Variational AutoEncoder (VAE) for synthetic insurance data

Insurance Data Science Conference (IDSC) 2024

Charlotte Jamotton[1], Donatien Hainaut

Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA)
Université catholique de Louvain (UCL), Belgium

Stockholm, 17-18 June 2024

[1] charlotte.jamotton@uclouvain.be

1. Objective & problematic

2. Pre-processing

3. Architecture modifications

4. Synthetic policies generation

Objective & problematic

Objective: use Variational AutoEncoders (VAEs) to reduce the input dimension and to generate new synthetic policies.

Challenge: VAEs were initially designed (Kingma & Welling, 2013) to model continuous data.

Problematic: how to adapt the VAE architecture to insurance datasets containing categorical (ordinal, nominal) and continuous variables with a variety of marginal distributions (multi-modal, long-tail, . . .)?

Variational AutoEncoder architecture

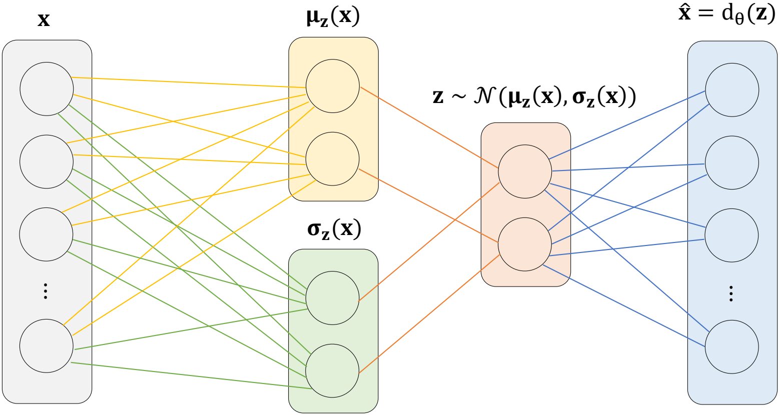
$\mathbf{x}$
$\boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x})$
$\boldsymbol{\sigma}_{\mathbf{z}}(\mathbf{x})$
$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{z}}(\mathbf{x}), \boldsymbol{\sigma}_{\mathbf{z}}(\mathbf{x}))$
$\hat{\mathbf{x}} = \mathrm{d}_{\theta}(\mathbf{z})$

Figure 1: Architecture of VAE with Gaussian prior.

Categorical data

How to deal with categorical data?

- we opted for the use of **one-hot encoding**

Policy	Vehicle color
1	Black
2	Black
3	Red
⋮	⋮
N	Gray

→

Policy	Black	Red	Gray
1	1	0	0
2	1	0	0
3	0	1	0
⋮	⋮	⋮	⋮
N	0	0	1

Table 1: Example of one-hot encoding with categorical variable `color` $\in \{\mathit{Black}, \mathit{Red}, \mathit{Gray}\}$

- entity embeddings (Guo & Berkhahn, 2016; Delong & Kozak, 2023) addresses the lack of semantic significance
  - takes the detour of an ordinal or one-hot encoding
  - first layer of VAE similar to an embedding layer
- vector quantization via codebooks (Van Den Oord et al., 2017) focuses on the scalability issue and generalizes well to unseen categories
  - additional complexity

Continuous data

How to deal with continuous data?

- challenging with multi-modal non-Gaussian distributions and the mode-seeking behavior of the KL divergence used to train the model:

$$-D_{KL}\left(q_{\phi}\left(\boldsymbol{z}|\boldsymbol{x}\right)||p_{\theta}(\boldsymbol{z})\right)+\ln p_{\theta}\left(\boldsymbol{z}|\boldsymbol{x}\right)$$

  - collapse of the posterior towards a single mode
  - biased generation of samples towards this (↑) specific mode

Solutions?

- mode-specific normalization using Gaussian mixture models (Bishop & Nasrabadi, 2006)
  - increases the input dimension by the nb of Gaussians used to *approximate* the variable distribution
- we propose a **quantile transformation**

Quantile transformation

Quantile transformer $\Phi^{-1}(F(X))$

If X is a random variable with a continuous CDF F then $U = F(X)$ is uniformly distributed on $[0,1]$. If $U \sim \mathcal{U}(0,1)$, then $\Phi^{-1}(U) = \Phi^{-1}(F(X))$ has distribution Φ.

In practise,

- applied on each feature X independently
- CDF of X, $F(X) = P(X \leq x)$, estimated using a reference set of quantiles and an estimation of the data percentiles and CDF
- apply $F^{-1}(X)$ to map the feature values to $\mathcal{U}(0,1)$ using the estimated percentiles
- use Φ^{-1} to map u to the desired output distribution $\Phi = \mathcal{N}(0,1)$

Quantile transformation

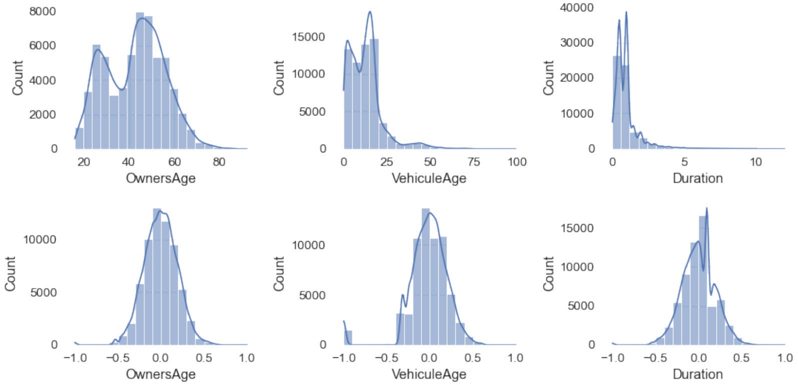

Figure 2: Continuous variables before and after quantile transformation.

Impact of $\Phi^{-1}(F(X))$ on $\hat{X}$

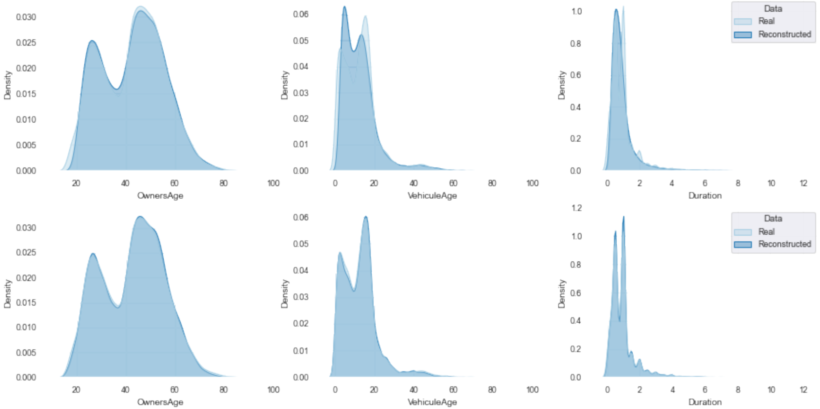

Figure 3: Reconstructed continuous variables, with(out) the quantile transformation pre-processing step.

Heterogeneous loss and activation functions

- softmax loss for the categorical variables with C_k modalities

$$\ln p\left(\mathrm{x}|\mathrm{z}\right) = \sum_{k=1}^{\#cat} \sum_{j=1}^{C_k} x_j \ln y_j\left(\boldsymbol{z}\right)$$

where $y_j\left(\boldsymbol{z}\right)$ are the outputs of the decoder network:

$$y\left(\boldsymbol{z}\right) = \mathtt{softmax}(\boldsymbol{W}_2 \boldsymbol{h}_x + \boldsymbol{b}_2)$$

- mean square error for the continuous variables

$$\ln p\left(\mathrm{x}|\mathrm{z}\right) = \sum_{j=1}^{\#cont} \|x_j - y_j\left(\boldsymbol{z}\right)\|^2$$

where $y_j\left(\boldsymbol{z}\right)$ are the outputs of the decoder network:

$$y\left(\boldsymbol{z}\right) = \mathtt{tanh}(\boldsymbol{W}_2 \boldsymbol{h}_x + \boldsymbol{b}_2)$$

Architecture

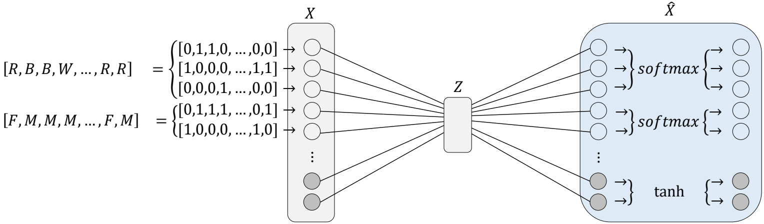
X
$[R, B, B, W, \ldots, R, R] = \begin{cases} [0,1,1,0,\ldots,0,0] \\ [1,0,0,0,\ldots,1,1] \\ [0,0,0,1,\ldots,0,0] \end{cases}$
$[F, M, M, M, \ldots, F, M] = \begin{cases} [0,1,1,1,\ldots,0,1] \\ [1,0,0,0,\ldots,1,0] \end{cases}$
Z
$\hat{X}$
$softmax$
$softmax$
tanh

Figure 4: Illustration of the slicing technique applied on the decoder output of the Variational AutoEncoder with two categorical variables `Colour` and `Gender` whose categories Red (R), Black (B), White (W), and Female (F), Male (M), are dummy encoded. The darker nodes represent continuous variables.

Synthetic policies generation

How to generate synthetic policies?

- sample from $p(\boldsymbol{z}) \sim \mathcal{N}(0, I)$
- apply inverse $\Phi^{-1}\left(F\left(Z^{(true)}\right)\right)$
- $d_\theta\left(z^{new}\right)$

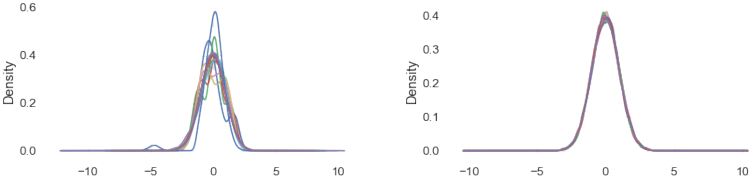

Figure 5: Densities of the $d = 15$ latent variables, before and after quantile transformation.

Impact of $\Phi^{-1}(F(X))$ on $\hat{X}^{synthetic}$

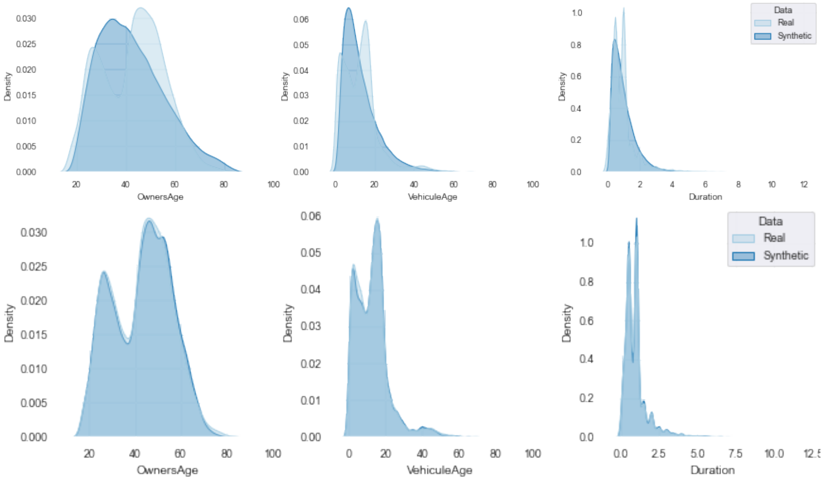

Figure 6: Synthetic continuous variables, with(out) the quantile transformation pre-processing step.

Synthetic data quality

Can the synthetic insurance portfolio replace the original portfolio in actuarial analyses?

$$D_{Poisson} = 2\sum_{i=1}^{n} N_i \left(\frac{\nu_i}{N_i}\hat{\lambda}_i - \left(\log \frac{\hat{\lambda}_i \nu_i}{N_i} + 1 \right) \mathit{1}_{\{N_i \geq 1\}} \right)$$

GLM trained on the original/synthetic/reconstructed insurance portfolio and used to predict the original policies claim frequencies:

	Poisson deviance	%
Original data	5 691.01	
Synthetic data	5 961.84	4.76%
Reconstructed data	5 713.09	0.39%

Check for duplicates:

- 2.55% of the synthetic policies are duplicate
- 1.82% of the synthetic policies (duplicates excluded) are copies of the original policies

Thank you!

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (No. 4). Springer.

Delong, Ł., & Kozak, A. (2023). The use of autoencoders for training neural networks with mixed categorical and numerical features. *ASTIN Bulletin: The Journal of the IAA*, *53*(2), 213–232.

Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*.

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, *30*.