Ganblr Plus Plus (GANBLR++)

Introduction

GANBLR++ is an advanced synthetic data generation framework designed to handle complex tabular datasets containing both numerical and categorical variables. It extends the capabilities of traditional GANs by integrating specialized preprocessing and postprocessing steps that cater to the unique challenges posed by mixed data types, ensuring the synthetic data maintains statistical fidelity to the original data while adhering to privacy constraints.

Detailed Architecture and Components:

GANBLR Core Model

Purpose: At the heart of GANBLR++ lies the GANBLR model, which is responsible for generating synthetic categorical data.

Components:

- Generator: Creates synthetic data from random noise. It learns to mimic the distribution of the transformed (discretized) real data.
- Discriminator: Attempts to distinguish between real and synthetic data, providing feedback to the generator on its performance.
- Classifier: Ensures that the generated data maintains the conditional distributions of the original data, particularly focusing on the label or target variable distributions.

DMMDiscritizer

Purpose: Converts numerical data into a categorical format that can be processed by the GANBLR model, using a Bayesian Gaussian Mixture Model.

Functionality:

- Fit: Learns the parameters of Gaussian mixtures that best represent the distribution of each numerical feature.
- Transform: Assigns each numerical value to a category based on the learned Gaussian mixture parameters, effectively discretizing the numerical data.
- Inverse Transform: Reverts the discretized data back to its original numerical form by sampling from the corresponding Gaussian distribution, guided by the categorical representation.

Preprocessing and Postprocessing

- Ordinal and Label Encoding: Converts categorical variables into a numerical format suitable for processing by neural networks.
- Scaling: Applies MinMax scaling to numerical features to normalize their range, which is crucial for effective learning in neural network models.

Training and Operation

Training Process

Step 1: Data Preparation

- Numerical features are discretized using the DMMDiscritizer.
- Categorical features are encoded using ordinal or label encoders.

Step 2: GAN Training

- The GANBLR model is trained on the transformed dataset. The generator learns to produce data that the discriminator cannot distinguish from real, transformed data.
- The classifier component ensures that the synthetic data adheres to the label distributions of the original dataset.

Step 3: Post-Processing

- The synthetic data generated by the GANBLR model is converted back to its original form:
- Discretized numerical data is transformed back using the inverse transform method of DMMDiscritizer.
- Categorical data is decoded using the inverse of the applied encodings.

Key Functionalities

- Synthetic Data Generation: Generates high-quality synthetic data that can be used for various purposes such as training machine learning models where data privacy is a concern.
- Privacy Preservation: By transforming data and generating new samples from learned distributions, GANBLR++ ensures that the synthetic data does not contain identifiable information from the original dataset.
- Handling Mixed Data Types: Through its preprocessing and postprocessing steps, GANBLR++ effectively manages datasets that contain a mix of numerical and categorical data, a common scenario in real-world datasets.

Parameter Influence and Experimentation

- GAN Parameters: Learning rate, number of epochs, and the architecture of the generator and discriminator significantly affect the quality of the generated data.
- Mixture Model Parameters: The number of components and the convergence criteria of the Bayesian Gaussian Mixture Model influence how well the numerical data is discretized and reconstructed.
- Privacy Settings: Adjusting parameters related to privacy, such as noise levels and differential privacy settings, can help balance the trade-off between data utility and privacy.

Usage Scenarios

- Data Augmentation: Enhances training datasets in scenarios where data is limited or imbalanced.
- Privacy-Preserving Data Sharing: Facilitates the sharing of data across organizations without compromising sensitive information, suitable for collaborative research or cross-organizational data analysis.