Understanding and Running the TabDDPM Model(Prepared by Siddharth Yadav)

1. Introduction

TabDDPM (Tabular Denoising Diffusion Probabilistic Model) is a generative model developed by Yandex Research to generate high-quality synthetic tabular data. Unlike traditional GAN-based approaches, TabDDPM leverages the diffusion model framework, which progressively adds and then removes noise from data. It is particularly well-suited for datasets with complex distributions and mixed types.

2. How It Works

The core idea behind TabDDPM is based on denoising diffusion probabilistic models (DDPMs). These models work in two phases:

- Forward process: Noise is gradually added to the input data.
- Reverse process: A neural network is trained to recover the original data from the noisy version.

3. Components of TabDDPM

Key components of the TabDDPM model include:

- A simple MLP denoising neural network.
- A beta schedule that controls the amount of noise added at each timestep.
- Sampling methods to generate new data points by reversing the noise process.

4. Installation Instructions

To install and run TabDDPM on your machine or Google Colab:

1. Clone the repository:

```
git clone https://github.com/yandex-research/tab-ddpm.git
cd tab-ddpm
```

2. Install the dependencies:

```
pip install -r requirements.txt
```

3. Download dataset:

```
wget "https://www.dropbox.com/s/rpckvcs3vx7j605/data.tar?dl=1" -O data.tar
```

tar -xf data.tar

4. Train the model (example using Adult dataset):
python scripts/train_tab_ddpm.py --config configs/adult.yaml

5. Running on Google Colab

Google Colab provides a convenient environment to run TabDDPM without local setup:
- Enable GPU: Runtime > Change Runtime Type > GPU
- Use cells to clone, install, download data, and train the model.
- You can visualize results or modify YAML files for custom datasets.

6. Using Your Own Dataset

To use your own dataset:
- Upload your CSV file.
- Modify or create a YAML config in the configs/ folder.
- Adjust feature types, target columns, and paths accordingly.
- Run the training script using your custom config.

7. Model Evaluation and Output

After training, TabDDPM generates synthetic data that can be compared to real data using:
- Correlation matrices (heatmaps).
- Statistical measures like MSE and R^2.
- Downstream performance on ML models trained on synthetic data.

8. Conclusion

TabDDPM is a robust and state-of-the-art model for generating tabular data. Its use of diffusion processes allows it to outperform many GAN-based models in terms of quality and stability. The provided GitHub repo includes everything needed to get started.