Gradient Importance Learning for Incomplete Observations

Qitong Gao[*] Dong Wang[*] Joshua D. Amason[*] Siyang Yuan[*] Chenyang Tao[*]

Ricardo Henao[*] Majda Hadziahmetovic[*] Lawrence Carin[*,†] Miroslav Pajic[*]

[*]Duke University

[†]King Abdullah University of Science and Technology

Contact: qitong.gao@duke.edu, miroslav.pajic@duke.edu

Motivation

- Learning from **incomplete datasets** is common and inevitable
  - There are three major causes for missing data: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR)

	MCAR	MAR	MNAR
Observable Variables	Independent	Dependent	Dependent
Unobservable Variables	Independent	Independent	Dependent

- Existing works mostly rely on imputation methods to fill in missing data before inference
  - GAIN (ICML'18) and MIWAE (ICML'19) use generative models (VAEs & GANs) to estimate missing entries
  - BRITS (NeurIPS'18) imposes both imputation and prediction losses to be jointly optimized during learning

Motivation

- However, imputations may not be necessary for downstream analysis; Sometimes the missingness **speaks for itself!**

- Both patients A and B are admitted to ICU for infections. From the records:
  - Doctors issued more PCR tests (specific to viral infections) to A,
  - More blood culture (specific to bacterial infections) to B,
  - By just looking at missingness patterns: A had undergone viral infections while B had bacterial infections.

Patient A

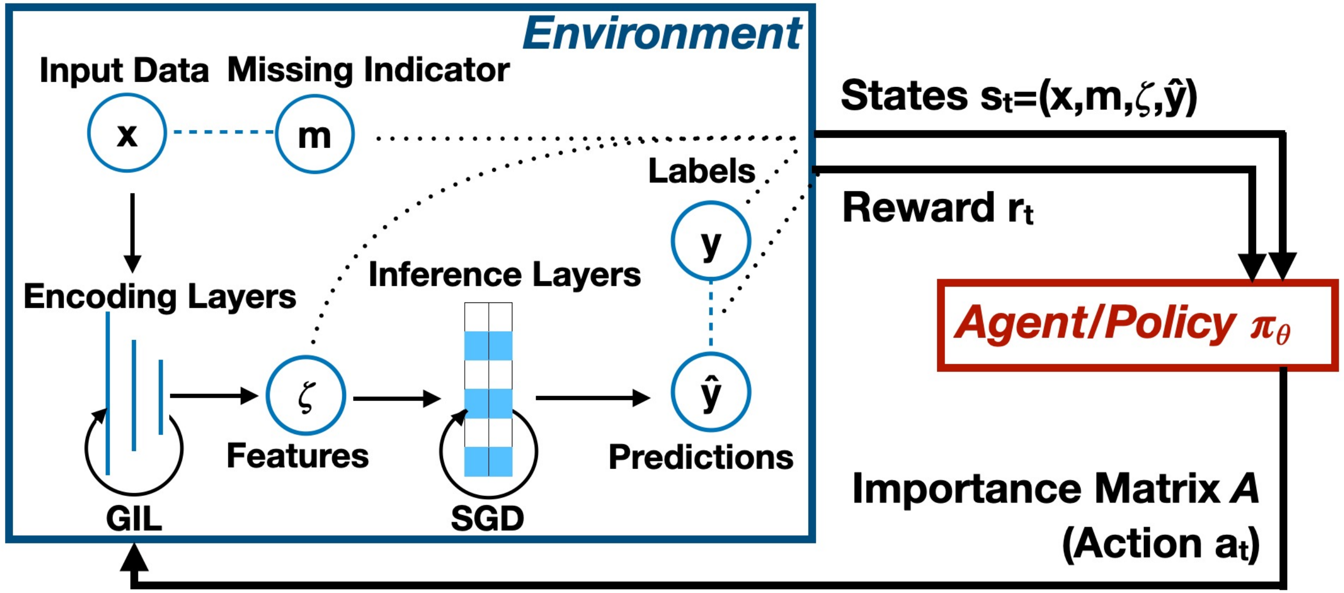
	Blood Culture	PCR Panel
02/11	☑	☑
02/12	N/A	☑
02/13	☑	☑
02/14	N/A	☑
02/15	N/A	☑
02/16	N/A	☑
02/17	N/A	☑

Patient B

	Blood Culture	PCR Panel
02/11	☑	☑
02/12	☑	N/A
02/13	☑	N/A
02/14	☑	N/A
02/15	☑	N/A
02/16	☑	N/A
02/17	☑	N/A

Gradient Importance Learning (GIL)

- Our method, GIL, trains models to
  - better capture the information underlying missingness, and
  - make accurate predictions over incomplete data,
  - **without** using any imputation losses/algorithms.



- **Main Idea**: Reinforcement learning (RL) is used to discover underlying information and infuse them into the gradients for training of downstream prediction models.

Gradient Importance Learning (GIL)

- Consider multi-layered perceptrons (MLPs), or LSTMs followed by dense layers as the prediction model
  - The 1st hidden layer (or the LSTM layer) – *encoding layer*,
  - The ones that follows – *inference layers*.

- The gradients used to train the encoding layer can be formulated as **outer products** between the layer inputs and gradients propagated from deeper layers,

$$\frac{\partial L}{\partial W_{enc}} = \Delta \cdot x^{\intercal};$$

  L is the loss function, W_{enc} the weights of encoding layers, Δ the gradients propagated from inference layers and x is the inputs.

- The i^{th} column in $\frac{\partial L}{\partial W_{enc}}$ is weighted by the i^{th} element of x, **which could be missing** => *directly propagating such gradients may not be meaningful*.

Gradient Importance Learning (GIL)

- Hence, the gradients can be re-weighted by an importance matrix $\boldsymbol{A}$ **elementwise** during back-propagation, i.e.,

$$W_{enc} \leftarrow W_{enc} - \alpha \cdot (\Delta \cdot x^{\top}) \odot \boldsymbol{A}.$$

- As $\boldsymbol{A}$ is introduced into the back-propagation after the original SGD gradients are calculated, its elements cannot be figured using the same back-propagation solver.

- **Idea**: Use RL to solve $\boldsymbol{A}$ by
  - First, formulating the back-propagation process for training the prediction model as an MDP,
  - Then, leverage actor-critic methods to generate an RL policy that can adapt elements of $\boldsymbol{A}$ in response to the changes of x, W_{enc} during training.

Experimental Results

GIL (imputation-free) vs existing 2-step (imputation THEN prediction) methods

Duke PRATT SCHOOL *of* ENGINEERING

- Comparison two real-world, large-scaled healthcare datasets, and MNIST digits
- GIL achieved **the best performance** on MIMIC-III dataset for septic shock prediction
  - A survey (Fleuren et al., 2020) over septic shock predictions reported that *the highest AUC* of existing domain-expert-designed models is around 96%,
  - However, GIL **does not require any domain expertise** for modeling and learning.

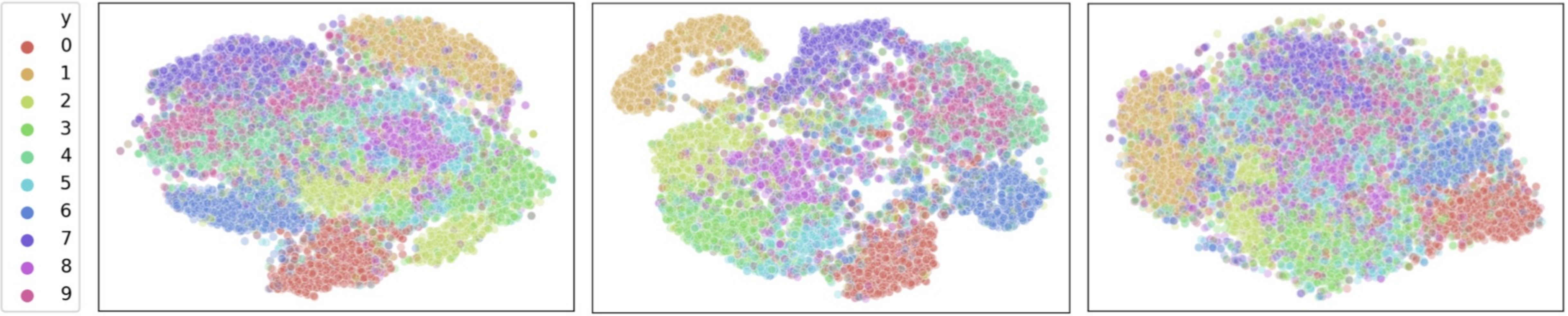
Table 1: Accuracy and AUC obtained from the MIMIC-III dataset.

		GIL	-D	-H	GAIN	MIWAE	GP-VAE	BRITS	MICE	Mean	CF	kNN	MF	EM
Var-l.	Acc.	**93.32**	93.09	89.17	90.32	88.71	-	-	92.17	88.02	87.32	84.79	75.81	68.20
	AUC	**96.10**	**96.79**	92.96	<u>95.57</u>	94.28	-	-	<u>95.97</u>	92.56	91.78	91.86	81.73	75.23
Fix-l.	Acc.	**91.47**	91.01	88.25	88.48	86.18	76.50	80.24	90.09	86.41	86.87	85.48	78.11	70.51
	AUC	**95.29**	**95.57**	92.99	91.94	93.10	81.47	92.13	94.02	91.69	91.98	92.38	84.54	79.97

Experimental Results

Duke
PRATT SCHOOL of ENGINEERING

GIL learns more expressive feature representations (outputs from the encoding layer)!



(a) GIL (b) MIWAE (c) GAIN

t-SNE visualizations of the feature space learned by GIL, MIWAE (ICML’19) and GAIN (ICML’18) on the **MNIST dataset with 90% missing rate**.

Thank you

Code available at https://github.com/gaoqitong/gradient-importance-learning