UniformGAN: generative adversarial networks in uniform probability spaces

Marc Visser, Supervisors: Zilong Zhao *, Lydia Chen *

* supervisors

1. Motivation and Background

- **Privacy** is a **key** challenge for **sharing data** in industry due to GDPR.
- Generating **synthetic tabular data** using generative adversarial networks is one **solution** to empower big discovery while respecting the constraints of data privacy.
- Existing solutions try to model the **cross-correlation** in the GAN [1], but the data can be **enhanced** with the **integral probability transform** leverged in copulaGAN[3] in order to better capture the local dependency structure and improve training time.

UniformGAN builds on state-of-the-art CTAB-GAN[2] in order to improve cross correlation of synthetic data.

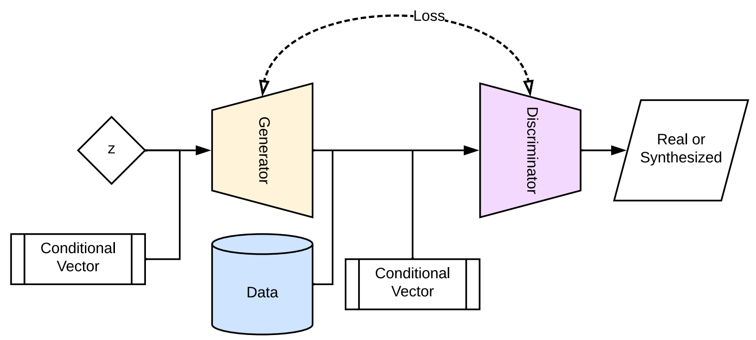

Figure 1. CTAB-GAN Model

2. Method
What is Uniform GAN?

UniformGAN is a tabular data generator which is based on CTAB-GAN designed to improve modeling speed by transforming continuous variables into uniform probability space in order for the GAN to make learning the underlying distribution easier.

- data is **encoded** to represent **boolean**, **categorical datetime** and **numerical types** as numerical values.
- Fit distributions and convert data using integral probabiliy transform, mapping it into uniform probability space.
- Then the discriminator in CTAB-GAN is fed the transformed data.

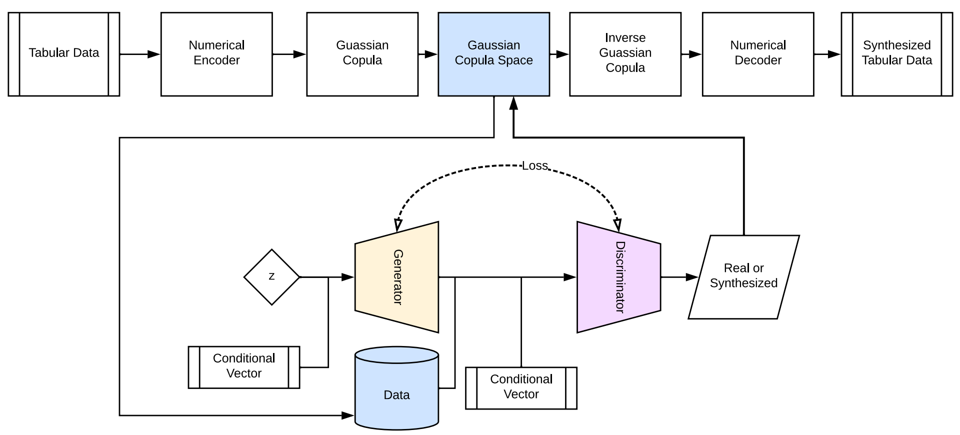

Figure 2. Uniform GAN architechture

3. Pipeline

To asses how well the synthesized data performs compared to the real data we created a utility pipeline as seen in figure 3. A similar pipeline is created for statistical similarity and privacy preservation.

Three metrics are considered with respect to **machine learning utility**;

- **Accuracy difference**, **Area Under Curve** (AUC) difference, and **F1-score** difference.
- To assess the **statistical similarity** we consider the average **Wasserstein distance**, average **Jensen–Shannon divergence**, and **correlation distance**.
- In order to assess the **privacy** we run a **nearest neighbour analysis**.

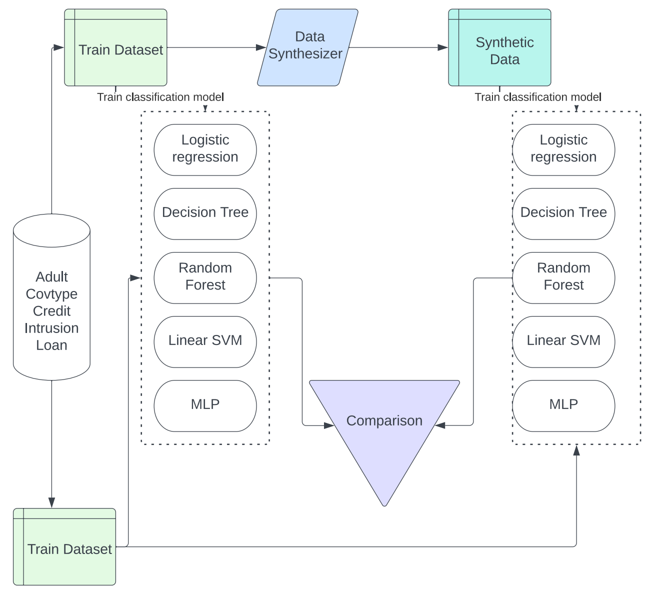

Figure 3. Utility Pipeline

4. Results

Model	ML Utility Difference			Statistical Similarity			Privacy Preservation					
							DCR			NNDR		
	Accuracy	AUC	F1-score	Avg JSD	Avg WD	Diff .Corr.	R&S	R	S	R&S	R	S
UniformGAN	**8.708**	**0.115**	**0.176**	**0.013**	**0.0761**	**3.210**	**1.373**	**0.308**	**0.958**	**0.782**	**0.421**	**0.623**
CTAB-GAN	11.205	0.134	0.205	0.331	0.070	1.900	1.260	0.3088	1.0840	0.751	0.4219	0.620
Copulas	18.998*	0.189	0.323	0.0172	0.126	3.703	1.759	0.308	1.584	0.826	0.421	0.745
CopulaGAN	29.97	0.21	0.371	0.082	0.294	5.814	1.424	0.201	0.535	0.815	0.337	0.538
CTGAN	35.442*	0.232	0.356	0.047	0.221	4.57	1.304	0.232	0.831	0.749	0.347	0.61

Table 4: Results 50 epochs: Average over Adult, Covtype, Intrusion and Insurance

REFERENCES

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proceedings of the 27th NIPS - Volume 2, page 2672–2680, Cambridge, MA, USA, 2014.

[2] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen. Ctab-gan: Effective table data synthesizing. In Proceedings of The 13th Asian Conference on Machine Learning, volume 157, pages 97–112, 17–19 Nov 2021.

[3] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. GaussianCopula - The synthetic data vault SDV. Proceedings - 3rd IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, pages 399–410, 2016.