**Ganblr++ Documentation**

**1. Overview**

**Ganblr++** is an advanced generative adversarial network tailored for synthesizing high-quality tabular data with mixed feature types. It introduces sophisticated mechanisms like Bayesian Gaussian Mixture-based discretization, truncation-based sampling for numerical features, and a Training-on-Synthetic-Testing-on-Real (TSTR) evaluation for model benchmarking.

Ganblr++ is designed for tasks requiring high-fidelity synthetic data generation, making it particularly suitable for sensitive domains such as healthcare, finance, and fraud detection. Its modular design allows it to handle complex datasets with numerical and categorical features seamlessly.

**2. Installation**

**Steps to Set Up:**

1. Fork the repository.
2. Clone the repository:

bash

cd Katabatic

1. Create a virtual environment:

bash

python -m venv venv

1. Activate the virtual environment:
   * On Windows:

bash

.\venv\Scripts\Activate

1. Install dependencies:

bash

pip install -r requirements.txt

**Key Dependencies:**

* **Scikit-learn:** Preprocessing, Gaussian Mixture Model, and evaluation.
* **Pandas:** Data manipulation and analysis.
* **Numpy:** Efficient numerical computations.
* **Scipy:** For truncated normal sampling.
* **Tqdm:** Progress monitoring during sampling.

**3. Architecture and Workflow**

**Main Components:**

1. **DMM Discretizer:**
   * Utilizes Bayesian Gaussian Mixture models for numerical feature discretization.
   * Encodes numerical features into ordinal representations, supporting smoother integration with GAN training.
2. **GANBLR:**
   * The core GAN model responsible for generating synthetic tabular data.
3. **Synthetic Sampling:**
   * Generates synthetic data by combining ordinal categorical and reconstructed numerical features.
4. **Evaluation:**
   * Implements TSTR evaluation to measure the utility of synthetic data in real-world machine learning tasks.

**Workflow:**

1. **Data Preprocessing:**
   * Numerical columns are discretized using the DMMDiscretizer.
   * Categorical data is encoded for GAN training compatibility.
2. **Training:**
   * The model is trained using adversarial loss on both numerical and categorical features.
3. **Synthetic Data Generation:**
   * Ordinal and numerical data are synthesized separately and combined for high-quality outputs.
4. **Evaluation:**
   * Accuracy-based benchmarks validate the synthetic data's effectiveness.

**4. Running the Model**

**Key Methods:**

* **fit():** Trains the GANBLR++ model on input data.
* **sample():** Generates synthetic data based on learned distributions.
* **evaluate():** Performs TSTR evaluation using logistic regression, random forests, or multi-layer perceptrons.

**Usage Example:**

python

from ganblrpp import GANBLRPP

# Initialize Ganblr++

model = GANBLRPP(numerical\_columns=[0, 2, 4], random\_state=42)

# Train model

X\_train, y\_train = <your\_data>, <your\_labels>

model.fit(X\_train, y\_train, k=1, batch\_size=64, epochs=20)

# Generate synthetic data

synthetic\_data = model.sample(size=1000)

print(synthetic\_data)

# Evaluate model performance

accuracy = model.evaluate(X\_test, y\_test, model='lr')

print("Evaluation Accuracy:", accuracy)

**5. Configuration**

**Configuration File Example (config.json):**

json

{

"numerical\_columns": [0, 2, 4],

"random\_state": 42,

"batch\_size": 64,

"epochs": 20,

"k": 1,

"warmup\_epochs": 1

}

**6. Example Workflow**

1. **Data Preparation:**
   * Identify numerical columns.
   * Prepare datasets for training and evaluation.
2. **Model Training:**
   * Fit the model with specified parameters:

python

model.fit(X\_train, y\_train, k=2, batch\_size=32, epochs=10)

1. **Synthetic Data Generation:**
   * Generate synthetic data:

synthetic\_data = model.sample(size=500)

1. **Evaluation:**
   * Evaluate synthetic data quality:

accuracy = model.evaluate(X\_test, y\_test, model='rf')

**7. Evaluation**

**Metrics:**

* **TSTR Accuracy:** Measures the performance of synthetic data when used to train machine learning models and tested on real data.
* **Categorical Matching:** Ensures generated categories match real-world distributions.

**Visualization:**

* Feature distributions for numerical and categorical variables.
* Heatmaps for correlation comparison.

**8. Use Cases**

**Applications:**

1. **Healthcare:** Generate synthetic patient records for safe data sharing.
2. **Finance:** Produce synthetic data for fraud detection model training.
3. **Retail:** Simulate customer behavior for predictive analytics.

**9. Troubleshooting**

**Common Issues:**

* **Training Errors:** Ensure numerical columns are accurately specified.
* **Poor Data Quality:** Validate input data preprocessing steps.
* **Evaluation Failures:** Confirm synthetic data dimensions align with test data.

**10. Contribution Guidelines**

To contribute:

1. Fork the repository:

bash

git fork https://github.com/DataBytes-Organisation/Katabatic.git

1. Create a feature branch:

bash

git checkout -b feature/ganblrpp

1. Submit a pull request for review.

**11. Research Context**

**Ganblr++** builds on the following concepts:

1. Bayesian Gaussian Mixture Models for numerical feature discretization.
2. GAN-based data generation for high-fidelity synthetic datasets.
3. Evaluation methodologies like TSTR for real-world performance benchmarking.

**12. References**

1. Goodfellow et al., 2014. "Generative Adversarial Networks."
2. Xu et al., 2019. "Modeling Tabular Data Using Conditional GAN."
3. Scikit-learn Documentation.

**Ganblr++ Documentation**

**1. Overview**

**Ganblr++** is an advanced generative adversarial network tailored for synthesizing high-quality tabular data with mixed feature types. It introduces sophisticated mechanisms like Bayesian Gaussian Mixture-based discretization, truncation-based sampling for numerical features, and a Training-on-Synthetic-Testing-on-Real (TSTR) evaluation for model benchmarking.

Ganblr++ is designed for tasks requiring high-fidelity synthetic data generation, making it particularly suitable for sensitive domains such as healthcare, finance, and fraud detection. Its modular design allows it to handle complex datasets with numerical and categorical features seamlessly.

**2. Installation**

**Steps to Set Up:**

1. Fork the repository.
2. Clone the repository:

bash

git clone <enter-your-forked-repo-link>

cd Katabatic

1. Create a virtual environment:

bash

python -m venv venv

1. Activate the virtual environment:
   * On Windows:

bash

.\venv\Scripts\Activate

1. Install dependencies:

bash

pip install -r requirements.txt

**Key Dependencies:**

* **Scikit-learn:** Preprocessing, Gaussian Mixture Model, and evaluation.
* **Pandas:** Data manipulation and analysis.
* **Numpy:** Efficient numerical computations.
* **Scipy:** For truncated normal sampling.
* **Tqdm:** Progress monitoring during sampling.

**3. Architecture and Workflow**

**Main Components:**

1. **DMM Discretizer:**
   * Utilizes Bayesian Gaussian Mixture models for numerical feature discretization.
   * Encodes numerical features into ordinal representations, supporting smoother integration with GAN training.
2. **GANBLR:**
   * The core GAN model responsible for generating synthetic tabular data.
3. **Synthetic Sampling:**
   * Generates synthetic data by combining ordinal categorical and reconstructed numerical features.
4. **Evaluation:**
   * Implements TSTR evaluation to measure the utility of synthetic data in real-world machine learning tasks.

**Workflow:**

1. **Data Preprocessing:**
   * Numerical columns are discretized using the DMMDiscretizer.
   * Categorical data is encoded for GAN training compatibility.
2. **Training:**
   * The model is trained using adversarial loss on both numerical and categorical features.
3. **Synthetic Data Generation:**
   * Ordinal and numerical data are synthesized separately and combined for high-quality outputs.
4. **Evaluation:**
   * Accuracy-based benchmarks validate the synthetic data's effectiveness.

**4. Running the Model**

**Key Methods:**

* **fit():** Trains the GANBLR++ model on input data.
* **sample():** Generates synthetic data based on learned distributions.
* **evaluate():** Performs TSTR evaluation using logistic regression, random forests, or multi-layer perceptrons.

**Usage Example:**

from ganblrpp import GANBLRPP

# Initialize Ganblr++

model = GANBLRPP(numerical\_columns=[0, 2, 4], random\_state=42)

# Train model

X\_train, y\_train = <your\_data>, <your\_labels>

model.fit(X\_train, y\_train, k=1, batch\_size=64, epochs=20)

# Generate synthetic data

synthetic\_data = model.sample(size=1000)

print(synthetic\_data)

# Evaluate model performance

accuracy = model.evaluate(X\_test, y\_test, model='lr')

print("Evaluation Accuracy:", accuracy)

**5. Configuration**

**Configuration File Example (config.json):**

json

{

"numerical\_columns": [0, 2, 4],

"random\_state": 42,

"batch\_size": 64,

"epochs": 20,

"k": 1,

"warmup\_epochs": 1

}

**6. Example Workflow**

1. **Data Preparation:**
   * Identify numerical columns.
   * Prepare datasets for training and evaluation.
2. **Model Training:**
   * Fit the model with specified parameters:

python

model.fit(X\_train, y\_train, k=2, batch\_size=32, epochs=10)

1. **Synthetic Data Generation:**
   * Generate synthetic data:

python

synthetic\_data = model.sample(size=500)

1. **Evaluation:**
   * Evaluate synthetic data quality:

accuracy = model.evaluate(X\_test, y\_test, model='rf')

**7. Evaluation**

**Metrics:**

* **TSTR Accuracy:** Measures the performance of synthetic data when used to train machine learning models and tested on real data.
* **Categorical Matching:** Ensures generated categories match real-world distributions.

**Visualization:**

* Feature distributions for numerical and categorical variables.
* Heatmaps for correlation comparison.

**8. Use Cases**

**Applications:**

1. **Healthcare:** Generate synthetic patient records for safe data sharing.
2. **Finance:** Produce synthetic data for fraud detection model training.
3. **Retail:** Simulate customer behavior for predictive analytics.

**9. Troubleshooting**

**Common Issues:**

* **Training Errors:** Ensure numerical columns are accurately specified.
* **Poor Data Quality:** Validate input data preprocessing steps.
* **Evaluation Failures:** Confirm synthetic data dimensions align with test data.

**10. Contribution Guidelines**

To contribute:

1. Fork the repository:

bash

git fork https://github.com/DataBytes-Organisation/Katabatic.git

1. Create a feature branch:

bash

git checkout -b feature/ganblrpp

1. Submit a pull request for review.

**11. Research Context**

**Ganblr++** builds on the following concepts:

1. Bayesian Gaussian Mixture Models for numerical feature discretization.
2. GAN-based data generation for high-fidelity synthetic datasets.
3. Evaluation methodologies like TSTR for real-world performance benchmarking.

**12. References**

1. Goodfellow et al., 2014. "Generative Adversarial Networks."
2. Xu et al., 2019. "Modeling Tabular Data Using Conditional GAN."
3. Scikit-learn Documentation.