Synthetic Tabular Data with CTGAN for ML Projects

Nishchay Krishnappa — Project Repository:-https://github.com/nishchaykrishnappaiu/hybrid-gan-synth-data-nishchay-krishnappa-42308956-CSEMCSPCSP01

1 Problem Statement

Currently, privacy concerns and the difficulty of locating or gathering high-quality data make it difficult for researchers and students to obtain enough of it for machine learning projects. When using tabular datasets for classification jobs, I encountered the same issues myself. This inspired me to figure out how to provide practical, meaningful data when there isn't enough.

2 Abstract

To create synthetic tabular data, I want to employ a Generative Adversarial Network (GAN). Instead of attempting to cover all types (such as text or photos), I'm concentrating on CSV tables, such as those found in the UCI repository or on Kaggle. The `CTGAN` library is what I'm using to create the project in Python because it's made for tabular data with mixed columns.

In my approach, the genuine dataset is cleaned and preprocessed, a CTGAN model is trained on it, and a new "fake" dataset that isn't a copy but looks similar is created. Initially, everything is run in Jupyter notebooks so that I can troubleshoot and get results immediately.

3 Proposed Solution

Workflow

1. **Preprocessing:** involves loading the raw CSV file into Pandas, handling null values, and automatically recognizing continuous (numeric) and categorical (string, few unique values) columns.
2. **Model Training (CTGAN):** Use the processed dataset to train a CTGAN model. CTGAN creates fake rows after learning the data patterns. I start with the default settings and want to tinker later.
3. **Generating Data:** Store the trained model in a distinct `synthetic` folder and utilize it to generate synthetic data in CSV format.
4. **Evaluation:** Compare distributions — use bar charts for categorical data and histograms for numerical columns to compare the distributions of synthetic and actual data. Another statistical method for determining similarity is the Kolmogorov–Smirnov test.

4 Expected Outcome

After completing this project, I want to be able to:

- Make artificial replicas of any tabular dataset that is appropriate for machine learning research.
- Using charts and metrics, show that the synthetic data is statistically comparable to the real data.
- Create a workflow that is well-structured, reusable, and adaptable to new datasets.