Learning and Sampling from the Causal Graph of Multi Categorical Distributions Via the Generative Adversarial Network

Qiang Fu, Computer Science
Mentor: Huan Liu, Faculty Mentor, Raha Moraffah, Graduate Mentor
Arizona State University

Introduction

Generating fake data is a common problem for data scientists. A GAN is often used for such a task and have achieved great results for many tasks. Our problem is to generate multi categorical data. An example of it is fake medical records for data research. (Choi et al., 2017). This kind of data is usually discrete, and GAN cannot be directly trained on them. A previous research attempted to solve this problem. (Ramiro et.al., 2017) Our research builds upon this previous research and attempts to improve the quality of generated data.

Methods

Our research is based on a model MC-WGAN-GP proposed by a previous research (Ramiro et.al., 2017). In that model, the researcher added one dense layer network with Gumbel-Softmax function for each feature existed in the input dataset. A continuous sample is generated from the generator, then separated and fed into these individual networks. Finally, the output from these network are concatenated to form the final data. This method avoided the discrete problem and made it possible for GAN to train on multi categorical data.

In our research, we wanted to learn the casual relationships between variables and use the information to improve generated data quality. We used an adjacency matrix to represent the relationships between the variables and try to learn them to better learn the real data distribution. The matrix information is then used to generate the fake data D*sample*. The following functions are the updated loss functions in our model.

$$\mathcal{L}_D = \mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] - \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)] + \lambda \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x}) - 1\|)^2],$$

$$\mathcal{L}_G = -\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})] = -\mathop{\mathbb{E}}_{z\sim\mathbb{P}_z}[D(G(z;\theta,A))]$$

$$\text{s.t.} \quad tr[(I + \beta A \odot A)^n] - n = 0$$

Methods(cont.)

To examine the effectiveness of our model, we trained both our model and MC-WGAN-GP with 4 synthetic datasets (Ramiro et.al., 2017) as well as a real-life dataset from US census 2010, then used 3 metrics to measure the data quality generated by the trained model. Datasets are divided into D*train* and D*test* with 90%-10% ratio for testing purposes.
The first two methods can be found in (Choi et al., 2017).
The easiest metric is to calculate the frequency of ones for both D*sample* and D*test*.
The first metric does not take account into dependencies between variables, in the second metric, one logistic regression model for each feature is trained with D*train* and D*sample*, then we calculate each model's f1-score with D*test*.
The third metric is proposed in (Ramiro et.al., 2017). It modifies the second metric to predict one categorical variable at a time and replaces f-1 score by accuracy score to avoid a few problems existed in the second metric.
At last, we calculate the mean squared error(MSE) for each metric to get a numerical value for analyze.

Results

	Dataset	MSE1	MSE2	MSE3
Original	Fixed 2	0.000728	0.000067	0.000861
our method	Fixed 2	**0.000557**	**0**	**0.000696**
Original	Fixed 10	0.000365	0.040832	0.001678
our method	Fixed 10	**0.000362**	**0.027288**	**0.001003**
Original	Mix Small	0.000601	0.01715	**0.000405**
our method	Mix Small	**0.000444**	**0.013605**	0.000832
Original	Mix Big	0.000585	0.012216	0.00141
our method	Mix Big	**0.000398**	**0.009462**	**0.000967**
Original	US Census	0.000080	**0.129054**	0.000330
our method	US Census	**0.000029**	0.129732	**0.000327**

Table 1. MSE results from synthetic datasets, lower values are highlighted

Results(cont.)

In table 1 we gathered the measured results of our method and MC-WGAN-GP, the original model that our model is based on. The error value is better when lower. Our method outperforms the original method for almost every dataset and measurement.

Discussion

- Most of the results from synthetic datasets shows improvement of data quality.
- In two cases, our model has performed worse than the original method. It may due to variance in network initialization. More work is needed to find out the flaws in our method and training process.
- Our model has shown significant improvement with the first measuring method in all datasets.
- In the real-life dataset, the data shows significant improvement in the first measure but not in the other measures.

References

Camino, Ramiro, et al. "Generating Multi-Categorical Samples with Generative Adversarial Networks." ArXiv:1807.01202 [Cs, Stat], July 2018. arXiv.org, http://arxiv.org/abs/1807.01202.

Choi, Edward, Biswal, Siddharth, Malin, Bradley, Duke, Jon, Stewart, Walter F., and Sun, Jimeng. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. arXiv:1703.06490 [cs], March 2017. URL http://arxiv.org/abs/1703.06490. arXiv: 1703.06490.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative Adversarial Nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

FURI

Ira A. Fulton Schools of Engineering
Arizona State University