Nonlinear Stein Variational Gradient Descent for Learning Diversified Mixture Models

Dilin Wang   Qiang Liu

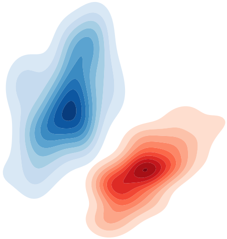
Department of Computer Science
The University of Texas at Austin

Learning Mixture Models

- Learning mixture models by maximum likelihood:

$$\max_{\Theta} \quad F(\Theta) := \mathbb{E}_{x \sim \mathcal{D}} \left[\log \left(\frac{1}{m} \sum_{i=1}^{m} p(x \mid \theta_i) \right) \right], \qquad \Theta = \{\theta_i\}_{i=1}^{m}.$$

- **Challenges**:
  - Optimization highly non-convex.
  - Promoting diversification increases robustness [e.g., Borodin, 2009; xie et al., 2018].



- **Our work**:
  - A variational view + entropic regularization.
  - Optimized by generalizing stein variational gradient descent [Liu, Wang 16].

Learning Diversified Infinite Mixtures

- **Step 1**: Relaxing to **learning infinite mixtures**:

$$\max_{\rho} \mathcal{F}[\rho] := \mathbb{E}_{x\sim D}\left[\log\left(\underbrace{\mathbb{E}_{\theta\sim\rho}[\ p(x \mid \theta)\ ]}_{\text{infinite mixture models}} \right)\right]$$

  - Reduces to finite case when $\rho := \sum_{i=1}^{m} \delta_{\theta_i}/m$

- **Step 2**: Add **entropy regularization** to enforce diversity:

$$\max_{\rho} \quad \mathcal{J}[\rho] := \mathcal{F}[\rho] + \alpha\mathcal{H}[\rho],$$

  - Entropy: $\mathcal{H}[\rho] = -\int \rho \log \rho$.

Learning Diversified Infinite Mixtures

- **Step 1**: Relaxing to **learning infinite mixtures**:

$$\max_{\rho} \mathcal{F}[\rho] := \mathbb{E}_{x\sim D}\Big[\log\Big(\underbrace{\mathbb{E}_{\theta\sim\rho}[\, p(x \mid \theta)\,]}_{\text{infinite mixture models}} \Big)\Big]$$

  - Reduces to finite case when $\rho := \sum_{i=1}^{m} \delta_{\theta_i}/m$

- **Step 2**: Add **entropy regularization** to enforce diversity:

$$\max_{\rho} \quad \mathcal{J}[\rho] = \underbrace{\mathcal{F}[\rho]}_{\substack{\text{likelihood} \\ \text{(nonlinear functional)}}} + \underbrace{\alpha\mathcal{H}[\rho]}_{\substack{\text{diversity} \\ \text{(entropy)}}},$$

  - A difficult problem to solve.
  - Achieved by generalizing **Stein variational gradient descent (SVGD)** [Liu, Wang 16].

Nonlinear SVGD: Derivation

- Want to approximate $\max_{\rho} \quad \mathcal{J}[\rho] = \mathcal{F}[\rho] + \alpha\mathcal{H}[\rho]$.
- Approximate it with $\rho := \sum_i \delta_{\theta_i}/m$.
- Iteratively update $\{\theta_i\}$ to yield steepest descent on $\mathcal{J}[\rho]$:

$$\theta_i' \leftarrow \theta_i + \epsilon\phi(\theta_i), \qquad \phi^* \approx \arg\max_{\phi\in\mathcal{F}}(J[\rho'] - J[\rho])$$

  - ρ' is the density of updated θ_i'.
  - $\mathcal{F}$ is the unit ball of a reproducing kernel Hilbert space (RKHS), with a positive definite kernel $k(\theta_i, \theta_j)$.

Yields a Simple Algorithm

- Starting from an initial $\{\theta_i\}$, repeat:

$$\theta_i \leftarrow \theta_i + \epsilon \hat{\mathbb{E}}_{\theta_j \sim \rho} \Big[\underbrace{\nabla_{\theta_j} F(\Theta) \; k(\theta_i, \theta_j)}_{weighted \;\; sum \;\; of \;\; gradient} \quad + \quad \underbrace{\alpha \nabla_{\theta_j} k(\theta_i, \theta_j)}_{repulsive \;\; force} \Big], \quad \forall i$$

  - $\nabla_{\theta_j} F(\Theta)$: the gradient of standard log likelihood.

- Return $\rho = \sum_i \delta_{\theta_i} / m$.

- In comparison, gradient descent of standard log likelihood is

$$\theta_i \leftarrow \theta_i + \epsilon \nabla_{\theta_i} F(\Theta), \qquad \forall i$$

Deep Embedded Clustering

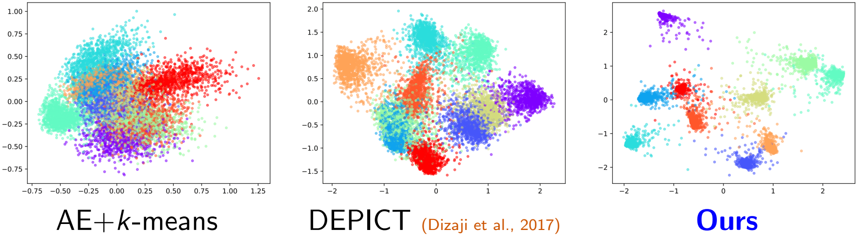


Figure: 2D-visualization with PCA on MNIST.

	DEC	JULE	DEPICT	**Ours**
	Xie et al., 2016	Yang et al., 2016	Dizaji et al., 2017	
NMI	0.816	0.913	0.917	**0.933**
ACC	0.844	0.964	0.965	**0.974**

Table: Results on MNIST.

Deep Anomaly Detection

- Applied our method to improve deep anomaly detection.

Method	Precision	Recall	F1
DSEBM Zhai et al., 2016	0.7369	0.7477	0.7423
DCN Yang et al., 2017	0.7696	0.7829	0.7762
DAGMM-p Zong et al., 2018	0.7579	0.7710	0.7644
DAGMM-NVI Zong et al., 2018	0.9290	0.9447	0.9368
DAGMM Zong et al., 2018	0.9297	0.9442	0.9369
Ours	**0.9659**	**0.9490**	**0.9573**

Table: Results on KDDCUP99 dataset

Conclusions

1. A new method to learn diversified mixture models
2. Generalizing Stein variational gradient descent (SVGD)
3. Simple and practical!

Poster #231. Today 06:30 – 09:00 PM @ Pacific Ballroom

Thank You