ZEUS: Zero-shot Embeddings for Unsupervised Separation of Tabular Data

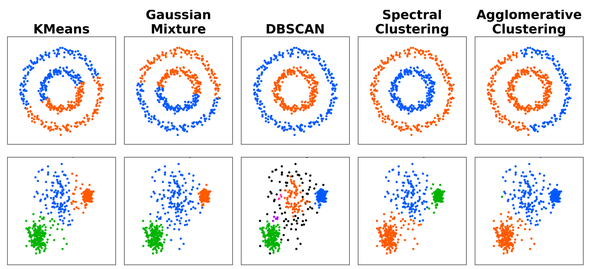
JAGIELLONIAN UNIVERSITY IN KRAKÓW

Patryk Marszałek, Tomasz Kuśmierczyk, Witold Wydmański, Jacek Tabor, Marek Śmieja

group of machine gmum learning research

Challenges in clustering tabular data

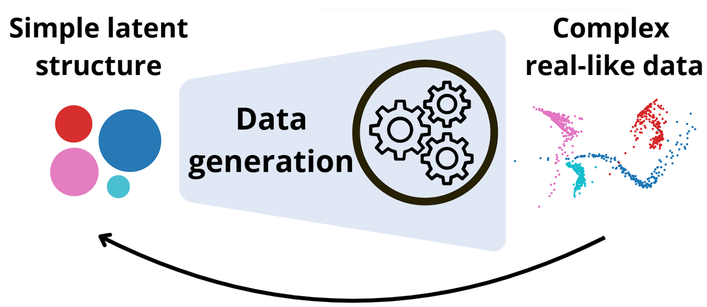
- How to characterize clusters?



- Trade-off: simple clustering vs time-consuming deep learning models
- How to set hyperparameters for unsupervised clustering?

Main idea

"What I cannot create, I do not understand."



Bayesian learning by pre-training transformer on synthetic data

Prior-data fitted network

$$\theta^* = \arg\min_\theta \mathbb{E}_{p(\mathcal{D})}\big[\mathcal{L}(\mathbf{y},\, f_\theta(\mathbf{x}))\big] \approx \arg\min_\theta \frac{1}{N_\mathcal{D}} \sum_{\mathcal{D}\sim p(\mathcal{D})} \mathcal{L}(\mathbf{y},\, f_\theta(\mathbf{x}))$$

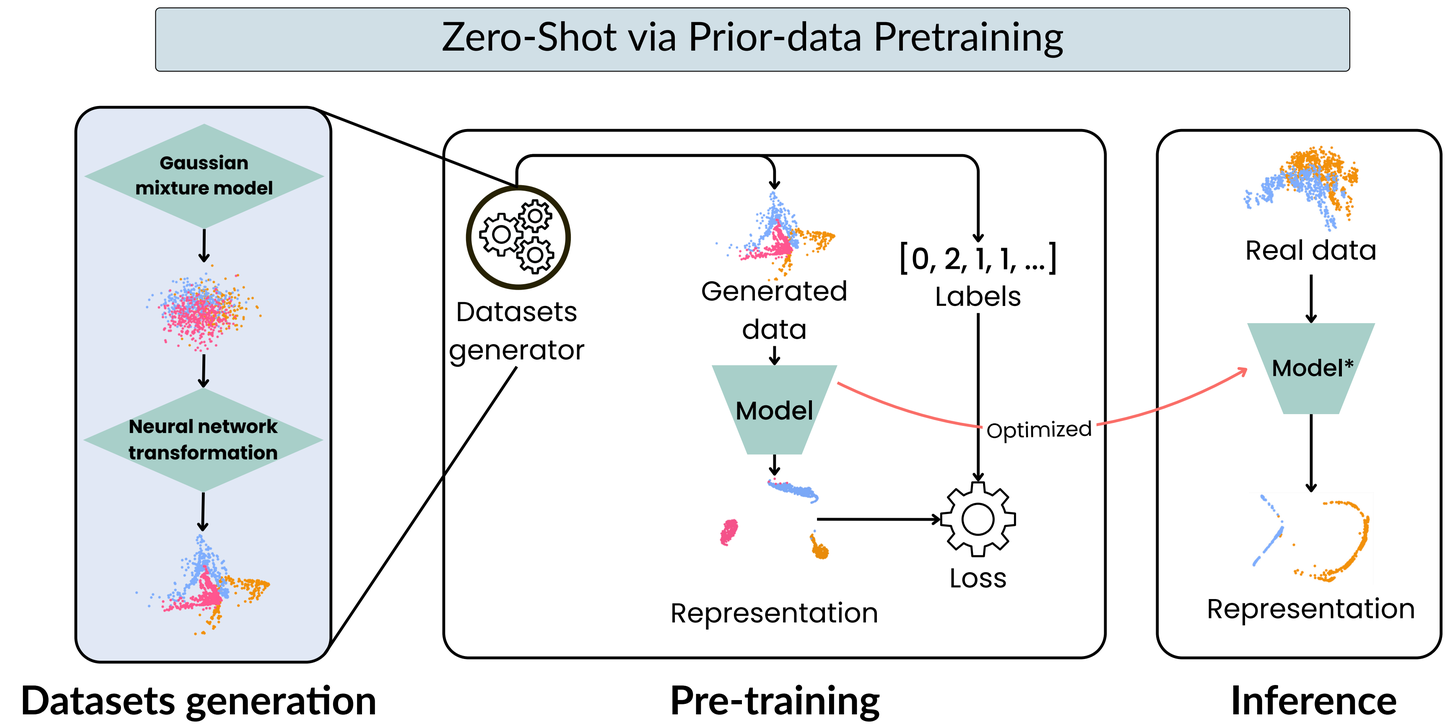
Clustering loss for ZEUS pretraining: $\mathcal{L} = \mathcal{L}_{prob} + \mathcal{L}_{cp} + \mathcal{L}_{sep}$

- Negative log-likelihood: $\mathcal{L}_{prob} = -\sum_i \log p_{y_i}(x_i)$
- Mean square error: $\mathcal{L}_{cp} = \sum_k \sum_{i:y_i=k} \|f_\theta(\mathbf{x})_i - \hat{c}_k\|^2$
- Separation term: $\mathcal{L}_{sep} = -\sum_{k=1}^K \sum_{j=k+1}^K \min(\|\hat{c}_k - \hat{c}_j\|^2, T)$

where

$$p_k(x_i) = \frac{\hat{\pi}_k \exp(-\|f_\theta(\mathbf{x})_i - \hat{c}_k\|^2)}{\sum_{j=1}^K \hat{\pi}_j \exp(-\|f_\theta(\mathbf{x})_j - \hat{c}_k\|^2)}$$

Method



Experimental results

Benchmarking ZEUS on real-world datasets, synthetic Gaussian mixtures, and ResNet-transformed Gaussian mixtures.

Clustering quality of ZEUS, assessed via the Adjusted Rand Index (ARI).

	KM	GMM	AE -KM	AE -GMM	DEC	IDEC	IDC	G-CEALS	Tab PFN	SCARF	ZEUS
Real	55.54	48.49	51.43	53.56	55.93	54.57	52.28	40.37	31.32	26.95	**57.43**
Syn. Gauss.	**89.90**	76.93	81.26	81.40	89.35	82.57	66.43	62.84	55.97	8.32	89.03
Syn. Transf.	75.04	75.88	60.45	71.29	79.94	61.26	66.78	49.17	15.66	2.48	**86.33**

Soft clustering quality of ZEUS vs. baselines, measured by Brier score.

	KM	GMM	AE-KM	AE-GMM	DEC	IDEC	G-CEALS	ZEUS
Real	0.44	0.48	0.46	0.47	0.39	**0.37**	0.47	0.38
Syn. Gauss.	**0.10**	0.31	0.21	0.25	0.39	0.23	0.39	0.13
Syn. Transf.	0.28	0.26	0.45	0.31	0.46	0.39	0.50	**0.18**

Conclusion

- ZEUS does not require model retraining on individual datasets.
- It builds a representation for clustering in a single forward pass.
- It outperforms all alternatives or attains a position within the top 3.