# CTGAN (Conditional Generative Adversarial Network)

**Overview**

CTGAN is a variant of the Generative Adversarial Network specifically designed for generating synthetic tabular data. It addresses the challenges posed by the discrete nature of many tabular datasets and the need for conditional generation based on specific column values. CTGAN has been particularly effective in scenarios where traditional GANs struggle due to the sparsity and imbalance of categorical data.

Architecture

CTGAN consists of three main components: the Generator, the Discriminator, and the Data Transformer. Each plays a crucial role in the synthetic data generation process.

1. Generator

- Purpose: To generate synthetic data samples that mimic the real data distribution.
- Architecture:
- Built using a series of residual blocks followed by a linear layer.
- Each residual block consists of a linear transformation, batch normalization, and a ReLU activation function, followed by a concatenation with the input (residual connection).
- The final output is produced by a linear transformation that reshapes the data into the desired dimension.

2. Discriminator

- Purpose: To distinguish between real and synthetic data samples.
- Architecture:
- Composed of a series of linear layers with LeakyReLU activations and dropout for regularization.
- The output layer is a single neuron with a sigmoid activation function to classify inputs as real or fake.
- Includes a gradient penalty function to stabilize training by enforcing the Lipschitz constraint.

3. Data Transformer

Purpose: To preprocess the data by normalizing continuous features and encoding categorical features, making them suitable for processing by the GAN.

Components:

- Continuous Columns: Transformed using a Gaussian mixture model to normalize and discretize the data.
- Discrete Columns: Encoded using one-hot encoding.
- The transformer also handles the inverse transformation to convert the synthetic data back to the original data format.

Training Process

CTGAN's training involves alternating updates between the Discriminator and the Generator:

1. Discriminator Training:

- Real data samples and their corresponding conditions are passed to the Discriminator.
- Synthetic data samples generated by the Generator, along with their conditions, are also passed to the Discriminator.
- The Discriminator is trained to maximize the probability of correctly classifying both real and synthetic samples.

2. Generator Training:

- The Generator receives a random noise vector and a conditional vector (if conditional generation is used).
- It aims to generate data that will be classified as real by the Discriminator.
- The Generator's training objective is to minimize the likelihood of the Discriminator correctly classifying its outputs as fake.

3. Gradient Penalty:

- Applied during Discriminator training to enforce the Lipschitz constraint, crucial for the convergence of GANs.

Key Functionalities

- Synthetic Data Generation: Capable of generating realistic synthetic data that adheres to the statistical properties of the original dataset.
- Conditional Generation: Can generate data conditioned on specific column values, useful for scenarios requiring balanced datasets across certain features.
- Handling Mixed Data Types: Effectively processes both continuous and categorical data, a common characteristic of tabular datasets.

Parameter Influence and Experimentation

- Network Architecture: The depth and width of the Generator and Discriminator can significantly impact the quality of the generated data.

- Learning Rates and Decay: These parameters control the speed and stability of the training process.
- Batch Size and Discriminator Steps: Influence the training dynamics and convergence behavior of the GAN.

Usage Scenarios

- Data Augmentation: Enhances existing datasets for improved machine learning model performance.
- Imbalanced Data Handling: Generates synthetic samples to balance datasets, particularly useful in training robust classification models.
- Privacy-Preserving Data Sharing: Generates data that can be shared without exposing sensitive information in scenarios where data privacy is crucial.