Orchestrating a brighter world NEC

IJCAI 2020, Yokohama, Japan

Measuring the Discrepancy between Conditional Distributions: Methods, Properties and Applications

Shujian Yu[1], Ammar Shaker[1], Francesco Alesiani[1], Jose C. Principe[2]

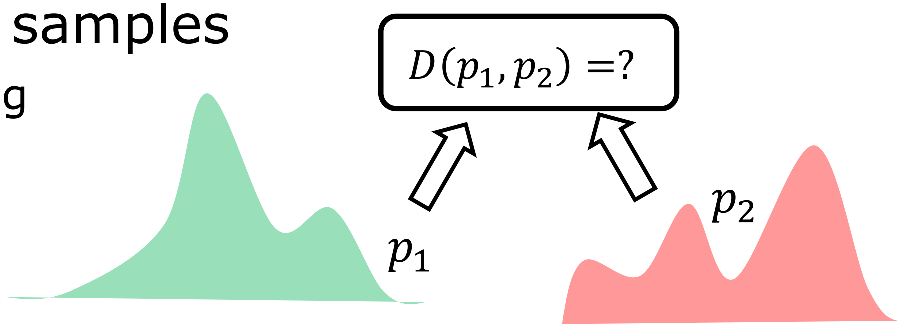
[1]NEC Laboratories Europe GmbH

[2]University of Florida

Contact: Shujian.Yu@neclab.eu

Motivation

Compare distributions with only samples



- Transfer Learning / Multi-Task Learning
- Deep Generative Models
- …

Divergence and Conditional Divergence in Machine Learning

- Kullback–Leibler (KL) divergence
  - $D_{\mathrm{KL}}(p_1(\boldsymbol{x})||p_2(\boldsymbol{x})) = \int p_1(\boldsymbol{x}) \log \frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} d\boldsymbol{x}$
- Maximum Mean Discrepancy (MMD)
  - $D_{\mathrm{MMD}}(p_1||p_2) = \left\| \mathbb{E}_{x\sim p_1}[\varphi(x)] - \mathbb{E}_{x'\sim p_2}[\varphi(x')] \right\|_{\mathcal{H}}, \varphi: \mathcal{X} \to \mathcal{H}$
- Wasserstein distance or optimal transport
  - $W_2^2(p_1||p_2) = \inf_{P \in \Pi[\mathrm{p}_1, \mathrm{p}_2]} \int \|x_2 - x_1\|^2 dP(x_1, x_2)$

But …

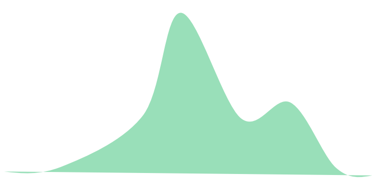
Bregman-Correntropy (Conditional) Divergence

Our Target

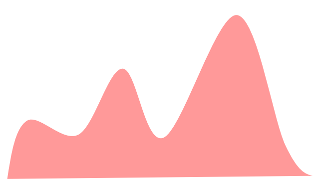
- A novel sample estimator to the divergence $D(p_1(\boldsymbol{x})||p_2(\boldsymbol{x}))$, $\boldsymbol{x} \in \mathbb{R}^p$
- Extension to conditional divergence $D(p_1(\boldsymbol{y}|\boldsymbol{x})||p_2(\boldsymbol{y}|\boldsymbol{x}))$, $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{y} \in \mathbb{R}^q$
- Easy to estimate (e.g., avoid density estimation)

Our General Idea

- Divergence on Matrix M_1 and M_2, $M \in \mathbb{S}_+^{p\times p}$
  - M_1 is a characterization of $p_1(\boldsymbol{x})$
  - M_2 is a characterization of $p_2(\boldsymbol{x})$
- Quantify divergence on $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ as the divergence on M_1 and M_2



$p_1 \Longrightarrow$ $$M_1 = \begin{bmatrix} M_{11} & \cdots & M_{1p} \\ \vdots & \ddots & \vdots \\ M_{p1} & \cdots & M_{pp} \end{bmatrix}$$



$p_2 \Longrightarrow$ $$M_2 = \begin{bmatrix} M_{11} & \cdots & M_{1p} \\ \vdots & \ddots & \vdots \\ M_{p1} & \cdots & M_{pp} \end{bmatrix}$$

$$D_{\mathcal{P}}(p_1||p_2) \Longrightarrow D_{\mathcal{M}}(M_1||M_2)$$

Bregman-Correntropy (Conditional) Divergence

Open problems

- How to construct M_1 and M_2 from P_1 and P_2?
- How to measure $D_{\mathcal{M}}(M_1||M_2)$?

Bregman-Correntropy (Conditional) Divergence

Open problems

- How to construct M_1 and M_2 from P_1 and P_2?
- Covariance matrix

$$\Sigma_x = \begin{bmatrix} \mathrm{var}(x_1) & \cdots & \mathrm{cov}(x_1, x_p) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(x_p, x_1) & \cdots & \mathrm{var}(x_p) \end{bmatrix} \in \mathbb{S}_+^{p\times p}$$

$$(\Sigma_x)_{ij} = \mathrm{cov}(x_i, x_j) = \mathbb{E}(x_i x_j) - \mathbb{E}(x_i)\mathbb{E}(x_j)$$

covariance: only linear relationship; 2nd-order statistics

- Correntropy matrix

$$C_x = \begin{bmatrix} U(x_1) & \cdots & U(x_1, x_p) \\ \vdots & \ddots & \vdots \\ U(x_p, x_1) & \cdots & U(x_p) \end{bmatrix} \in \mathbb{S}_+^{p\times p}$$

$$(C_x)_{ij} = U(x_i, x_j) = \mathbb{E}[\kappa(x_i, x_j)] - \mathbb{E}_{x_i}\mathbb{E}_{x_j}[\kappa(x_i, x_j)]$$

κ: a kernel function

centered correntropy[1,2]:

1. nonlinear counterpart of covariance in kernel space
2. contains all higher-order information (depends on kernel)

1. Rao, Murali, Sohan Seth, Jianwu Xu, Yunmei Chen, Hemant Tagare, and Jose C. Principe. "A test of independence based on a generalized correlation function." *Signal Processing*, vol. 91, no. 1, pp. 15-27, 2011.
2. Santamaría, Ignacio, Puskal P. Pokharel, and Jose C. Principe. "Generalized correlation function: definition, properties, and application to blind equalization." *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2187-2197, 2006.

Bregman-Correntropy (Conditional) Divergence

Open problems

- Given $\{C_1, C_2\}$ or $\{\Sigma_1, \Sigma_2\}$, how to measure $D_{\mathcal{M}}(M_1||M_2)$?
- Bregman matrix divergence[3] $D_{\varphi,B}$
  - φ: $\mathbb{S}_+ \rightarrow \mathbb{R}$ is a strictly convex, differentiable function
  - $D_{\varphi,B}(M_1||M_2) = \varphi(M_1) - \varphi(M_2) - \mathrm{tr}((\nabla_{\varphi}(M_2))^T(M_1 - M_2))$

- If $\varphi(M) = \mathrm{tr}(M \log M - M)$,
  - $D_{\varphi,B}(M_1||M_2) = \mathrm{tr}(M_1 \log M_1 - M_1 \log M_2 - M_1 + M_2)$
  - von Neumann divergence (D_{vN})

- If $\varphi(M) = -\log|M|$,
  - $D_{\varphi,B}(M_1||M_2) = \mathrm{tr}(M_1 M_2^{-1}) - \log\left|M_1 M_2^{-1}\right| - p$
  - Log-Determinant divergence (D_{lD})

3. Kulis, Brian, Mátyás A. Sustik, and Inderjit S. Dhillon. "Low-Rank Kernel Learning with Bregman Matrix Divergences." *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

Bregman-Correntropy (Conditional) Divergence

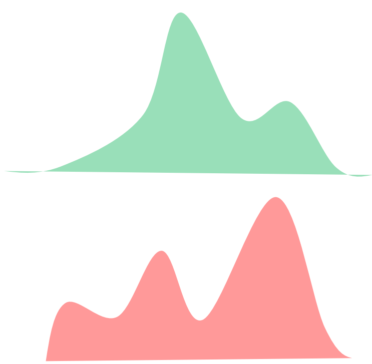
Our Measure on $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$

- $C_{\boldsymbol{x}_1}$ and $C_{\boldsymbol{x}_2}$: correntropy matrix evaluated at $P_1(\boldsymbol{x})$ and $P_2(\boldsymbol{x})$
- $D(P_1(\boldsymbol{x})||P_2(\boldsymbol{x})) = D_{\varphi,B}(C_{\boldsymbol{x}_1}||C_{\boldsymbol{x}_2})$

Our Measure on $p_1(\boldsymbol{y}|\boldsymbol{x})$ and $p_2(\boldsymbol{y}|\boldsymbol{x})$

- $C_{\boldsymbol{x}_1}$ and $C_{\boldsymbol{x}_2}$: correntropy matrix evaluated at $P_1(\boldsymbol{x})$ and $P_2(\boldsymbol{x})$
- $C_{\boldsymbol{x}_1\boldsymbol{y}_1}$ and $C_{\boldsymbol{x}_2\boldsymbol{y}_2}$: joint correntropy matrix evaluated at $P_1(\boldsymbol{x},y)$ and $P_2(\boldsymbol{x},y)$
- $D(P_1(\boldsymbol{y}|\boldsymbol{x})||P_2(\boldsymbol{y}|\boldsymbol{x})) = D_{\varphi,B}(C_{\boldsymbol{x}_1\boldsymbol{y}_1}||C_{\boldsymbol{x}_2\boldsymbol{y}_2}) - D_{\varphi,B}(C_{\boldsymbol{x}_1}||C_{\boldsymbol{x}_2})$

Bregman-Correntropy (Conditional) Divergence



$p_1 \Longrightarrow M_1 = \begin{bmatrix} M_{11} & \cdots & M_{1p} \\ \vdots & \ddots & \vdots \\ M_{p1} & \cdots & M_{pp} \end{bmatrix}$

$p_2 \Longrightarrow M_2 = \begin{bmatrix} M_{11} & \cdots & M_{1p} \\ \vdots & \ddots & \vdots \\ M_{p1} & \cdots & M_{pp} \end{bmatrix}$

$$D_{\mathcal{P}}(p_1||p_2) \Longrightarrow D_{\varphi,B}(M_1||M_2)$$

Bregman-Correntropy (Conditional) Divergence

Properties of Bregman-Correntropy (Conditional) Divergence

- Non-negative: $D_{\varphi,B}(C_{\boldsymbol{x}_1\boldsymbol{y}_1}||C_{\boldsymbol{x}_2\boldsymbol{y}_2}) - D_{\varphi,B}(C_{\boldsymbol{x}_1}||C_{\boldsymbol{x}_2}) \geq 0$

- Definiteness: suppose $\boldsymbol{y} = W^{\mathrm{T}}\boldsymbol{x}$, $D_{\varphi,B}(C_{\boldsymbol{x}_1\boldsymbol{y}_1}||C_{\boldsymbol{x}_2\boldsymbol{y}_2}) - D_{\varphi,B}(C_{\boldsymbol{x}_1}||C_{\boldsymbol{x}_2}) = 0$, iff $W_1 = W_2$

- Reduce to KL divergence on Gaussian data as a baseline, if
  - $\varphi(X) = -\log|X|$
  - Replace C (correntropy matrix) with Σ (covariance matrix)

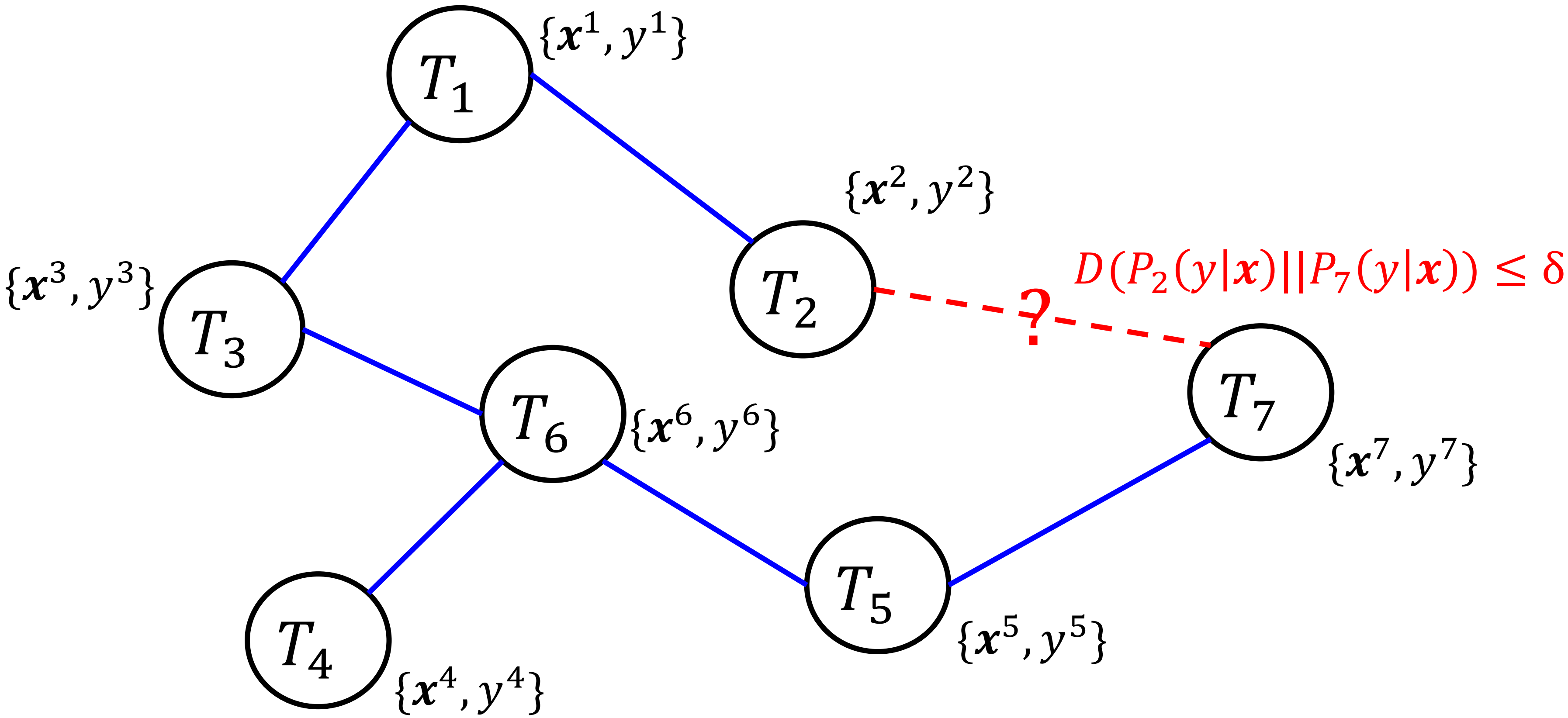
Bregman-Correntropy (Conditional) Divergence

Applications

- Task Similarity in Multi-Task Learning
- Concept Drift Detection
- Feature Selection

Bregman-Correntropy (Conditional) Divergence

Application: Task Similarity in Multi-Task Learning



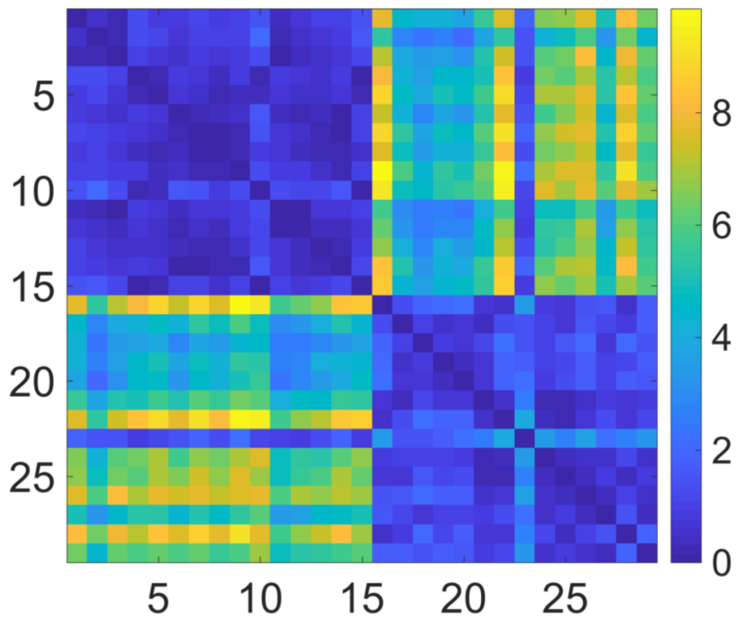
Joint learning of multiple related tasks, e.g., T_1, T_2, …

Learn from each task a $f: \boldsymbol{x} \to y$, usually $f \to p(y|\boldsymbol{x})$

Bregman-Correntropy (Conditional) Divergence

Application: Task Similarity in Multi-Task Learning

Task Similarity Visualization



Ground Truth: 29 tasks, tasks **1-15** are different from Tasks **16-29**

Bregman-Correntropy (Conditional) Divergence

Application: Concept Drift Detection

- Objective: identify the change of $p_t(y|\boldsymbol{x})$ in a data stream

- Traditional methods (DDM[4], PERM[5], etc.)
  - Train a classifier $f: \boldsymbol{x} \rightarrow y$
  - monitoring the distributional change of prediction error $e = y - f(\boldsymbol{x})$

- Our method
  - Classifier-free
  - Explicitly monitoring the change of $p_t(y|\boldsymbol{x})$ by $D_{\varphi,B}(P_t(y|\boldsymbol{x})||P_{t'}(y|\boldsymbol{x}))$

Method	Precision	Recall	Delay	Accuracy (%)
DDM	0.49	0.50	50	89.22
EDDM	0.69	0.82	230	92.60
HDDM	**1**	0.83	133	97.47
PERM	0.81	0.88	99	**97.81**
vN (Σ)	0.77	**1**	**43**	92.82
LD (Σ)	0.83	**1**	113	93.43
vN (C)	0.80	**1**	60	90.07
LD (C)	0.77	**1**	53	92.23

4. Gama, Joao, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. “Learning with drift detection.” In *Brazilian symposium on artificial intelligence*, pp. 286-295. Springer, Berlin, Heidelberg, 2004.
5. Harel, Maayan, Shie Mannor, Ran El-Yaniv, and Koby Crammer. “Concept drift detection through resampling.” In *International Conference on Machine Learning*, pp. 1009-1017. 2014.

Application: Feature Selection

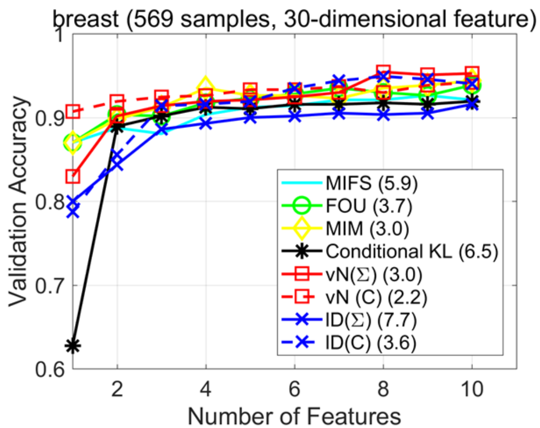
- Objective: Given a set of features $S = \{x_1, x_2, \dots, x_M\}$ and class label y, select a subset of features $S^\star \subset S$ ($|S^\star| \ll |S|$) to maximize classification accuracy.

- Traditional methods (from an information-theoretic perspective)
  - Maximize mutual information $\mathbf{I}(y; S^\star)$

- Our method
  - Maximize conditional divergence $D_{\varphi,B}\left(p(y|S^\star)||p\left(y|\tilde{S}\right)\right)$
  - $\tilde{S}$ is “useless” feature set that has no predictive power to y.

$$
\begin{aligned}
\mathbf{I}(y; S^\star) &= \iint P(y, S^\star) \log \frac{P(y, S^\star)}{P(y)P(S^\star)} \\
&= \iint \left(P(y|S^\star) \log \frac{P(y|S^\star)}{P(y)} \right) P(S^\star) \\
&= \mathbb{E}_S[D_{KL}(P(y|S^\star)||P(y))] \\
&= \mathbb{E}_S[D_{KL}(P(y|S^\star)||P(y|\tilde{S}))],
\end{aligned}
$$

Theoretical guarantee: the equivalence between our objective and maximizing mutual information $I(y; S^\star)$.

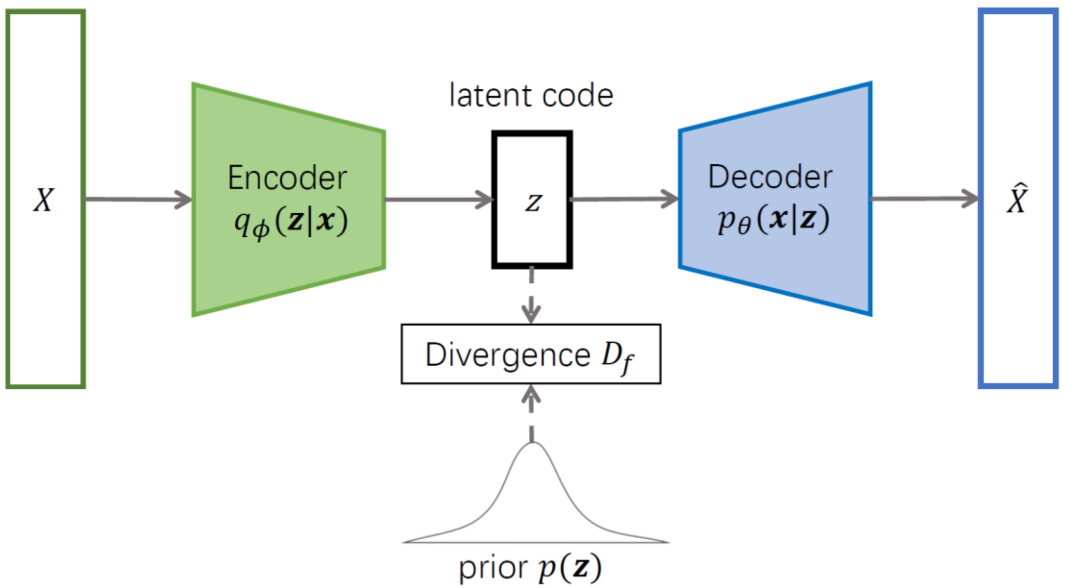


Practical performance: vN(C) refers to our D_{vN} on correntropy matrix.

Conclusions

New Estimators on Divergence and Conditional Divergence

- Easy to estimate (avoid density estimation)
- Statistically more powerful than most of existing ones (e.g., KL)
- Applicable to numerous real-world applications
- Automatically differentiable



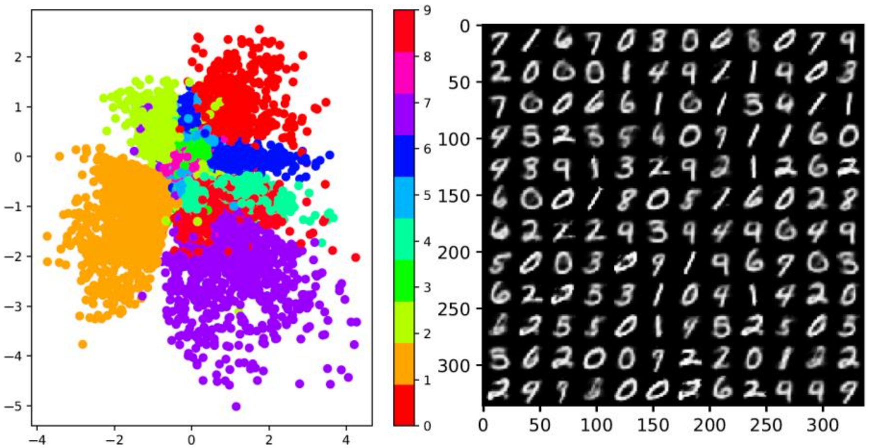
Deep Generative Autoencoder

$$L_{ours}(\theta,\phi) = \frac{1}{2}\mathbb{E}_{\hat{p}(x)}\left[\left\|x - D_\theta\left(E_\phi(x)\right)\right\|_2^2\right] + D_{\varphi,B}\left(C_{q_\phi(z)}||C_{p(z)}\right)$$

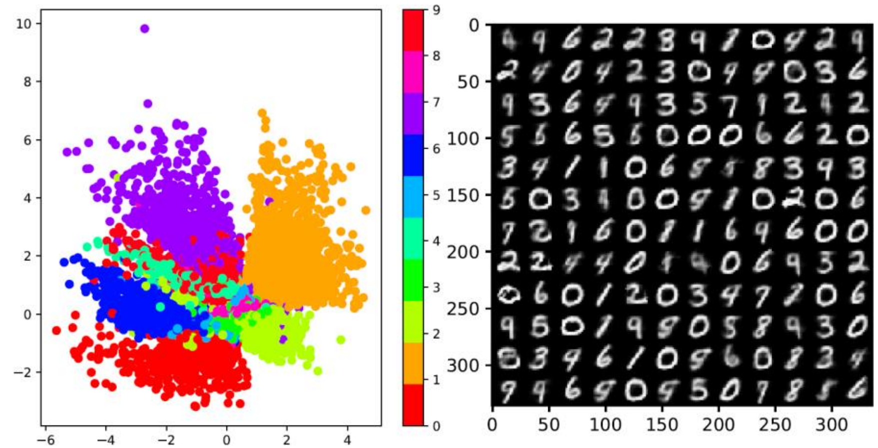
Conclusions

New Estimators on Divergence and Conditional Divergence

- Easy to estimate (avoid density estimation)
- Statistically more powerful than most of existing ones (e.g., KL)
- Applicable to numerous real-world applications
- Automatically differentiable



Gaussian prior $p(z)$



Laplacian prior $p(z)$

Conclusions

New Estimators on Divergence and Conditional Divergence

- Easy to estimate (avoid density estimation)
- Statistically more powerful than most of existing ones (e.g., KL)
- Applicable to numerous real-world applications
- Automatically differentiable
- More notes in arXiv: https://arxiv.org/abs/2005.02196

GitHub

WeChat

\Orchestrating a brighter world

NEC