Variational Inference for Discriminative Learning with Generative Modeling of Feature Incompletion

Kohei Miyaguchi, Takayuki Katsuki, Akira Koseki, Toshiya Iwamori (IBM Research)

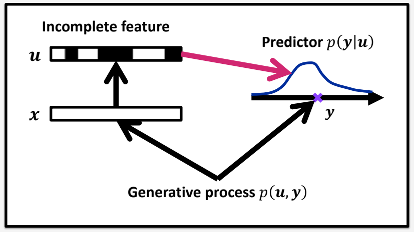
miyaguchi@ibm.com, kats@jp.ibm.com, akoseki@jp.ibm.com, iwamori@jp.ibm.com

Message

Problem of **discriminative learning** with **generative modeling** is solved with black-box variational inference (BBVI).

Summary

Task: Prediction with incomplete features



Example: Survival analysis

- x = patient's health state
- u = partially-missing electronic health records
- y = days to onset

Existing approaches

- *Generative approach*: Learn the generative process directly.
- *Discriminative approach*: Learn the predictor directly.
- *Hybrid approach*: Learn the predictor w/ generative modeling.

Approach	Objective	Missing-value model	Applicable models
Generative	Generative	Yes	Any generative models
Discriminative	Predictive	No	Any discriminative models
Hybrid	Predictive	Yes	Predictively-trainable generative models

Problem & our solution

- *Problem*: Poor applicability of hybrid approach, i.e., absence of model-independent algorithm.
- *Our solution*: Extension of black-box variational inference (BBVI).
- *Result*: Applicable to neural-network-based models (e.g., VAE).

Key challenge

Background: Objective function of hybrid approach

$$\mathcal{L}(\theta) := \ln p(\boldsymbol{y}|\boldsymbol{u},\theta) \quad \text{(Predictive objective)}$$

$$= \underbrace{\ln p(\boldsymbol{y},\boldsymbol{u}|\theta)}_{\text{Evidence lower bound (ELBO) maximization}} - \underbrace{\ln p(\boldsymbol{u}|\theta)}_{\text{Evidence upper bound (EUBO) minimization}} \quad \text{(Generative modeling)}$$

Challenge: Existing EUBOs are either biased or unstable†

	Unbiased	Stable
χ-bound [Dieng+ 2017]	No	Yes
Reversed KL bound [Ji & Shen 2019]	No	Yes
Tangent χ-bound [Kuleshov & Ermon 2017]	Yes	No

†: Unstable ⇔ unbounded gradient of variational lower bound:

$$VLB(\theta,\zeta) = ELBO(\theta,\zeta_1) - \left\{ f(\boldsymbol{u};\zeta_2) + \frac{1}{2} w^2(\theta,\zeta) + C \right\}$$

where

$$w(\theta,\zeta) := \frac{p(\boldsymbol{u},\boldsymbol{z}|\theta)}{\exp f(\boldsymbol{u};\zeta_2)\, q(\boldsymbol{z}|\boldsymbol{u},\zeta_3)}$$

Proposed stabilization technique

1. Partially transform the parameter (θ and ζ_2):

$$p(\boldsymbol{u},\boldsymbol{z}|\theta) \mapsto p(\boldsymbol{u},\boldsymbol{z}|\theta') := \frac{G(w(\theta,\zeta))}{Z(\theta,\zeta)} p(\boldsymbol{u},\boldsymbol{z}|\theta)$$

$$f(\boldsymbol{u};\zeta_2) \mapsto f(\boldsymbol{u};\zeta_2') := f(\boldsymbol{u};\zeta_2) - \ln Z(\theta,\zeta)$$

2. Divergent term is stabilized (if G(w) is appropriate)

$$w(\theta',(\zeta_1,\zeta_2',\zeta_3)) = \frac{p(\boldsymbol{u},\boldsymbol{z}|\theta')}{\exp f(\boldsymbol{u},\zeta_2')\, q(\boldsymbol{z}|\boldsymbol{u},\zeta_3)} = w(\theta,\zeta) G(w(\theta,\zeta))$$

Theoretical justification

Put $T(\theta,\zeta) := (\theta',(\zeta_1,\zeta_2',\zeta_3))$. Then,

1. The gradient becomes bounded:

$$\|\nabla(VLB \circ T)(\theta,\zeta)\| \le 9K$$

where

$$K := \|\nabla \ln p(\boldsymbol{y},\boldsymbol{u}|\theta)\| \vee \|\nabla \ln p(\boldsymbol{u}|\theta)\| \vee \|\nabla f(\boldsymbol{u},\zeta)\| \vee \|\nabla \ln q(\boldsymbol{z}|\boldsymbol{u},\zeta)\| \vee \|\nabla \ln q(\boldsymbol{z}'|\boldsymbol{y},\boldsymbol{u},\zeta)\|$$

2. "Effective" parameters are preserved and invariant:

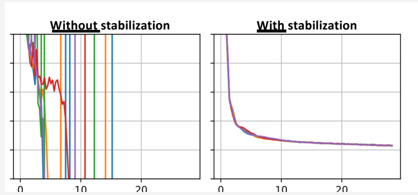
$$\Theta_{\mathrm{eff}}(T(\Omega)) = \Theta_{\mathrm{eff}}(\Omega),$$

where

$$\Theta_{\mathrm{eff}}(\Omega) := \{(\theta,\zeta) \in \Omega : \mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{z}}[VLB(\theta,\zeta)]\}.$$

Experimental results

Comparison of stability (training objective)



Comparison of predictive performance w/ VAE

		AQ-CO	AQ-NMHC	AQ-NOx	Boston	Diabetes	YearPred	Total
Discriminative	CVAE	**0.41** (0.02)	5.19 (0.14)	5.33 (0.02)	3.01 (0.18)	5.46 (0.03)	3.49 (0.06)	3.82 (0.04)
Hybrid (MNAR/MCAR)	DVAE	0.45 (0.03)	**5.16** (0.10)	**5.26** (0.02)	2.91 (0.23)	5.44 (0.08)	3.36 (0.04)	**3.76** (0.05)
	DVAE*	0.44 (0.03)	**5.16** (0.11)	5.27 (0.03)	2.92 (0.21)	**5.42** (0.07)	**3.35** (0.02)	**3.76** (0.04)
	Simple	0.49 (0.03)	5.47 (0.11)	5.40 (0.02)	3.13 (0.18)	5.47 (0.04)	3.62 (0.04)	3.93 (0.04)
	MICE	0.46 (0.06)	5.34 (0.23)	5.38 (0.02)	2.94 (0.10)	5.46 (0.03)	3.61 (0.08)	3.87 (0.05)
Generative (MNAR/MCAR)	VAE	0.47 (0.01)	5.21 (0.08)	5.49 (0.06)	2.95 (0.23)	5.48 (0.10)	3.59 (0.02)	3.87 (0.04)
	VAE*	0.46 (0.02)	5.20 (0.09)	5.47 (0.05)	**2.81** (0.20)	5.46 (0.05)	3.58 (0.03)	3.83 (0.04)

Negative test log-likelihood (std.)