Boosting with Structural Sparsity

John Duchi*  
UC Berkeley  

Yoram Singer  
Google

*Much of this work done at Google
Outline of talk

- Boosting and coordinate descent
- Losses and mixed norms
- Theory
- Experiments
- Future work and open issues
Coordinate Descent

- Popularized by Friedman, Hastie, Tibshirani
- Add features that do well on a loss function

$$L : \{ \mathcal{X} \times \{-1, +1\} \} \times \mathbb{R}^n \to \mathbb{R}$$

$$(j, \delta) = \operatorname*{argmin}_{j,\delta} \sum_{i=1}^{m} L(x_i, y_i, w + \delta e_j)$$

$$w' = w + \delta e_j$$
What we give

- Family of coordinate descent methods
- Multiclass, binary, regression
- Mixed-norm regularization
• Multiclass logistic loss

\[ L(W) = \sum_{i=1}^{m} \log \left( 1 + \sum_{r \neq y_i} \exp (w_r \cdot x_i - w_{y_i} \cdot x_i) \right) \]

\[ W = [w_1 \ w_2 \ \cdots \ w_k] \]
Mixed-norm regularization

- Goal: group sparsity in rows

\[ \| W \|_{\ell_1/\ell_p} = \sum_{j=1}^{n} \| w_j \|_p \]

\[ W = \begin{bmatrix}
    w_1^T \\
    w_2^T \\
    \vdots \\
    w_n^T
\end{bmatrix} \]
Example

• Contour of $\|w\|_2$
Example

- Contour of $\|w\|_2$
- Subgradients of $\|w\|_2$ at $w = 0$

$$\{ z : \|z\|_2 \leq 1 \}$$
Example

- Contour of $\|w\|_2$
- Subgradients of $\|w\|_2$ at $w = 0$

\[ \{ z : \|z\|_2 \leq 1 \} \]

- So if $\|\nabla L(0)\|_2 \leq \lambda$
Example

- Contour of $\|w\|_2$
- Subgradients of $\|w\|_2$ at $w = 0$
  \[
  \{ z : \|z\|_2 \leq 1 \}
  \]
- So if $\|\nabla L(0)\|_2 \leq \lambda$
  \[
  0 = \arg\min_w L(w) + \lambda \|w\|_2
  \]
Add regularization

- Multiclass logistic loss

\[ L(W) = \sum_{i=1}^{m} \log \left( 1 + \sum_{r \neq y_i} \exp(w_r \cdot x_i - w_{y_i} \cdot x_i) \right) \]

\[ + \lambda \sum_{j=1}^{n} \| \overrightarrow{w_j} \|_p \]
Loss Bounds (Multiclass)

• Idea:

\[ L(w + \delta e_j) \leq L(w) + \nabla L(w) \cdot e_j \delta + \frac{1}{2} \delta e_j \cdot D e_j \delta \]
Bound Illustration

\[ \log(1 + \exp(-x)) \]
Bound Illustration

\[ \log(2) - \frac{d}{1 + e} + \frac{1}{8}d^2 \]

\[ \log(1 + \exp(-x)) \]
Bound Illustration (3D)
Loss Bounds (Multiclass)

- Idea:
  
  \[ L(w + \delta e_j) \leq L(w) + \nabla L(w) \cdot e_j \delta + \frac{1}{2} \delta e_j \cdot D e_j \delta \]
Loss Bounds (Multiclass)

• Idea:

\[ L(w + \delta e_j) \leq L(w) + \nabla L(w) \cdot e_j \delta + \frac{1}{2} \delta e_j \cdot De_j \delta \]

• Bound:

\[
g_{r,j} = \frac{\partial}{\partial w_{r,j}} L(W), \quad \delta = \begin{bmatrix} \delta_1 & 0 & \cdots & \delta_k \\ 0 & \delta_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \delta_k \end{bmatrix}
\]

\[ L(W + \delta) \leq L(W) + \sum_{r=1}^{k} g_{r,j} \delta_r + \frac{1}{4} \sum_{r=1}^{k} \delta_r^2 x_{i,j} \]
The point of bounds

- Add regularization: easy minimization problems, easy updates

\[
\begin{align*}
\text{minimize} & \quad \sum_r \left( g_{r,j} - \frac{w^t_{r,j}}{2a_j} \right) w_r + \sum_{r=1}^k \frac{1}{4a_j} w_r^2 + \lambda \|w\|_p \\
\text{subject to} & \quad a_j = 1 / \sum_i x_{i,j}^2, \quad \delta_r = w_r - w^t_{r,j}
\end{align*}
\]
Update example

Multiclass GradBoost with $\ell_2$

\[ a_j = 1/\sum_i x_{i,j}^2, \quad g_j = \left[ \frac{\partial}{\partial w_{1,j}} L(W) \cdots \frac{\partial}{\partial w_{k,j}} L(W) \right]^T \]
Update example

Multiclass GradBoost with $\ell_2$

$$
\bar{w}_j^{t+1} = (\bar{w}_j^t - 2a_j g_j) \left[ 1 - \frac{2a_j \lambda}{\| \bar{w}_j^t - 2a_j g_j \|_2} \right] +
$$

$$
a_j = 1/\sum_i x_{i,j}^2, \quad g_j = \left[ \frac{\partial}{\partial w_{1,j}} L(W) \cdots \frac{\partial}{\partial w_{k,j}} L(W) \right]^T
$$
Update Benefits

$$
\overline{w}_{j}^{t+1} = (\overline{w}_{j}^{t} - 2a_j g_j) \left[ 1 - \frac{2a_j \lambda}{\|w_j^{t} - 2a_j g_j\|_2} \right] +
$$
Update Benefits

\[
\overline{w}_{j}^{t+1} = (\overline{w}_{j}^{t} - 2a_{j}g_{j}) \left[ 1 - \frac{2a_{j}\lambda}{\|\overline{w}_{j}^{t} - 2a_{j}g_{j}\|_2} \right] +
\]

- Totally corrective: feature pruning
Update Benefits

\[
\overline{w}_{j}^{t+1} = (\overline{w}_{j}^{t} - 2a_{j}g_{j}) \left[ 1 - \frac{2a_{j} \lambda}{\| \overline{w}_{j}^{t} - 2a_{j}g_{j} \|_2} \right] +
\]

- Totally corrective: feature pruning
Update Benefits

\[
\overline{w}_{j}^{t+1} = (\overline{w}_{j}^{t} - 2a_{j}g_{j}) \left[ 1 - \frac{2a_{j}\lambda}{\|\overline{w}_{j}^{t} - 2a_{j}g_{j}\|_2} \right] + \\
\]

- Totally corrective: feature pruning

- Feature scoring:
  \[
  \left( \frac{\left(\|g_{j}\|_2 - \lambda\right)}{\sum_{i} x_{i,j}} \right)^2
  \]
Theorem: If number of features/base hypotheses is finite, all presented algorithms converge to optimum of their respective losses.
Experiments

• Binary classification
• Multiclass classification
• Runtime properties
Binary Classification

MCAT Error Rates

MCAT Loss Rates

Dataset: Reuters RCV1, Lewis et al. 2004
Multiclass Classification

Dataset: StatLog Satellite
Dataset: PenDigit (UCI ML Repository)
Runtime Behavior

Error rate vs. train time

Non-zeros vs. train time
Concluding Remarks
Concluding Remarks

• Boosting variants for multiclass/binary/regression
Concluding Remarks

- Boosting variants for multiclass/binary/regression
- Prune features, perform induction, provably convergent
Concluding Remarks

• Boosting variants for multiclass/binary/regression
• Prune features, perform induction, provably convergent
• Further work:
Concluding Remarks

• Boosting variants for multiclass/binary/ regression

• Prune features, perform induction, provably convergent

• Further work:
  • Convergence Rates (Shalev-Shwartz and Tewari 2009)
Concluding Remarks

• Boosting variants for multiclass/binary/regression

• Prune features, perform induction, provably convergent

• Further work:
  • Convergence Rates (Shalev-Shwartz and Tewari 2009)
  • Consistency and generalization (Schapire et al. 1998; Zhang and Yu 2005)
Thanks very much!

• Any questions?