Trust Region Newton Method for Large-Scale Logistic Regression

Chih-Jen Lin
Department of Computer Science
National Taiwan University

Joint work with R. C. Weng and S. S. Keerthi
International Conference on Machine Learning, 2007
Large-scale Linear Classifiers

For areas such as document classification
- Usually linear classifiers as good as kernelized ones
  - Large \# of features
- Can solve larger problems than kernelized cases

Recently an active research topic
- [Keerthi and DeCoste, 2005]: modified Newton for L2-SVM
- [Joachims, 2006]: an approximation algorithm for L1-SVM
Regularized Logistic Regression

$$\min_w f(w) \equiv \frac{1}{2} w^T w + C \sum_{i=1}^{l} \log(1 + e^{-y_i w^T x_i})$$

L1-SVM: not differentiable

$$\frac{1}{2} w^T w + C \sum_{i=1}^{l} \max(0, 1 - y_i w^T x_i)$$

L2-SVM: differentiable, not twice differentiable

$$\frac{1}{2} w^T w + C \sum_{i=1}^{l} (\max(0, 1 - y_i w^T x_i))^2$$

Their performances usually similar, but logistic regression twice differentiable
Training Logistic Regression

Existing Methods

- Iterative scaling
  [Darroch and Ratcliff, 1972, Pietra et al., 1997, Goodman, 2002, Jin et al., 2003]
- Nonlinear conjugate gradient
- Limited memory quasi Newton
  [Liu and Nocedal, 1989, Benson and Moré, 2001]
- Truncated Newton
  [Komarek and Moore, 2005]

Limited memory quasi Newton considered the most efficient [Malouf, 2002]
Newton Method and Conjugate Gradient

- Newton direction $s^k$

$$\min_s \nabla f(w^k)^T s + \frac{1}{2} s^T \nabla^2 f(w^k) s$$

same as solving Newton linear system

$$\nabla^2 f(w^k) s = -\nabla f(w^k)$$

- Hessian matrix $\nabla^2 f(w^k)$ cannot be stored, but

$$\nabla^2 f(w) = \mathcal{I} + CX^TDX,$$

where $D$ diagonal, $X$: data, #instances $\times$ #feature

- CG: only Hessian-vector product needed

$$\nabla^2 f(w)s = s + C \cdot X^T(D(Xs))$$
Truncated Newton Methods

Inexactly solve the Newton linear system

- Also called inexact Newton method
- Early stopping of the conjugate gradient procedure
- An ad hoc way is in [Komarek and Moore, 2005]
  May not converge

Requirements:

- Not waste time in the beginning
- Accurate directions in the end
- Ensure convergence
A modern optimization technique; here we follow [Lin and Moré, 1999]

At each iteration, a size $\Delta_k$ of the trust region, and a quadratic model

$$q_k(s) = \nabla f(w^k)^T s + \frac{1}{2} s^T \nabla^2 f(w^k) s$$

as the approximation of $f(w^k + s) - f(w^k)$.

Solve trust region sub-problem

$$\min_s q_k(s) \quad \text{subject to } \|s\| \leq \Delta_k.$$ (1)
Trust Region Newton Method (Cont’d)

- Checking the reduction ratio
  \[
  \rho_k = \frac{f(w^k + s^k) - f(w^k)}{q_k(s^k)}
  \]

- Direction accepted if \( \rho_k \) large enough:
  \[
  w^{k+1} = \begin{cases} 
  w^k + s^k & \text{if } \rho_k > \eta_0, \\
  w^k & \text{if } \rho_k \leq \eta_0,
  \end{cases}
  \]

- \( \eta_0 > 0 \) is pre-specified

- Update \( \Delta_k \)
  - Reduced if reduction not good enough
  - Enlarged if reduction good enough
The Trust Region Algorithm

Given $w^0$.

For $k = 0, 1, \ldots$ (outer iterations)

- If $\nabla f(w^k) = 0$, stop.
- Find approximate solution $s^k$ of the trust region sub-problem

$$
\min_s q_k(s) \quad \text{subject to } \|s\| \leq \Delta_k
$$

- Compute reduction ratio $\rho_k$
- Update (or keep) $w^k$ to $w^{k+1}$
- Update $\Delta_{k+1}$
Trust Region Sub-Problem

- Approximately solve
  \[ \min_{\mathbf{s}} q_k(\mathbf{s}) \quad \text{subject to } \|\mathbf{s}\| \leq \Delta_k \]

- Note
  \[ q_k(\mathbf{s}) = \nabla f(\mathbf{w}^k)^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \nabla^2 f(\mathbf{w}^k) \mathbf{s} \]

  has minimum when
  \[ \nabla^2 f(\mathbf{w}^k) \mathbf{s} = -\nabla f(\mathbf{w}^k) \]

  i.e., the Newton linear system

- Additional condition
  \[ \|\mathbf{s}\| \leq \Delta_k \]
Trust Region Sub-Problem (Cont’d)

- Conjugate gradient procedure
  Inner iterations \( \{\bar{s}^1, \bar{s}^2, \ldots \} \)
- Output \( s^k \)
- Conjugate gradient stops if

\[
\| \nabla f(w^k) + \nabla^2 f(w^k) \bar{s}^i \| \leq 0.1 \| \nabla f(w^k) \|
\]

or \( \bar{s}^i \) reaches the boundary
- A careful design to ensure convergence
  Details not shown
## Experiments: Data

<table>
<thead>
<tr>
<th>Problem</th>
<th>#instances</th>
<th>#features</th>
<th># nonzeros</th>
</tr>
</thead>
<tbody>
<tr>
<td>a9a</td>
<td>32,561</td>
<td>123</td>
<td>451,592</td>
</tr>
<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>citeeseer</td>
<td>181,395</td>
<td>105,354</td>
<td>512,267</td>
</tr>
<tr>
<td>yahoo-japan</td>
<td>176,203</td>
<td>832,026</td>
<td>23,506,415</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>
Comparisons

TRON: trust region Newton method

SVMLIN [Keerthi and DeCoste, 2005]
- A Newton method for L2-SVM
- Newton linear system exactly solved

LBFGS: [Liu and Nocedal, 1989]
- Limited memory quasi Newton

SVM$^{perf}$ [Joachims, 2006]
- An approximation algorithm for L1-SVM
## Comparisons: Results

<table>
<thead>
<tr>
<th>Problem</th>
<th>Logistic regression</th>
<th>L2-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>TRON</td>
</tr>
<tr>
<td></td>
<td>CV</td>
<td>Time</td>
</tr>
<tr>
<td>rcv1</td>
<td>0.25</td>
<td>97.18%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>97.56%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>97.72%</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>97.69%</td>
</tr>
<tr>
<td>yahoo-korea</td>
<td>0.25</td>
<td>81.34%</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>84.03%</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>85.75%</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>86.40%</td>
</tr>
</tbody>
</table>

**CV**: Cross Validation Accuracy; Time in seconds
Comparisons: TRON (blue solid), $SVM^{perf}$ (red dotted)

Accuracy difference between the current model and that of optimum (real-sim and rcv1)
Conclusions

- Trust region Newton method very effective for large logistic regression
- Possible extensions:
  - L1-regularized logistic regression
  - Maximum entropy model, condition random fields
- A new package LIBLR released in April 2007
  http://www.csie.ntu.edu.tw/~cjlin/liblr
- All sources used for experiments are also available