Outline

- Sparse Representation
- Existing Optimization Methods
- Coordinate Descent Methods
- Other Methods
- Experiments
Outline

- Sparse Representation
- Existing Optimization Methods
- Coordinate Descent Methods
- Other Methods
- Experiments
Sparse Representation

A mathematical way to model a signal, an image, or a document is

$$y = Xw$$

$$= w_1 \begin{bmatrix} x_{11} \\ \vdots \\ x_{l1} \end{bmatrix} + \cdots + w_n \begin{bmatrix} x_{1n} \\ \vdots \\ x_{ln} \end{bmatrix}$$

- A signal is a linear combination of others
- $X$ and $y$ are given
- We would like to find $w$ with as few non-zeros as possible (sparsity)
Example: Image Deblurring

- Consider
  \[ y = Hz \]
- \( z \): original image, \( H \): blur operation
- \( y \): observed image
- Assume
  \[ z = Dw \]
  with known dictionary \( D \)
- Try to
  \[ \min_w \|y - HDw\| \] and get \( \hat{w} \)
We hope $\mathbf{w}$ has few nonzeros as each image is generated using only several columns of the dictionary.

The restored image is $D\hat{\mathbf{w}}$.
Example: Face Recognition

- Assume a face image is a combination of the same person’s other images

\[
\begin{bmatrix}
  x_{11} \\
  \vdots \\
  x_{l1}
\end{bmatrix} : 1\text{st image}, \quad \begin{bmatrix}
  x_{12} \\
  \vdots \\
  x_{l2}
\end{bmatrix} : 2\text{nd image}, \ldots
\]

\(l\): number of pixels in a face image

- Given a face image \(y\) and collections of two persons’ faces \(X_1\) and \(X_2\)
Example: Face Recognition (Cont’d)

- If

\[
\min_w \| y - X_1 w \| < \min_w \| y - X_2 w \|,
\]

predict \( y \) as the first person

- We hope \( w \) has few nonzeros as noisy images shouldn’t be used
Example: Feature Selection

- Given

\[
X = \begin{bmatrix}
  x_{11} & \ldots & x_{1n} \\
  \vdots & & \vdots \\
  x_{l1} & \ldots & x_{ln}
\end{bmatrix}, \quad [x_{i1} \ldots x_{in}] : \text{ith document}
\]

- \( y_i = +1 \) or \(-1\) (two classes)
- We hope to find \( w \) such that

\[
w^T x_i \begin{cases} 
  > 0 & \text{if } y_i = 1 \\
  < 0 & \text{if } y_i = -1
\end{cases}
\]
Example: Feature Selection (Cont’d)

- Try to

\[
\min_w \sum_{i=1}^{l} e^{-y_i w^T x_i}
\]

and hope that \( w \) is sparse

- That is, we assume that each document is generated from important features

\[ w_i \neq 0 \Rightarrow \text{important features} \]
Finding $w$ with the smallest number of non-zeros is difficult.

$$\|w\|_0: \text{number of nonzeros}$$

Instead, L1-norm minimization

$$\min_w C\|y - Xw\|^2 + \|w\|_1$$

$C$: a parameter given by users

This is a regression problem.
L1-norm Minimization II

- 1-norm versus 2-norm

\[ \| \mathbf{w} \|_1 = |w_1| + \cdots + |w_n| \]
\[ \| \mathbf{w} \|_2^2 = w_1^2 + \cdots + w_n^2 \]

- Two figures
L1-norm Minimization III

- If using 2-norm, all $w_i$ are non-zeros
- Using 1-norm, many $w_i$ may be zeros
- Smaller $C$, better sparsity
L1-regularized Classifier

- Training data \( \{y_i, x_i\}, x_i \in \mathbb{R}^n, i = 1, \ldots, l, y_i = \pm 1 \)
- \( l \): # of data, \( n \): # of features

\[
\min_{w} \|w\|_1 + C \sum_{i=1}^{l} \xi(w; x_i, y_i)
\]

- \( \xi(w; x_i, y_i) \): loss function
- Logistic loss:
  \[
  \log(1 + e^{-yw^T x})
  \]
- L1 and L2 losses:
  \[
  \max(1 - yw^T x, 0) \quad \text{and} \quad \max(1 - yw^T x, 0)^2
  \]
- We do not consider kernels
L1-regularized Classifier (Cont’d)

- $\|w\|_1$ not differentiable $\Rightarrow$ causes difficulties in optimization
- Loss functions: logistic loss twice differentiable, L2 loss differentiable, and L1 loss not differentiable
- We focus on logistic and L2 loss
- Sometimes bias term is added

\[ w^T x \Rightarrow w^T x + b \]
L1-regularized Classifier (Cont’d)

- Many available methods; we review existing methods and show details of some methods
- Notation:

\[
f(w) \equiv \|w\|_1 + C \sum_{i=1}^{l} \xi(w; x_i, y_i)
\]

is the function to be minimized, and

\[
L(w) \equiv C \sum_{i=1}^{l} \xi(w; x_i, y_i).
\]

- We do not discuss L1-regularized regression, which is another hot topic recently
Outline

- Sparse Representation
- Existing Optimization Methods
- Coordinate Descent Methods
- Other Methods
- Experiments
Decomposition Methods

Working on some variables at a time

Cyclic coordinate descent methods

- Working variables sequentially or randomly selected

One-variable case:

$$\min_z f(w + ze_j) - f(w)$$

$e_j$: indicator vector for the $j$th element

Examples: Goodman (2004); Genkin et al. (2007); Balakrishnan and Madigan (2005); Tseng and Yun (2007); Shalev-Shwartz and Tewari (2009); Duchi and Singer (2009); Wright (2010)
Decomposition Methods (Cont’d)

Gradient-based working set selection
- Higher cost per iteration; larger working set
- Examples: Shevade and Keerthi (2003); Tseng and Yun (2007); Yun and Toh (2009)

Active set method
- Working set the same as the set of non-zero $w$ elements
- Examples: Perkins et al. (2003)
Constrained Optimization

Replace $\mathbf{w}$ with $\mathbf{w}^+ - \mathbf{w}^-$:

$$\min_{\mathbf{w}^+, \mathbf{w}^-} \sum_{j=1}^{n} w_j^+ + \sum_{j=1}^{n} w_j^- + C \sum_{i=1}^{l} \xi(\mathbf{w}^+ - \mathbf{w}^-; x_i, y_i)$$

s. t. $w_j^+ \geq 0, \ w_j^- \geq 0, \ j = 1, \ldots, n.$

- Any bound-constrained optimization methods can be used
- Examples: Schmidt et al. (2009) used Gafni and Bertsekas (1984); Kazama and Tsujii (2003) used Benson and Moré (2001); we have considered Lin and Moré (1999); Koh et al. (2007): interior point method
Constrained Optimization (Cont’d)

Equivalent problem with non-smooth constraints:

\[
\min_w \sum_{i=1}^{l} \xi(w; x_i, y_i)
\]
subject to \( \|w\|_1 \leq K \).

- \( C \) replaced by a corresponding \( K \)
- Go back to LASSO (Tibshirani, 1996) if \( y \in R \) and least-square loss
- Examples: Kivinen and Warmuth (1997); Lee et al. (2006); Donoho and Tsaig (2008); Duchi et al. (2008); Liu et al. (2009); Kim and Kim (2004); Kim et al. (2008); Roth (2004)
Other Methods

- Stochastic gradient descent: Langford et al. (2009); Shalev-Shwartz and Tewari (2009)
- Modified quasi Newton: Andrew and Gao (2007); Yu et al. (2010)
- Hybrid: easy method first and then interior-point for faster local convergence (Shi et al., 2010)
Other Methods (Cont’d)

- Quadratic approximation followed by coordinate descent: Krishnapuram et al. (2005); Friedman et al. (2010); a kind of Newton approach
- Cutting plane method: Teo et al. (2010)
- Some methods find a solution path for different $C$ values; e.g., Rosset (2005), Zhao and Yu (2007), Park and Hastie (2007), and Keerthi and Shevade (2007).

Here we focus on a single $C$. 
Strengths and Weaknesses of Existing Methods

- **Convergence speed**: higher-order methods (quasi Newton or Newton) have fast local convergence, but fail to obtain a reasonable model quickly.

- **Implementation efforts**: higher-order methods usually more complicated.

- **Large data**: if solving linear systems is needed, use iterative (e.g., CG) instead of direct methods.

- **Feature correlation**: methods working on some variables at a time (e.g., decomposition methods) may be efficient if features are almost independent.
Outline

- Sparse Representation
- Existing Optimization Methods
- Coordinate Descent Methods
- Other Methods
- Experiments
Minimizing the one-variable function

\[ g_j(z) \equiv |w_j + z| - |w_j| + L(w + z e_j) - L(w), \]

where

\[ e_j \equiv [0, \ldots, 0, 1, 0, \ldots, 0]^T. \]

No closed form solution

Genkin et al. (2007), Shalev-Shwartz and Tewari (2009), and Yuan et al. (2010)
Coordinate Descent Methods II

They differ in how to minimize this one-variable problem.

While $g_j(z)$ is not differentiable, we can have a form similar to Taylor expansion:

$$g_j(z) = g_j(0) + g'_j(0)z + \frac{1}{2}g''_j(\eta z)z^2,$$

Another representation (for our derivation)

$$\min_z g_j(z) = |w_j + z| - |w_j| + L_j(z; w) - L_j(0; w),$$
where
\[ L_j(z; w) \equiv L(w + ze_j). \]
is a function of \( z \).
They rewrite $g_j(z)$ as

$$g_j(z) = g_j(0) + g'_j(0)z + \frac{1}{2}g''_j(\eta z)z^2,$$

where $0 < \eta < 1$,

$$g'_j(0) \equiv \begin{cases} 
L'_j(0) + 1 & \text{if } w_j > 0, \\
L'_j(0) - 1 & \text{if } w_j < 0,
\end{cases} \quad (1)$$

and

$$g''_j(\eta z) \equiv L''_j(\eta z).$$
BBR (Genkin et al., 2007) II

- $g_j(z)$ is not differentiable if $w_j = 0$
- BBR finds an upper bound $U_j$ of $g_j''(z)$ in a trust region:
  \[ U_j \geq g_j''(z), \quad \forall |z| \leq \Delta_j. \]
- Then $\hat{g}_j(z)$ is an upper-bound function of $g_j(z)$:
  \[ \hat{g}_j(z) \equiv g_j(0) + g_j'(0)z + \frac{1}{2}U_jz^2. \]
- Any step $z$ satisfying $\hat{g}_j(z) < \hat{g}_j(0)$ leads to
  \[ g_j(z) - g_j(0) = g_j(z) - \hat{g}_j(0) \leq \hat{g}_j(z) - \hat{g}_j(0) < 0, \]
BBR (Genkin et al., 2007) III

- Convergence not proved (no sufficient decrease condition via line search)
- Logistic loss

\[ U_j \equiv C \sum_{i=1}^{l} x_{ij}^2 F(y_i w^T x_i, \Delta_j |x_{ij}|), \]

where

\[ F(r, \delta) = \begin{cases} 
0.25 & \text{if } |r| \leq \delta, \\
\frac{1}{2 + e(|r| - \delta) + e(\delta - |r|)} & \text{otherwise.} 
\end{cases} \]
The sub-problem solved in practice:

$$\min_z \hat{g}_j(z)$$

s. t. $|z| \leq \Delta_j$ and $w_j + z \begin{cases} 
\geq 0 & \text{if } w_j > 0, \\
\leq 0 & \text{if } w_j < 0.
\end{cases}$

Update rule:

$$d = \min \left( \max \left( P \left( -\frac{g_j'(0)}{U_j} , w_j \right) , -\Delta_j \right) , \Delta_j \right),$$
where

\[ P(z, w) \equiv \begin{cases} 
  z & \text{if } \text{sgn}(w + z) = \text{sgn}(w), \\
  -w & \text{otherwise}. 
\end{cases} \]
SCD (Shalev-Shwartz and Tewari, 2009) I

- SCD: stochastic coordinate descent
- \( \mathbf{w} = \mathbf{w}^+ - \mathbf{w}^- \)
- At each step, randomly select a variable from
  \[ \{ w_1^+, \ldots, w_n^+, w_1^-, \ldots, w_n^- \} \]
- One-variable sub-problem:

\[
\min_z g_j(z) \equiv z + L_j(z; \mathbf{w}^+ - \mathbf{w}^-) - L_j(0; \mathbf{w}^+ - \mathbf{w}^-),
\]

subject to

\[
w_j^{k,+} + z \geq 0 \quad \text{or} \quad w_j^{k,-} + z \geq 0,
\]
SCD (Shalev-Shwartz and Tewari, 2009) II

- Second-order approximation similar to BBR:

\[ \hat{g}_j(z) = g_j(0) + g_j'(0)z + \frac{1}{2}U_jz^2, \]

where

\[ g_j'(0) = \begin{cases} 1 + L_j'(0) & \text{for } w_j^+ \\ 1 - L_j'(0) & \text{for } w_j^- \end{cases} \]

and \( U_j \geq g_j''(z), \forall z. \)

- BBR: \( U_j \) an upper bound of \( g_j''(z) \) in the trust region
- SCD: global upper bound
SCD (Shalev-Shwartz and Tewari, 2009)

- For logistic regression,
  \[
  U_j = 0.25C \sum_{i=1}^{l} x_{ij}^2 \geq g''_j(z), \forall z.
  \]

- Shalev-Shwartz and Tewari (2009) assume
  \(-1 \leq x_{ij} \leq 1, \forall i, j\), so a simple upper bound is
  \[
  U_j = 0.25Cl.
  \]
CDN (Yuan et al., 2010) I

- Newton step:

\[
\min_z \ g_j'(0)z + \frac{1}{2} g_j''(0)z^2.
\]

That is,

\[
\min_z \ |w_j + z| - |w_j| + L_j'(0)z + \frac{1}{2} L_j''(0)z^2.
\]

- Second-order term not replaced by an upper bound
- Function value may not be decreasing
- Tighter second-order term is useful (will see later)
CDN (Yuan et al., 2010) II

- Assume $z$ is the optimal solution; need line search
- Following Tseng and Yun (2007)

\[ g_j(\lambda z) - g_j(0) \leq \sigma \lambda (L'_j(0)z + |w_j + z| - |w_j|), \]

- This is slightly different from the traditional form of line search. Now

\[ |w_j + z| - |w_j| \]

must be taken into consideration
- Convergence can be proved
We have

\[ L'_j(0) = \left. \frac{dL(w + ze_j)}{dz} \right|_{z=0} = \nabla_j L(w) \]

\[ L''_j(0) = \left. \frac{d^2L(w + ze_j)}{dzdz} \right|_{z=0} = \nabla_{jj}^2 L(w) \]
Calculating First and Second Order Information II

- For logistic loss:

\[ L'_j(0) = C \sum_{i=1}^{l} y_i x_{ij} \left( \tau(y_i (w)^T x_i) - 1 \right), \]

\[ L''_j(0) = C \sum_{i=1}^{l} x_{ij}^2 \left( \tau(y_i (w)^T x_i) \right) \left( 1 - \tau(y_i (w)^T x_i) \right), \]

where

\[ \tau(s) \equiv \frac{1}{1 + e^{-s}}. \]
Outline

- Sparse Representation
- Existing Optimization Methods
- Coordinate Descent Methods
- Other Methods
- Experiments
GLMNET (Friedman et al., 2010)

- A quadratic approximation of $L(w)$:

$$f(w + d) - f(w) = (\|w + d\|_1 + L(w + d)) - (\|w\|_1 + L(w))$$

$$\approx \nabla L(w)^T d + \frac{1}{2} d^T \nabla^2 L(w) d + \|w + d\|_1 - \|w\|_1.$$

- Then

$$w \leftarrow w + d$$

- Line search is needed for convergence
GLMNET (Friedman et al., 2010) II

- But how to handle quadratic minimization with some one-norm terms?
- GLMNET uses coordinate descent
- For logistic regression:

\[
\nabla L(w) = C \sum_{i=1}^{l} \left( \tau(y_i w^T x_i) - 1 \right) y_i x_i
\]

\[
\nabla^2 L(w) = CX^T DX,
\]

where \( D \in R^{l\times l} \) is a diagonal matrix with

\[
D_{ii} = \tau(y_i w^T x_i) \left( 1 - \tau(y_i w^T x_i) \right)
\]
Bundle Methods (Teo et al., 2010)

- Also called cutting plane method
- $L(w)$ a convex loss function
- If $w^k$ the current solution,

$$L(w) \geq \nabla L(w^k)^T(w - w^k) + L(w^k)$$

$$= a_k^T w + b_k, \quad \forall w,$$

where

$$a_k \equiv \nabla L(w^k) \quad \text{and} \quad b_k \equiv L(w^k) - a_k^T w^k.$$
Maintains all earlier cutting planes to form a lower-bound function for $L(w)$:

$$L(w) \geq L^\text{CP}_k(w) \equiv \max_{1 \leq t \leq k} a_t^T w + b_t, \ \forall w.$$ 

Obtaining $w^{k+1}$ by solving

$$\min_w \|w\|_1 + L^\text{CP}_k(w).$$
Bundle Methods (Teo et al., 2010) III

- This is a linear program using \( w = w^+ - w^- \):

\[
\begin{align*}
\min_{w^+, w^-, \zeta} & \quad \sum_{j=1}^{n} w^+_j + \sum_{j=1}^{n} w^-_j + \zeta \\
\text{subject to} & \quad a_t^T (w^+ - w^-) + b_t \leq \zeta, \quad t = 1, \ldots, k, \\
& \quad w^+_j \geq 0, \quad w^-_j \geq 0, \quad j = 1, \ldots, n.
\end{align*}
\]
Outline

- Sparse Representation
- Existing Optimization Methods
- Coordinate Descent Methods
- Other Methods
- Experiments
**Data**

<table>
<thead>
<tr>
<th>Data set</th>
<th>$l$</th>
<th>$n$</th>
<th># of non-zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td>real-sim</td>
<td>72,309</td>
<td>20,958</td>
<td>3,709,083</td>
</tr>
<tr>
<td>news20</td>
<td>19,996</td>
<td>1,355,191</td>
<td>9,097,916</td>
</tr>
<tr>
<td>rcv1</td>
<td>677,399</td>
<td>47,236</td>
<td>49,556,258</td>
</tr>
<tr>
<td>yahoo-korea</td>
<td>460,554</td>
<td>3,052,939</td>
<td>156,436,656</td>
</tr>
</tbody>
</table>

- $l$: number of data, $n$: number of features
- They are all document sets
- 4/5 for training and 1/5 for testing
- Select best $C$ by cross validation on training
Compared Methods

Software using $\mathbf{w}^T \mathbf{x}$ without $b$

- BBR (Genkin et al., 2007)
- SCD (Shalev-Shwartz and Tewari, 2009)
- CDN: our coordinate descent implementation
- TRON: our Newton implementation for bound-constrained formulation
- OWL-QN (Andrew and Gao, 2007)
- BMRM (Teo et al., 2010)
Compared Methods (Cont’d)

Software using $\mathbf{w}^T \mathbf{x} + b$

- CDN: our coordinate descent implementation
- BBR (Genkin et al., 2007)
- CGD-GS (Yun and Toh, 2009)
- IPM (Koh et al., 2007)
- GLMNET (Friedman et al., 2010)
- Lassplore (Liu et al., 2009)
Convergence of Objective Values (no $b$)

- **real-sim**
- **news20**
- **rcv1**
- **yahoo-korea**
Test Accuracy

- real-sim
- news20
- rcv1
- yahoo-korea

Chih-Jen Lin (National Taiwan Univ.)
Convergence of Objective Values (with $b$)

Experiments

- real-sim
- news20
- rcv1
- yahoo-korea
Decomposition methods better in the early stage

One-variable sub-problem in coordinate descent
Use tight approximation if possible

Newton (IPM, GLMNET) and quasi Newton (OWL-QN): fast local convergence in the end

We also checked gradient and sparsity

Complete results (of more data sets) and programs are in Yuan et al. (2010); JMLR 2010 (11), 3183–3234


References II


References IV


