Fraud Analytics Assignment 4 : Synthetic data generation using GAN

Team Members:
CS22BTECH11012 : Bolla Nehasree
CS22BTECH11043 : Nethi Keerthana
ES22BTECH11025 : N. Krishna Chaitanya

1. Objective

This project implements a Wasserstein GAN with Gradient Penalty (WGAN-GP) using PyTorch to generate synthetic tabular data that mimics the structure and distribution of a real dataset. The model is trained to produce data with similar statistical properties and inter-feature correlations.

2. Code Explanation

2.1 Importing Libraries

The code begins by importing essential libraries such as PyTorch, NumPy, Pandas, Matplotlib, and Seaborn for model building, data manipulation, and visualization.

2.2 Data Preparation

- Real dataset is loaded into a Pandas DataFrame.
- Normalization is applied for stable GAN training.
- The dataset is converted into a PyTorch TensorDataset and loaded in batches using DataLoader.

2.3 Model Architecture

Generator:

- Takes a noise vector z as input.
- Passes through multiple fully connected layers with LeakyReLU activations.
- Outputs synthetic samples with the same shape as real data.

Critic (Discriminator):

- Accepts real or generated data.
- Uses a similar multi-layer structure.
- Outputs a scalar Wasserstein score instead of probability.

2.4 Gradient Penalty

To enforce the Lipschitz constraint (a requirement for WGANs), the gradient penalty is computed between real and fake samples:

$$\text{GP} = \lambda \cdot (\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2$$

This term penalizes the critic when its gradient norm deviates from 1.

2.5 Training Procedure

- For each epoch:
  - Train the critic for n_{critic} steps per generator update.
  - Calculate WGAN loss: $D(\text{real}) - D(\text{fake})$.
  - Apply gradient penalty.
  - Update generator to maximize $D(\text{fake})$.
- Optimizer: Adam ($\beta_1 = 0.5, \beta_2 = 0.9$).

2.6 Generating Data

After training, noise vectors are passed through the generator to create synthetic samples, which are then evaluated against the real dataset.

3. Results and Evaluation

3.1 Feature Distribution

Feature-wise histograms or KDE plots(Kernel density estimation) show a high overlap between real and synthetic data distributions, indicating successful learning.

3.2 Correlation Structure

Correlation matrices for real and synthetic data are computed and compared by plotting heatmaps.

Analysis of Correlation Metrics : The comparison between the real and synthetic correlation matrices is summarized below:

- **Mean Squared Error (MSE):** 0.00427
  This low MSE value indicates that the synthetic correlation matrix is numerically very close to the real correlation matrix. Most pairwise feature relationships are being accurately captured by the WGAN-GP model.
- **Pearson Correlation Coefficient:** 0.98206
  A high Pearson correlation coefficient indicates a very strong linear relationship between the real and synthetic correlation matrices. This means that the model is able to preserve the direction and strength of relationships among features across datasets.
- **Structural Similarity Index (SSIM):** 0.98415
  SSIM values close to 1 indicate high structural similarity. A value of 0.984 confirms that the overall layout and texture of correlations in the synthetic data match very closely with the real data.

- **Mean of Correlation Matrices:**

  - Real: 0.1293
  - Synthetic: 0.1262

  The average correlation magnitude in both datasets is nearly the same, suggesting that the overall degree of feature interdependence is well preserved in the generated data.

Observation

The synthetic data generated using WGAN-GP exhibits high fidelity with respect to the real dataset, not only in terms of marginal distributions but also in capturing inter-feature relationships. The low MSE, high Pearson correlation, and high SSIM all support the conclusion that the synthetic correlation structure is very similar to the real one. These results highlight the effectiveness of WGAN-GP in generating structurally and statistically coherent tabular data.

4. Conclusion

The WGAN-GP model generates realistic synthetic data. Visual and quantitative evaluations confirm that the synthetic data matches the real data in terms of distribution and correlation structure.