Modeling Tabular Data using Conditional GAN
Lei Xu1, Maria Skoulardiou2, Alfredo Cuesta-Infante3, Kalyan Veeramachaneni1
1MIT LIDS, 2University of Cambridge, 3Universidad Rey Juan Carlos

Introduction

- GANs can generate realistic synthetic images. But in generating synthetic tabular data, state-of-the-art GAN-based models cannot outperform simple Bayesian network models as shown on Table 1.
- The challenges of generating synthetic data using GANs are the non-Gaussian multimodal distribution of continuous columns and imbalanced discrete columns.

We design CTGAN to address these challenges. CTGAN uses mode-specific normalization to effectively represent continuous values from different distribution; and uses a conditional generator and a training-by-sampling method to learn imbalanced discrete columns.

Mode-specific Normalization

For real data, we train classifiers or regressors on the synthetic data and evaluate prediction metrics on real test data.

Conditional Generator

For simulated data, we evaluate (1) the likelihood of test data on learned distribution as L_{test}, (2) and the likelihood of synthetic data on original data distribution as L_{syn}

Evaluation Metrics

- For simulated data, we evaluate (1) the likelihood of test data on learned distribution as L_{test}, (2) and the likelihood of synthetic data on original data distribution as L_{syn}
- For real data, we train classifiers or regressors on the synthetic data and evaluate prediction metrics on real test data.

Conclusion

In this paper we attempt to find a flexible and robust model to learn the distribution of columns with complicated distributions. We observe that none of the existing deep generative models can outperform Bayesian networks which discretize continuous values and learn greedily. We show that our model can learn better distributions than Bayesian networks. Mode-specific normalization can convert continuous values of arbitrary range and complicated distribution into a bounded vector representation suitable for neural networks. And our conditional generator and training-by-sampling can overcome the imbalanced training data issue. As future work, we would derive a theoretical justification on why GANs can work on a distribution with both discrete and continuous data.