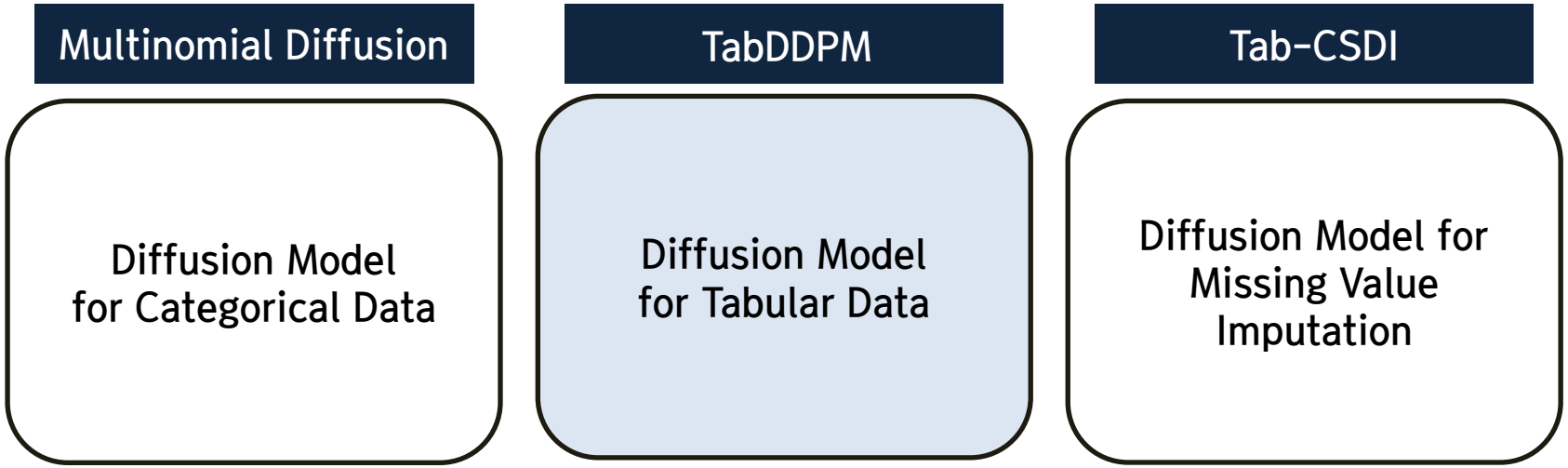
```
mexico city the aztec stadium estadio aztecs home of club america is on
e of the world s largest stadiums with capacity to seat approximately o
ne one zero zero zero zero fans mexico hosted the football world cup in
 one nine seven zero and one nine eight six
```

(c) Suggested, prediction by the model.

Hoogeboom, E., Nielsen, D., Jaini, P., Forré, P., & Welling, M. (2021). Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*

Diffusion models

Toward categorical data generation



Diffusion Model based on Tabular Data

Generating Tabular Data

❖ **TabDDPM : Modeling Tabular Data with Diffusion Models(2023, ICML)**

- Motivation : 일반적 tabular 문제들을 위한 DDPM 기반의 diffusion framework를 제안해보자
- Idea : Multinomial diffusion을 활용하여 원래의 데이터 분포와 유사한 형태의 새로운 tabular 데이터를 생성해보자!

TABDDPM: MODELLING TABULAR DATA WITH DIFFUSION MODELS

Akim Kotelnikov HSE, Yandex ya@akotelnikov.ru
Dmitry Baranchuk Yandex
Ivan Rubachev HSE, Yandex
Artem Babenko Yandex

ABSTRACT

Denoising diffusion probabilistic models are currently becoming the leading paradigm of generative modeling for many important data modalities. Being the most prevalent in the computer vision community, diffusion models have also recently gained some attention in other domains, including speech, NLP, and graph-like data. In this work, we investigate if the framework of diffusion models can be advantageous for general tabular problems, where datapoints are typically represented by vectors of heterogeneous features. The inherent heterogeneity of tabular data makes it quite challenging for accurate modeling, since the individual features can be of completely different nature, i.e., some of them can be continuous and some of them can be discrete. To address such data types, we introduce TabDDPM — a diffusion model that can be universally applied to any tabular dataset and handles any type of feature. We extensively evaluate TabDDPM on a wide set of benchmarks and demonstrate its superiority over existing GAN/VAE alternatives, which is consistent with the advantage of diffusion models in other fields. Additionally, we show that TabDDPM is eligible for privacy-oriented setups, where the original datapoints cannot be publicly shared. The source code of TabDDPM and our experiments is available at https://github.com/rotot0/tab-ddpm.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023, July). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning* (pp. 17564-17579). PMLR.

Diffusion Model based on Tabular Data

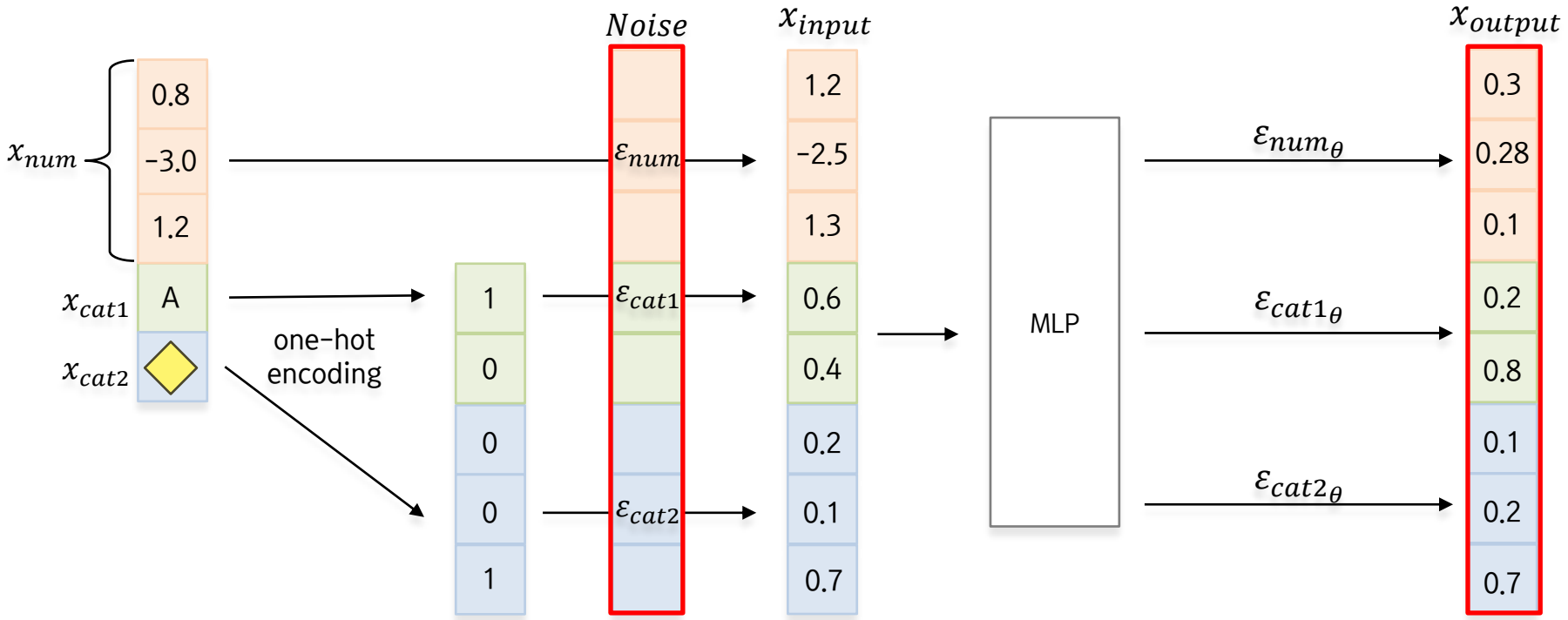
Generating Tabular Data

❖ **TabDDPM : Modeling Tabular Data with Diffusion Models(2023, ICML)**

- Model structure

$L_t^{simple} = \mathbb{E}_{x_0,\varepsilon,t}||\varepsilon - \varepsilon_\theta(x_t, t)||_2^2$: Gaussian diffusion term

$L_t^i = KL[q(x_{t-1}^i | x_t^i, x_0^i) | p_\theta(x_{t-1}^i | x_t^i)]$: Multinomial diffusion term



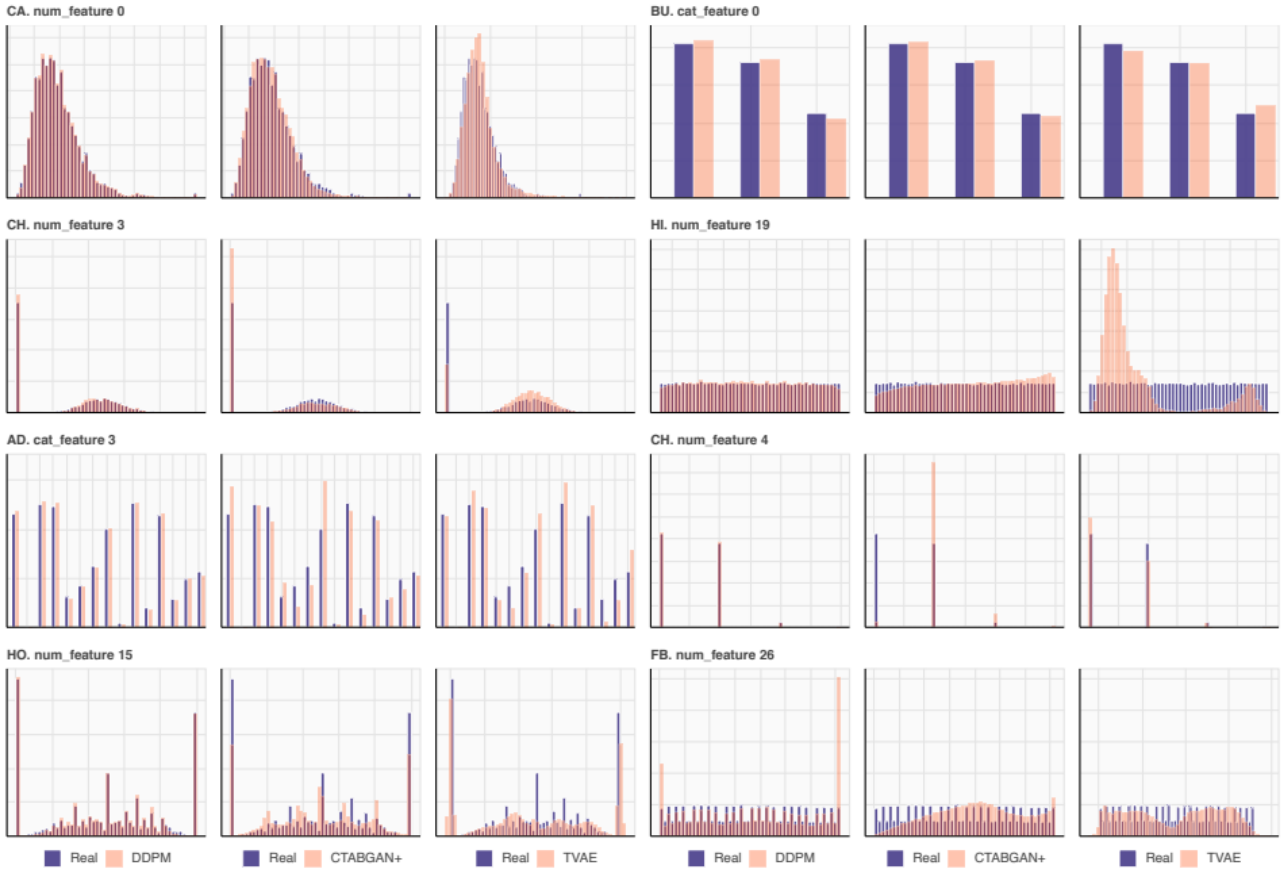
Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023, July). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning* (pp. 17564-17579). PMLR.

Diffusion Model based on Tabular Data

Generating Tabular Data

❖ **TabDDPM : Modeling Tabular Data with Diffusion Models(2023, ICML)**

- Experiments 1) Qualitative comparison (TabDDPM , CTABGAN+, TVAE 기반 데이터 생성 분포 비교)



Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023, July). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning* (pp. 17564-17579). PMLR.

Diffusion Model based on Tabular Data

Generating Tabular Data

❖ **TabDDPM : Modeling Tabular Data with Diffusion Models(2023, ICML)**

- Experiments 2) Machine Learning Efficiency

  – 전반적으로 SMOTE, TabDDPM 모델이 높은 성능을 가지는 것을 확인

	AB ($R2$)	AD ($F1$)	BU ($F1$)	CA ($R2$)	CAR ($F1$)	CH ($F1$)	DE ($F1$)	DI ($F1$)
TVAE	0.433±.008	0.781±.002	0.864±.005	0.752±.001	0.717±.001	0.732±.006	0.656±.007	**0.714±.039**
CTABGAN	–	0.783±.002	0.855±.005	–	0.717±.001	0.688±.006	0.644±.011	**0.731±.022**
CTABGAN+	0.467±.004	0.772±.003	0.884±.005	0.525±.004	0.733±.001	0.702±.012	0.686±.004	**0.734±.020**
SMOTE	**0.549±.005**	0.791±.002	0.891±.003	**0.840±.001**	0.732±.001	0.743±.005	**0.693±.003**	0.683±.037
TabDDPM	**0.550±.010**	**0.795±.001**	**0.906±.003**	0.836±.002	**0.737±.001**	**0.755±.006**	**0.691±.004**	**0.740±.020**
Real	0.556±.004	0.815±.002	0.906±.002	0.857±.001	0.738±.001	0.740±.009	0.688±.003	0.785±.013

	FB ($R2$)	GE ($F1$)	HI ($F1$)	HO ($R2$)	IN ($R2$)	KI ($R2$)	MI ($F1$)	WI ($F1$)
TVAE	0.685±.003	0.434±.006	0.638±.003	0.493±.006	0.784±.010	0.824±.003	0.912±.001	0.501±.012
CTABGAN	–	0.392±.006	0.575±.004	–	–	–	0.889±.002	**0.906±.019**
CTABGAN+	0.509±.011	0.406±.009	0.664±.002	0.504±.005	0.797±.005	0.444±.014	0.892±.002	0.798±.021
SMOTE	**0.803±.002**	**0.658±.007**	**0.722±.001**	0.662±.004	**0.812±.002**	**0.842±.004**	0.932±.001	**0.913±.007**
TabDDPM	0.713±.002	0.597±.006	**0.722±.001**	**0.677±.010**	0.809±.002	**0.833±.014**	**0.936±.001**	**0.904±.009**
Real	0.837±.001	0.636±.007	0.724±.001	0.662±.003	0.814±.001	0.907±.002	0.934±.000	0.898±.006

Table 4: The values of machine learning efficiency computed with regards to the state-of-the-art tuned CatBoost model.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023, July). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning* (pp. 17564-17579). PMLR.

Diffusion Model based on Tabular Data

Generating Tabular Data

❖ **TabDDPM : Modeling Tabular Data with Diffusion Models(2023, ICML)**

- Experiments 3) Privacy

  – Distance to Closest Record(DCR) : 실제의 데이터포인트와 합성된 데이터포인트까지의 거리

  – DCR 값이 클 수록 privacy regulation 문제를 피해 데이터를 유연하게 사용할 수 있음

	AB		AD		BU		CA		CAR		CH		DE		DI	
	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR
SMOTE	0.549	0.014	0.791	0.024	0.891	0.054	0.840	0.014	0.732	0.007	0.743	0.077	0.693	0.027	0.683	0.068
TabDDPM	0.550	**0.050**	0.795	**0.104**	0.906	**0.143**	0.836	**0.041**	0.737	**0.012**	0.755	**0.157**	0.691	**0.112**	0.740	**0.204**

	FB		GE		HI		HO		IN		KI		MI		WI	
	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR	score	DCR
SMOTE	0.803	0.027	0.658	0.023	0.722	0.319	0.662	0.056	0.812	0.030	0.842	0.066	0.932	0.016	0.913	0.007
TabDDPM	0.713	**0.112**	0.597	**0.059**	0.722	**0.449**	0.677	**0.086**	0.809	**0.041**	0.833	**0.189**	0.936	**0.022**	0.904	**0.016**

Table 5: ML efficiency CatBoost scores and privacy scores for SMOTE and TabDDPM models.

Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023, July). Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning* (pp. 17564-17579). PMLR.

Diffusion Model based on Tabular Data

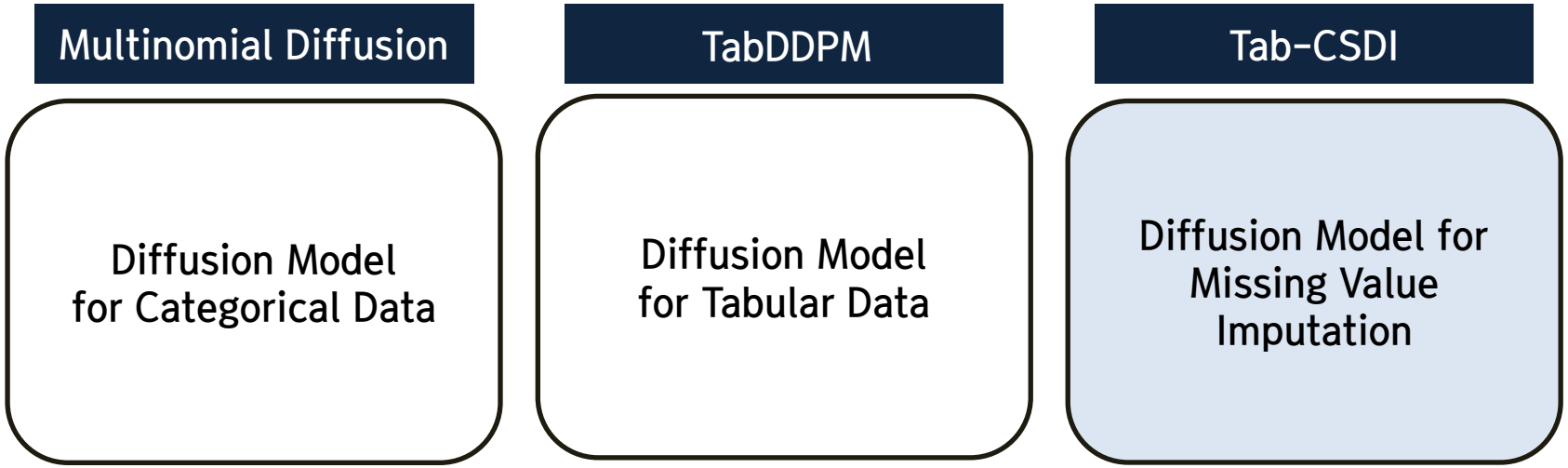
Generating Tabular Data

❖ TabDDPM : Modeling Tabular Data with Diffusion Models(2023, ICML)

- Conclusion
  - ✓ Tabular data에 적용가능한 diffusion model 방법론을 제안함
  - ✓ 실제 데이터가 아니라 실제 데이터에서 생성 되어 실제 데이터와 통계 속성이 동일한 데이터 생성
    - 실제 데이터를 수집해 가공하는 것보다 시간/경제적 비용 적음
    - 개인정보, 데이터 편향성 등의 문제를 우회해 데이터를 생성할 수 있음
    - 정보 보호와 privacy 문제를 피할 수 있음
  - ✓ 데이터 분포의 유사성, privacy의 관점에서의 차별성을 trade off 로 적당한 값을 가지는 데이터 생성

Diffusion models

Toward categorical data generation



Diffusion Model based on Tabular Data

Missing Value Imputation

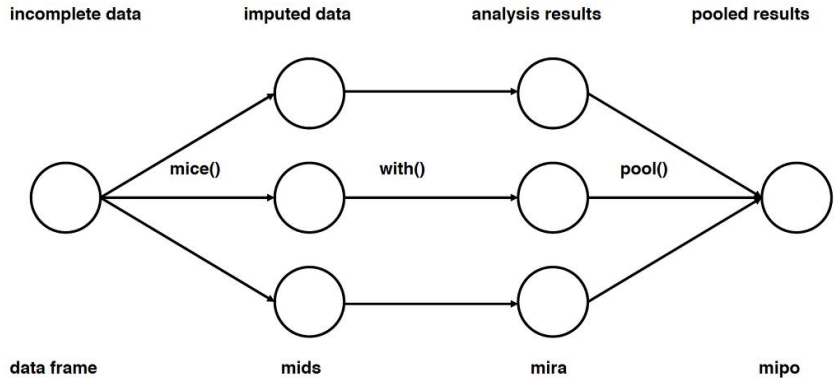
❖ How to fill the missing values in tabular data?

- 실제 데이터에는 다양한 유형의 결측치가 존재
- 전통적 결측치 처리 방법

  – 통계값으로 대체 , K-NN 알고리즘 기반 근접 데이터 대체 , MICE 기반 대치 세트 평균값으로 결과 대체 등

생성 모델을 활용하여 결측치를 채울 수 없을까?

Yes	Male	?	…	29	1.72
No	Female	4.9	…	37	1.62
?	Female	2.8	…	?	1.63
?	Male	3.7	…	?	?
Yes	?	2.4	…	42	1.80

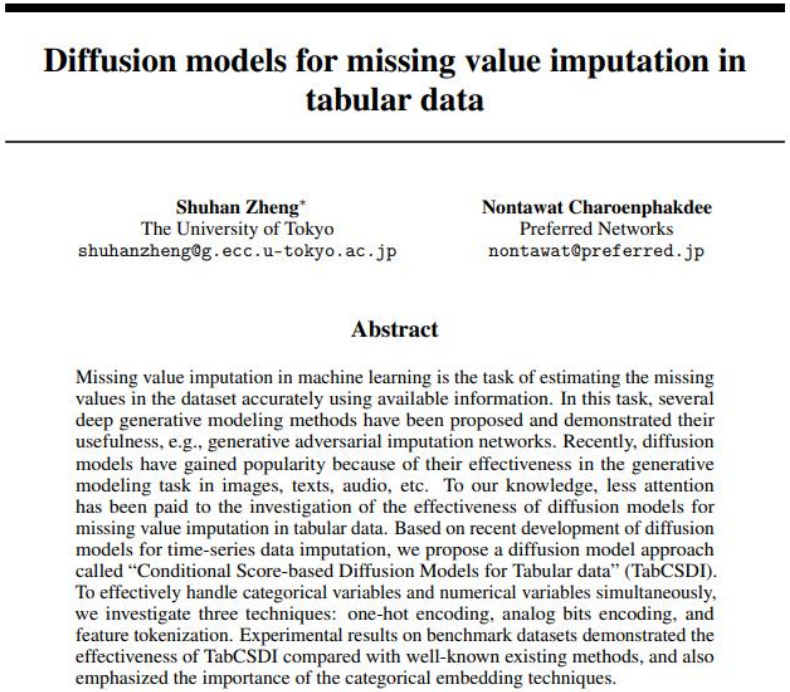


Diffusion Model based on Tabular Data

Missing Value Imputation

❖ **Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)**

- Diffusion model을 활용한 tabular 데이터 결측치 imputation 연구
- 시계열 데이터 imputation 연구인 CSDI를 정형 데이터 imputation 연구로 확장
- CSDI는 범주형 변수를 다룰 수 없다는 한계를 가짐



Diffusion models for missing value imputation in tabular data

Shuhan Zheng*
The University of Tokyo
shuhanzheng@g.ecc.u-tokyo.ac.jp

Nontawat Charoenphakdee
Preferred Networks
nontawat@preferred.jp

Abstract

Missing value imputation in machine learning is the task of estimating the missing values in the dataset accurately using available information. In this task, several deep generative modeling methods have been proposed and demonstrated their usefulness, e.g., generative adversarial imputation networks. Recently, diffusion models have gained popularity because of their effectiveness in the generative modeling task in images, texts, audio, etc. To our knowledge, less attention has been paid to the investigation of the effectiveness of diffusion models for missing value imputation in tabular data. Based on recent development of diffusion models for time-series data imputation, we propose a diffusion model approach called "Conditional Score-based Diffusion Models for Tabular data" (TabCSDI). To effectively handle categorical variables and numerical variables simultaneously, we investigate three techniques: one-hot encoding, analog bits encoding, and feature tokenization. Experimental results on benchmark datasets demonstrated the effectiveness of TabCSDI compared with well-known existing methods, and also emphasized the importance of the categorical embedding techniques.

Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. Table Representation Learning Workshop at NeurIPS 2022

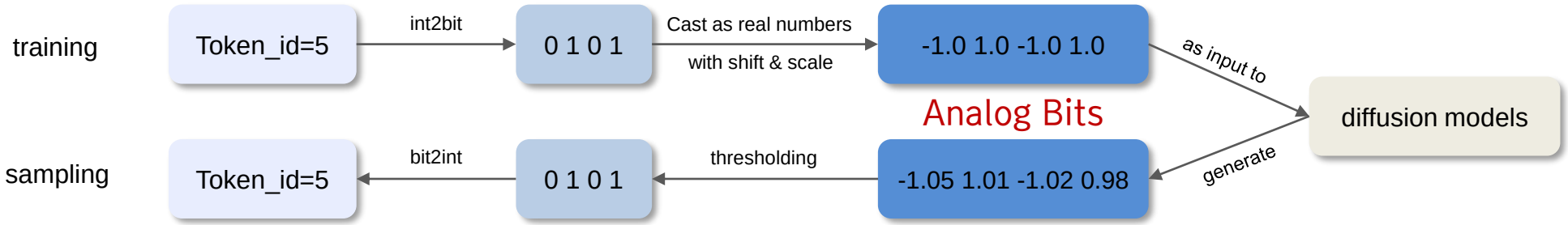
Diffusion Model based on Tabular Data

Missing Value Imputation

❖ **Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)**

- 범주형 변수를 처리하는데 세가지 방법 사용

(1) analog bits encoding (2) feature tokenization (3) one-hot encoding



- Analog Bits : discrete한 데이터를 나타내는 비트를 모델링하는데 사용되는 실수
- 이진형 bits $\{0,1\}^n$를 단순하게 실수 R^n으로 cast 함으로써 연속형 diffusion 모델로부터 직접 모델링 될 수 있음
  - 이산형 state space / re-formulate diffusion process 불필요

Chen, T., Zhang, R., & Hinton, G. (2022). Analog bits: Generating discrete data using diffusion models with self-conditioning. *ICLR 2023*

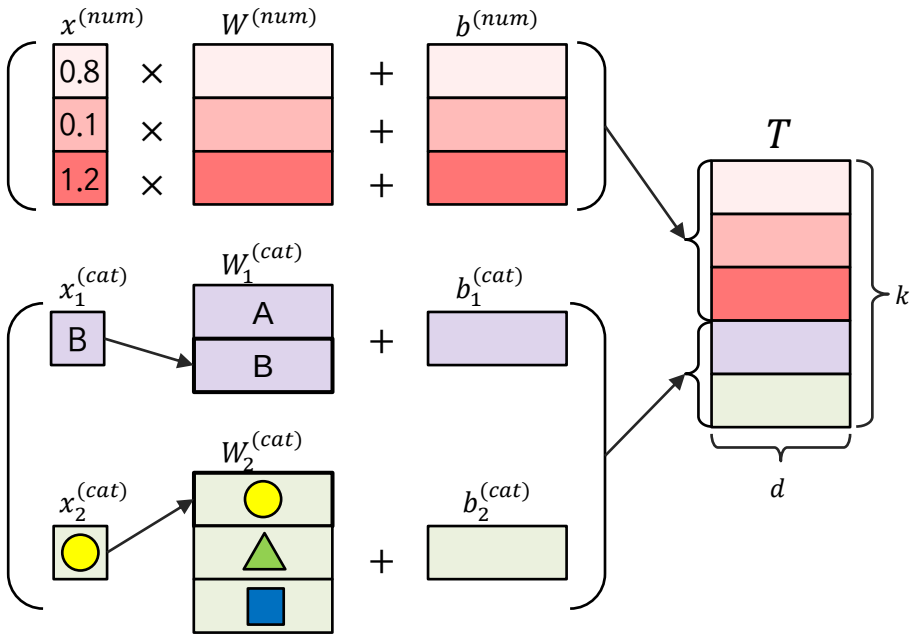
Diffusion Model based on Tabular Data

Missing Value Imputation

❖ Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)

- 범주형 변수를 처리하는데 세가지 방법 사용

(1) analog bits encoding **(2) feature tokenization** (3) one-hot encoding



Feature Tokenizer

- 모든 특성(범주형, 숫자형)을 임베딩
- 입력 x를 임베딩 $T \in R^{k \times d}$로 변환
- 주어진 특성 x_j에 대한 임베딩
  - $T_j = b_j + f_j(x_j) \in R^d$

Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. Table Representation Learning Workshop at NeurIPS 2022

Diffusion Model based on Tabular Data

Missing Value Imputation

❖ **Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)**

- 범주형 변수를 처리하는데 세가지 방법 사용

  (1) analog bits encoding (2) feature tokenization **(3) one-hot encoding**

  - 각각의 범주를 독립적인 이진변수로 변환

Color
Red
Yellow
Yellow
Green
Red

→

Red	Yellow	Green
1	0	0
0	1	0
0	1	0
0	0	1
1	0	0

Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. Table Representation Learning Workshop at NeurIPS 2022

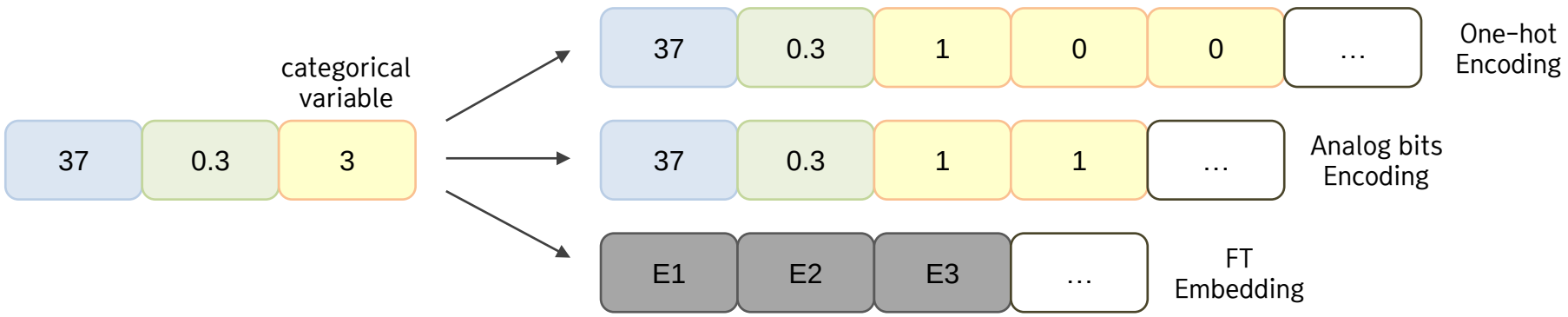
Diffusion Model based on Tabular Data

Missing Value Imputation

❖ **Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)**

- 범주형 변수를 처리하는데 세가지 방법 사용

(1) analog bits encoding (2) feature tokenization (3) one-hot encoding



- Analog bits 인코딩 : 원핫인코딩에 비해 칼럼 수가 적지만 인코딩 된 벡터의 복잡성은 상대적으로 큼
- FT Embedding : E1, E2, E3 모두 같은 길이를 가질 수 있음

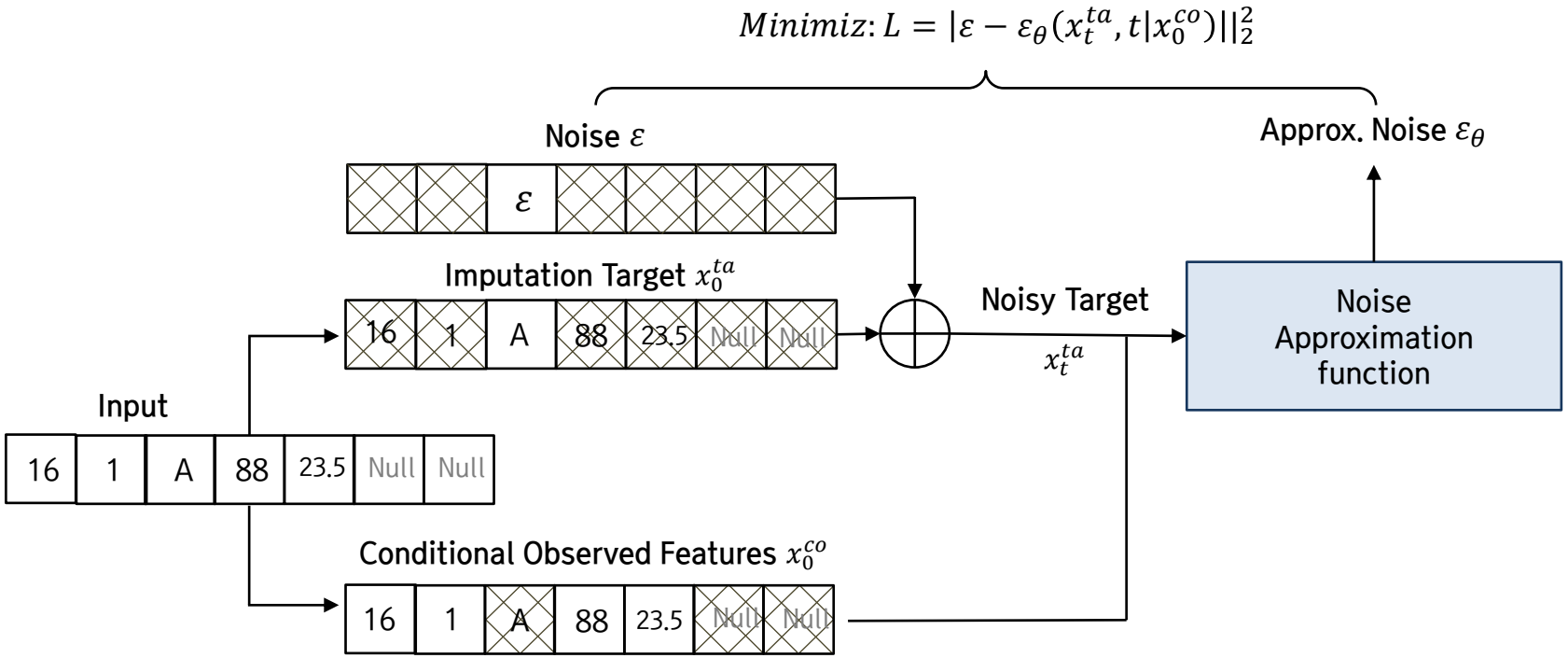
Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. Table Representation Learning Workshop at NeurIPS 2022

Diffusion Model based on Tabular Data

Missing Value Imputation

❖ Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)

- Imputation task : $x_0^{co} \rightarrow x_0^{ta}$



Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. Table Representation Learning Workshop at NeurIPS 2022

Diffusion Model based on Tabular Data

Missing Value Imputation

❖ Diffusion model for missing value imputation in tabular data(2022, NeurIPS Workshop)

- Experiments

  – 7개의 벤치마크 데이터셋 사용

Table 1: RMSE and error rate performance for comparison methods on three mixed variable datasets. Note that one-hot and analog bits are equivalent for a dataset without multi-categorical variables.

	Diabetes		COVID-19		Census	
	RMSE	Error rate	RMSE	Error rate	RMSE	Error rate
Mean / Mode	0.222 (0.003)	0.260 (0.004)	0.138 (0.002)	0.144 (0.002)	0.120 (0.003)	0.424 (0.003)
MICE (linear)	0.263 (0.002)	0.270 (0.004)	0.125 (0.003)	0.300 (0.038)	0.101 (0.002)	0.530 (0.011)
MissForest	0.216 (0.003)	**0.214 (0.001)**	**0.120 (0.002)**	0.131 (0.002)	0.112 (0.004)	**0.300 (0.014)**
GAIN	0.202 (0.003)	0.282 (0.005)	0.127 (0.002)	0.217 (0.011)	0.123 (0.057)	0.412 (0.012)
TabCSDI/ one-hot	**0.197 (0.001)**	0.222 (0.005)	0.122 (0.003)	0.111 (0.012)	0.099 (0.004)	0.400 (0.033)
TabCSDI/ analog bits	**0.197 (0.001)**	0.222 (0.005)	0.122 (0.003)	0.111 (0.012)	0.103 (0.004)	0.376 (0.013)
TabCSDI/ FT	0.206 (0.002)	0.224 (0.004)	0.123 (0.002)	**0.107 (0.002)**	**0.098 (0.003)**	0.345 (0.002)

Table 2: RMSE performance of comparison methods on four pure numerical datasets.

Methods	Wine	Concrete	Libras	Breast
Mean	0.076 (0.003)	0.217 (0.007)	0.099 (0.001)	0.263 (0.009)
MICE (linear)	0.065 (0.003)	0.153 (0.006)	0.034 (0.001)	0.154 (0.011)
MissForest	**0.060 (0.002)**	0.173 (0.005)	0.024 (0.001)	0.163 (0.014)
GAIN	0.072 (0.004)	0.203 (0.007)	0.089 (0.006)	0.165 (0.006)
TabCSDI	0.065 (0.004)	**0.131 (0.008)**	**0.011 (0.001)**	**0.153 (0.003)**

Zheng, S., & Charoenphakdee, N. (2022). Diffusion models for missing value imputation in tabular data. Table Representation Learning Workshop at NeurIPS 2022

Conclusion

Diffusion models for tabular data

❖ **Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions**

- Image segmentation, text 와 같은 categorical한 성격을 가지는 데이터를 위한 diffusion 방법론 제안
- Loss 계산 시 categorical한 정보를 제공하는 term 부여

❖ **TabDDPM : Modeling Tabular Data with Diffusion Models**

- Tabular data에 diffusion 모델 방법론을 적용
- 실제 데이터가 아니기 때문에, 정보 보호와 privacy 문제를 피할 수 있음

❖ **Diffusion model for missing value imputation in tabular data**

- 정형데이터에 여러 인코딩/임베딩을 적용하여 데이터 imputation 성능 비교

Thank you