Task-Driven Dictionary Learning

Julien Mairal, Francis Bach, and Jean Ponce

Presented by: Bailey Kong

Outline

- Unsupervised Dictionary Learning
- Supervised Dictionary Learning
  - Basic Formulation
  - Extensions
- Optimization
- Results

Unsupervised Dictionary Learning

$$\min_{\mathbf{D}\in\mathcal{D}} g(\mathbf{D}) \text{ s.t. } \|\mathbf{d}_i\|_2 \leq 1 \quad \text{for } i = 1, \ldots, p$$

$$g(\mathbf{D}) \triangleq \mathbb{E}_{\mathbf{x}}[\ell_u(\mathbf{x}, \mathbf{D})] \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ell_u(\mathbf{x}_i, \mathbf{D})$$

$\mathbf{x}_i \in \mathbb{R}^m$ image patch, $\mathbf{D} \in \mathbb{R}^{m\times p}$ dictionary

Unsupervised Dictionary Learning

$$\arg\min_{\mathbf{D}\in\mathcal{D}} g(\mathbf{D}) \text{ s.t. } \|\mathbf{d}_i\|_2 \leq 1 \quad \text{for } i = 1, \ldots, p$$

$$g(\mathbf{D}) \triangleq \sum_{i=1}^{n} \ell_u(\mathbf{x}_i, \mathbf{D})$$

$\mathbf{x}_i \in \mathbb{R}^m$ image patch, $\mathbf{D} \in \mathbb{R}^{m \times p}$ dictionary

Unsupervised Dictionary Learning

$$\min_{\mathbf{D}\in\mathcal{D}} g(\mathbf{D}) \text{ s.t. } \|\mathbf{d}_i\|_2 \leq 1 \quad \text{for } i = 1, \ldots, p$$

$$g(\mathbf{D}) \triangleq \mathbb{E}_{\mathbf{x}}[\ell_u(\mathbf{x}, \mathbf{D})] \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ell_u(\mathbf{x}_i, \mathbf{D})$$

$\mathbf{x}_i \in \mathbb{R}^m$ image patch, $\mathbf{D} \in \mathbb{R}^{m\times p}$ dictionary

ℓ_u loss function -LASSO (basis pursuit)

$$\ell_u(\mathbf{x}, \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

Unsupervised Dictionary Learning

$$\min_{\mathbf{D}\in\mathcal{D}} g(\mathbf{D}) \text{ s.t. } \|\mathbf{d}_i\|_2 \leq 1 \quad \text{for } i = 1, \ldots, p$$

$$g(\mathbf{D}) \triangleq \mathbb{E}_{\mathbf{x}}[\ell_u(\mathbf{x}, \mathbf{D})] \overset{\text{a.s.}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ell_u(\mathbf{x}_i, \mathbf{D})$$

$\mathbf{x}_i \in \mathbb{R}^m$ image patch, $\mathbf{D} \in \mathbb{R}^{m\times p}$ dictionary

ℓ_u loss function - elastic net

$$\ell_u(\mathbf{x}, \mathbf{D}) \triangleq \min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2$$

Elastic Net

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2+\lambda_1\|\boldsymbol{\alpha}\|_1+\frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2$$

Authors choose for stability reasons

- Strongly convex when $\lambda_2 > 0$
- Critical for some tasks beyond reconstruction

Limitations of LASSO [1]

- High-dimensional data with few examples (p>n)
- Highly correlated variables: LASSO tends to select one variable from a group and ignore others

[1] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005): 301-320.

Unsupervised Dictionary Learning

Classification learning

$$\min_{\mathbf{W}\in\mathcal{W}} f(\mathbf{W}) + \frac{\nu}{2}\|\mathbf{W}\|_F^2$$

$$f(\mathbf{W}) \triangleq \mathbb{E}_{\mathbf{y},\mathbf{x}}[\ell_s(\mathbf{y}, \mathbf{W}, \mathbf{a}^\star(\mathbf{x}, \mathbf{D}))]$$

$\mathbf{y}$ class labels or regression target

ℓ_s convex loss function - square, logistic, hinge…

$\mathbf{a}^\star(\mathbf{x}, \mathbf{D})$ optimal sparse codes given the dictionary

Supervised Dictionary Learning

Basic formulation

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} f(\mathbf{D},\mathbf{W}) + \frac{\nu}{2}\|\mathbf{W}\|_F^2$$

$$f(\mathbf{D},\mathbf{W}) \triangleq \mathbb{E}_{\mathbf{y},\mathbf{x}}[\ell_s(\mathbf{y},\mathbf{W},\mathbf{a}^\star(\mathbf{x},\mathbf{D}))]$$

$\mathbf{y}$ class labels or regression target

ℓ_s convex loss function - square, logistic, hinge…

$\mathbf{a}^\star(\mathbf{x},\mathbf{D})$ optimal sparse codes given the dictionary

Supervised Dictionary Learning

Extensions

1. Learning a linear transform of the input data

$$f(\mathbf{D}, \mathbf{W}, \mathbf{Z}) \triangleq \mathbb{E}_{\mathbf{y},\mathbf{x}}[\ell_s(\mathbf{y}, \mathbf{W}, \boldsymbol{\alpha}^\star(\mathbf{Z}\mathbf{x}, \mathbf{D}))]$$

2. Semi-supervised learning

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} (1-\mu)\mathbb{E}_{\mathbf{y},\mathbf{x}}[\ell_s(\mathbf{y}, \mathbf{W}, \boldsymbol{\alpha}^\star(\mathbf{x}, \mathbf{D}))] + \mu\mathbf{E}_{\mathbf{x}}[\ell_u(\mathbf{x}, \mathbf{D})] + \frac{\nu}{2}\|\mathbf{W}\|_F^2$$

Applications

1. Regression

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} \mathbb{E}\left[\frac{1}{2}\|\mathbf{y}-\mathbf{W}\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D})\|_2^2\right]+\frac{\nu}{2}\|\mathbf{W}\|_F^2$$

2. Classification

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} \mathbb{E}\left[\log(1+e^{-y\mathbf{w}^T\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D})})\right]+\frac{\nu}{2}\|\mathbf{w}\|_2^2$$

3. Compressed Sensing

$$\min_{\mathbf{D}\in\mathcal{D},\mathbf{W}\in\mathcal{W}} \mathbb{E}\left[\frac{1}{2}\|\mathbf{y}-\mathbf{W}\boldsymbol{\alpha}^\star(\mathbf{Z}\mathbf{x},\mathbf{D})\|_2^2\right]+\frac{\nu_1}{2}\|\mathbf{W}\|_F^2+\frac{\nu_2}{2}\|\mathbf{Z}\|_F^2$$

Optimization

$$\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D}) \triangleq \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\alpha}\|_2^2$$

$$\nabla_{\mathbf{W}} f(\mathbf{D},\mathbf{W}) = \mathbb{E}_{\mathbf{y},\mathbf{x}}[\nabla_{\mathbf{W}} \ell_s(\mathbf{y},\mathbf{W},\boldsymbol{\alpha}^\star)]$$

$$\nabla_{\mathbf{D}} f(\mathbf{D},\mathbf{W}) = \mathbb{E}_{\mathbf{y},\mathbf{x}}[-\mathbf{D}\boldsymbol{\beta}^\star\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D})^T + (\mathbf{x}-\mathbf{D}\boldsymbol{\alpha}^\star(\mathbf{x},\mathbf{D}))\boldsymbol{\beta}^{\star T}]$$

$$\boldsymbol{\beta}^\star_{\Lambda^C} = 0 \qquad \boldsymbol{\beta}^\star_{\Lambda} = (\mathbf{D}_\Lambda^T\mathbf{D}_\Lambda + \lambda_2\mathbf{I})^{-1}\nabla_{\boldsymbol{\alpha}\Lambda}\ell_s(\mathbf{y},\mathbf{W},\boldsymbol{\alpha}^\star)$$

Results - Handwritten Digit Classification

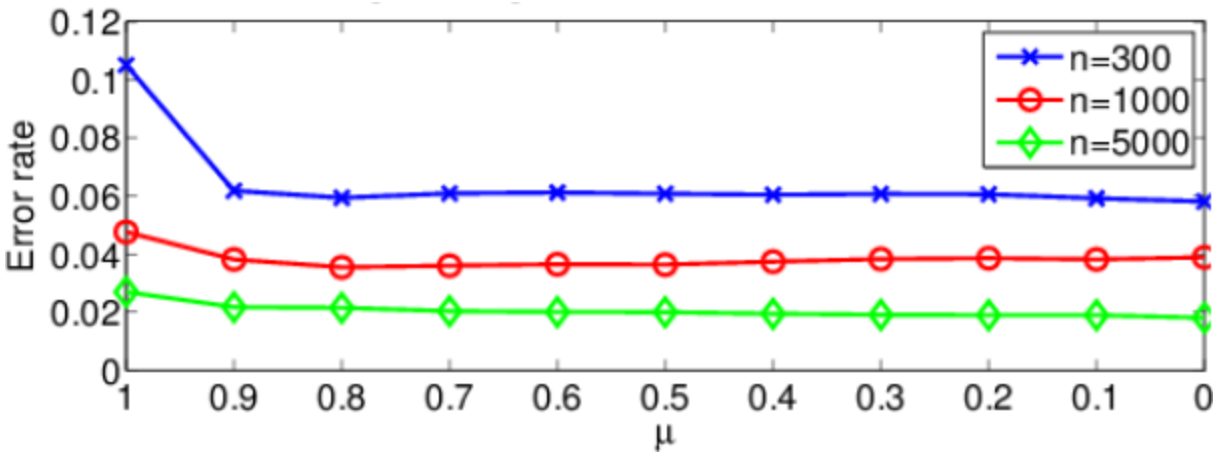
Dictionary Size

D	unsupervised				supervised			
p	50	100	200	300	50	100	200	300
MNIST	5.27	3.92	2.95	2.36	.96	.73	.57	.54
USPS	8.02	6.03	5.13	4.58	3.64	3.09	2.88	2.84

MNIST: 28x28 images; 60K train; 10K test

USPS: 16x16 images; 7,291 train; 2,007 test

Results - Handwritten Digit Classification



mu=0 fully supervised

mu=1 fully unsupervised

Results - Nonlinear Image Mapping

	Validation set				Test set							
Image	1	2	3	4	5	6	7	8	9	10	11	12
FIHT2	30.8	25.3	25.8	31.4	24.5	28.6	29.5	28.2	29.3	26.0	25.2	24.7
WInHD	31.2	26.9	26.8	31.9	25.7	29.2	29.4	28.7	29.4	28.1	25.6	26.4
LPA-ICI	31.4	27.7	26.5	32.5	25.6	29.7	30.0	29.2	30.1	28.3	26.0	27.2
SA-DCT	32.4	28.6	27.8	**33.0**	**27.0**	30.1	30.2	29.8	30.3	28.5	**26.2**	27.6
Ours	**33.0**	**29.6**	**28.1**	**33.0**	26.6	**30.2**	**30.5**	**29.9**	**30.4**	**29.0**	**26.2**	**28.0**

Formulated as a regular regression problem, no prior on task.

Results - Nonlinear Image Mapping

Results - Nonlinear Image Mapping

Results - Compressed Sensing

Z	RANDOM			SL1	PCA			SL2
D	DCT	UL	SL	SL	DCT	UL	SL	SL
$r = 5$	77.3 ± 4.0	76.9 ± 4.0	76.7 ± 4.0	54.1 ± 1.3	49.9 ± 0.0	47.6 ± 0.0	47.5 ± 0.1	47.3 ± 0.3
$r = 10$	57.8 ± 1.5	56.5 ± 1.5	55.7 ± 1.4	36.5 ± 0.7	33.7 ± 0.0	32.3 ± 0.0	32.3 ± 0.1	31.9 ± 0.2
$r = 20$	37.1 ± 1.2	35.4 ± 1.0	34.5 ± 0.9	21.4 ± 0.1	20.4 ± 0.0	19.7 ± 0.0	19.6 ± 0.1	19.4 ± 0.2
$r = 40$	19.3 ± 0.8	18.5 ± 0.7	18.0 ± 0.6	10.0 ± 0.3	9.2 ± 0.0	9.1 ± 0.0	9.0 ± 0.0	9.0 ± 0.0

Thanks!