Online Learning for Group Lasso

Haiqin Yang

Department of Computer Science & Engineering
The Chinese University of Hong Kong

March 29, 2010

Outline

1. Introduction
2. Motivations and Contributions
3. Algorithm and Regret Bound
4. Experiments
5. Conclusions

Group Lasso

Introduction

- ✓ A natural extension of Lasso (Tibshirani, 1996)
- ✓ Find important explanatory factors in a group manner (Yuan & Lin, 2006)

Applications with structured sparsity

- ✓ Speech and signal processing (McAuley et al., 2005)
- ✓ Bioinformatics (Lanckriet et al., 2004; Meier et al., 2008)
- ✓ Computer vision (Harchaoui and Bach, 2007; Huang et al., 2009)

Group Lasso

Data

$\mathbf{X} : \mathbb{R}^{N\times d}$

$\mathbf{Y} : \mathbb{R}^{N}$, or $\{\pm 1\}^N$

G groups

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_i^1 \\ \vdots \\ \mathbf{x}_i^G \end{pmatrix}$$

Models

Lasso (Tibshirani, 1996):

$$\min_{\mathbf{w}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \lambda\|\mathbf{w}\|_1$$

Group Lasso (Yuan & Lin, 2006):

$$\min_{\mathbf{w}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \lambda\sum_{g=1}^{G}\sqrt{d_g}\|\mathbf{w}^g\|_2$$

Sparse Group Lasso (Friedman et al., 2010):

$$\min_{\mathbf{w}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \lambda\sum_{g=1}^{G}(\sqrt{d_g}\|\mathbf{w}^g\|_2 + r_g\|\mathbf{w}^g\|_1)$$

Illustrations

(a) Group Lasso

(b) Sparse Group Lasso

Motivations and Contributions

Limitations

- ✓ Learned by the batch-mode training; training data may appear sequentially
- ✓ Only handle data up to several thousands of instances or features
- ✓ Yield solutions with sparsity in the group level

Contributions

- ✓ First proposed online learning algorithm for the Group Lasso algorithms
- ✓ Efficiency: $\mathcal{O}(d)$ memory and computation at each step
- ✓ Sparse solutions on both group level and elemental levels
- ✓ Provide regret bound on the online learning algorithm

Algorithm Framework

Objective: $$\min_{\mathbf{w}} \sum_{i=1}^{N} l(\mathbf{w},\ \mathbf{z}_i) + \Omega_\lambda(\mathbf{w}),$$

Algorithm 1 Online learning algorithm for group lasso

Initialization: $\mathbf{w}_1 = \mathbf{w}_0$, $\bar{\mathbf{u}}_0 = \mathbf{0}$.

for $t = 1, 2, 3, \ldots$ **do**

Given the function l_t, compute the subgradient on $\mathbf{w}_t$, $\mathbf{u}_t \in \partial\, l_t$.

Update the average subgradient $\bar{\mathbf{u}}_t$:

$$\bar{\mathbf{u}}_t = \frac{t-1}{t}\bar{\mathbf{u}}_{t-1} + \frac{1}{t}\mathbf{u}_t.$$

Calculate the next iteration $\mathbf{w}_{t+1}$:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \Upsilon(\mathbf{w}) \triangleq \left\{\bar{\mathbf{u}}_t^\top \mathbf{w} + \Omega_\lambda(\mathbf{w}) + \frac{\gamma}{\sqrt{t}} h(\mathbf{w})\right\}$$

end for

Update rules

Group Lasso: $\Omega_\lambda(\mathbf{w}) = \lambda \sum_{g=1}^{G} \sqrt{d_g}\|\mathbf{w}^g\|_2$, $h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$

$$\mathbf{w}_{t+1}^{g} = -\frac{\sqrt{t}}{\gamma}\left[1 - \frac{\lambda\sqrt{d_g}}{\|\bar{\mathbf{u}}_t^g\|_2}\right]_+ \cdot \bar{\mathbf{u}}_t^g$$

Sparse Group Lasso: $\Omega_{\lambda,\mathbf{r}}(\mathbf{w}) = \lambda \sum_{g=1}^{G} \left(\sqrt{d_g}\|\mathbf{w}^g\|_2 + r_g\|\mathbf{w}^g\|_1\right)$, $h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$

$$\mathbf{w}_{t+1}^{g} = -\frac{\sqrt{t}}{\gamma}\left[1 - \frac{\lambda\sqrt{d_g}}{\|\mathbf{c}_t^g\|_2}\right]_+ \cdot \mathbf{c}_t^g, \quad c_t^{g,j} = \left[|\bar{u}_t^{g,j}| - \lambda r_g\right]_+ \cdot \mathrm{sign}\left(\bar{u}_t^{g,j}\right)$$

Enhanced Sparse Group Lasso: $\Omega_{\lambda,\mathbf{r}}(\mathbf{w}) = \lambda \sum_{g=1}^{G} \left(\sqrt{d_g}\|\mathbf{w}^g\|_2 + r_g\|\mathbf{w}^g\|_1\right)$,
$h(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 + \rho\|\mathbf{w}\|_1$

$$\mathbf{w}_{t+1}^{g} = -\frac{\sqrt{t}}{\gamma}\left[1 - \frac{\lambda\sqrt{d_g}}{\|\tilde{\mathbf{c}}_t^g\|_2}\right]_+ \cdot \tilde{\mathbf{c}}_t^g, \quad \tilde{c}_t^{g,j} = \left[|\bar{u}_t^{g,j}| - \lambda r_g - \frac{\gamma\rho}{\sqrt{t}}\right]_+ \cdot \mathrm{sign}\left(\bar{u}_t^{g,j}\right)$$

Theoretical results

Average regret

$$\bar{R}_T(\mathbf{w}) := \frac{1}{T}\sum_{t=1}^{T}(\Omega_\lambda(\mathbf{w}_t) + l_t(\mathbf{w}_t)) - S_T(\mathbf{w})$$

Theoretical bounds

Given $h(\mathbf{w}^\star) \le D^2$ and $\|\bar{\mathbf{u}}_T\|_*^2 \le L^2$

$$\bar{R}_T \le \left(\gamma\sqrt{T}D^2 + \frac{L^2}{2\gamma}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\right)/T \le \left(\gamma D^2 + \frac{L^2}{\gamma}\right)/\sqrt{T}$$

$$\frac{1}{2}\|\mathbf{w}_{T+1} - \mathbf{w}^\star\|^2 \le D^2 + \frac{L^2}{\gamma^2} - \frac{\sqrt{T}}{\gamma}\bar{R}_T$$

Experimental setup

Data

- ★ Synthetic data
- ★ Realworld data for gene finding

Comparison algorithms

- ★ Lasso
- ★ Group Lasso (GL)
- ★ L_1-RDA
- ★ DA-GL
- ★ DA-SGL

Synthetic data

Data generation scheme: sparsity on both group and element levels

- ✓ $\mathbf{w} \in \mathbb{R}^{100}$, $w_i = \pm 1$
- ✓ $G = 10$, $\#$ NNZ $= \{10, 8, 6, 4, 2, 1, 0, 0, 0, 0\}$
- ✓ $\mathbf{x}_i = L\mathbf{v}_i$,
  L: Cholesky decomposition of the correlation matrix, $\Sigma^g_{i,j} = 0.2^{|i-j|}$
- ✓ $y_i = \text{sign}\left(\mathbf{w}^\top \mathbf{x}_i + \epsilon\right)$

Measurement

- ✓ Accuracy
- ✓ Average F1 score: measure true weight

Synthetic data results

Accuracy

- ★ Accuracies increase with the increase of the number of training samples
- ★ DA-SGL achieves the best accuracy, especially when the number of training sample is small
- ★ DA-GL achieves slightly worse results than the DA-SGL and the GL when the number of training sample is large
- ★ Two batch-trained algorithms achieve nearly the same accuracy when the number of training samples is large

	Lasso	GL	L_1-RDA	DA-GL	DA-SGL
25	54.2± 14.1	54.2± 11.4	56.6± 9.9	57.0± 11.6	**57.6**± 11.0
50	58.2± 7.7	60.0± 6.3	59.5± 6.9	**60.9**± 6.2	**60.9**± 6.0
100	62.7± 5.5	64.0± 5.1	61.7± 4.8	64.5± 4.1	**64.6**± 4.5
250	71.1± 4.5	72.1± 4.5	64.9± 3.7	71.6± 2.7	**72.3**± 2.8
500	75.6± 2.4	75.7± 2.3	66.2± 3.0	74.8± 2.3	**75.9**± 2.2
1000	77.7± 1.5	77.8± 1.5	65.9± 2.0	76.3± 1.4	**77.9**± 1.6
2000	**79.0**± 0.7	78.9± 0.7	67.4± 1.6	77.7± 0.9	**79.0**± 1.4
5000	**79.4**± 0.4	**79.4**± 0.3	67.8± 1.5	78.2± 0.6	**79.4**± 0.8

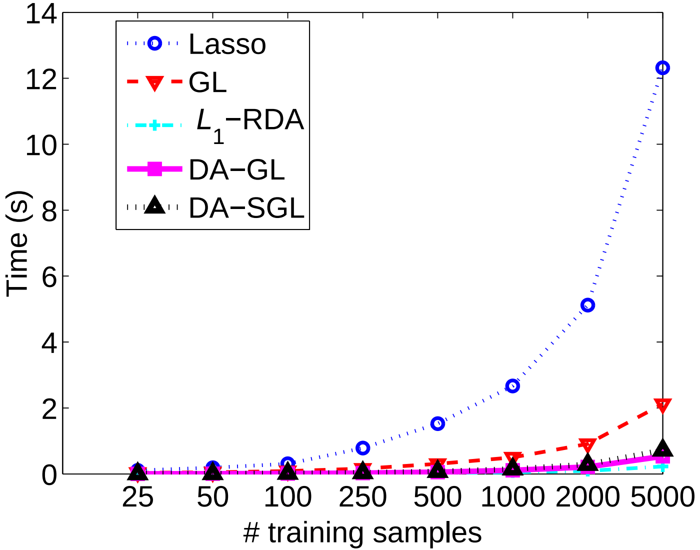
Synthetic data results

Averaged F1 score

- ★ DA-SGL outperforms all other four algorithms
- ★ The DA-SGL combines both the advantages of the lasso and the GL
- ★ GL and the DA-GL got similar average F1 scores

	Lasso	GL	L_1-RDA	DA-GL	DA-SGL
25	23.6± 8.5	37.3± 13.6	35.6± 6.3	37.2± 3.0	**37.9**± 4.5
50	35.0± 9.3	**49.8**± 6.0	39.7± 6.5	49.7± 3.0	**49.8**± 4.9
100	47.0± 7.2	**57.4**± 2.4	46.5± 9.7	57.1± 2.7	**57.4**± 5.9
250	60.0± 3.0	60.4± 2.0	59.0± 9.6	60.7± 4.0	**65.5**± 7.5
500	65.0± 2.5	65.5± 2.1	63.6± 9.7	65.2± 6.8	**81.9**± 5.3
1000	70.1± 2.4	67.2± 2.1	64.9± 8.7	67.2± 4.7	**87.3**± 4.3
2000	76.0± 2.0	68.0± 1.5	65.7± 7.4	68.2± 3.3	**91.4**± 3.0
5000	88.2± 2.4	68.2± 2.0	66.8± 8.0	68.3± 2.9	**93.7**± 2.5

Efficiency

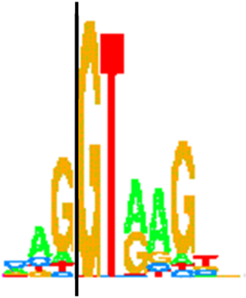


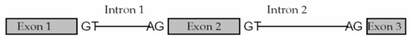
Lasso
GL
L_1−RDA
DA−GL
DA−SGL
Time (s)
0
2
4
6
8
10
12
14
25
50
100
250
500
1000
2000
5000
training samples

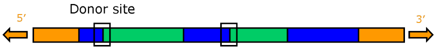
Splice Site Detection

Description

- Splice sites: regions between coding (exons) and non-coding (introns) DNA segments
- Donor splice site: 5' end of an intron
- Training set: 8,415 true, 179,438 false donor site
- Test set: 4,208 true, 89,717 false donor site
- Remove consensus "GT", length = 7







Results

% Non-zero	L1-RDA	DA-GL	DA-SGL
10	0.5632	**0.5656**	**0.5656**
40	0.6056	0.6071	**0.6082**
60	0.6481	0.6496	**0.6501**
80	0.6494	**0.6520**	**0.6520**

Conclusions

Conclusions

- A novel online learning algorithm framework for the group lasso
- Apply this framework for several group lasso models
- Provides closed-form solutions to update the models
- Give the convergence rate of the average regret
- Experimental results demonstrate the proposed algorithms in both efficiency and effectiveness

Future work

- Evaluate on other online learning algorithms, e.g., FOBOS
- Study lazy update schemes to handle high-dimensional data
- Derive a faster convergence rate for the online learning algorithm
- Extend the framework to solve other related problems

Questions ?

Haiqin Yang
www.cse.cuhk.edu.hk/~hqyang
hqyang@cse.cuhk.edu.hk