Accurate Max-Margin Training for Structured Output Spaces

Sunita Sarawagi        Rahul Gupta
IIT Bombay

http://www.cse.iitb.ac.in/~{sunita,grahul}
Structured learning

**Score of a prediction** $y$ for input $x$:
- $s(x, y) = w \cdot f(x, y)$
- **Predict**: $y_\star = \text{argmax}_y s(x, y)$
  - Exploit decomposability of feature functions
    - $f(x, y) = \sum_d f(x, y_d, d)$

**Feature function vector**

$f(x, y) = f_1(x, y), f_2(x, y), \ldots, f_K(x, y)$,
Training structured models

- Given
  - N input output pairs \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\)

- Error of output: \(E_i(y)\)
  - Also decomposes over smaller parts: \(E_i(y) = \sum_c E_{i,c}(y_c)\)
  - Example: \(\text{Hamming}(y_i, y) = \sum_c [y_{i,c} \neq y_c]\)

- Find \(w\)
  - Small training error
  - Generalizes to unseen instances
  - Efficient for structured models
Related work: max-margin training of structured models


- Several others…
Max-margin formulations

- Margin scaling

\[
\begin{align*}
\text{s.t. } w : f(x_i; y_i) & \geq E_i(y) + w : f(x_i; y) & \quad & \forall y \notin y_i; \forall i \\
\end{align*}
\]
Max-margin formulations

- Margin scaling
  
  \[
  \min_w \left\{ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \right\}_i \\
  \text{s.t.: } w : f(x_i; y_i) \geq E_i(y) + w : f(x_i; y) \quad \forall y \notin y_i; \forall i \\
  \right\}_i \geq 0 \quad \forall i
  \]

- Slack scaling
  
  \[
  \min_w \left\{ \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \right\}_i \\
  \text{s.t.: } w : f(x_i; y_i) \geq 1 + w : f(x_i; y) \ \frac{\right\}_i}{E_i(y)} \quad \forall y \notin y_i; \forall i \\
  \right\}_i \geq 0 \quad \forall i
  \]

Exponential number of constraints ➔ Use cutting plane
Margin Vs Slack scaling

- **Margin**
  - Easy inference of most violated constraint for decomposable $f$ and $E$
    
    $$
y^M = \arg\max_y (w : f(x_i; y) + E_i(y))$$

    - Too much importance to $y$ far from margin

- **Slack**
  - Difficult inference of violated constraint
    
    $$
y^S = \arg\max_y (w : f(x_i; y) i \frac{\gamma_i}{E_i(y)})$$

    - Zero loss of everything outside margin of 1
Accuracy comparison

Slack scaling up to 25% better than Margin scaling.
Approximating Slack inference

- Slack inference: \( \max_y s(y) - \xi / E(y) \)
  - Decomposability of \( E(y) \) cannot be exploited.

- \(-\xi / E(y)\) is concave in \( E(y) \)

- Variational method to rewrite as linear function

\[
i \left( \frac{\xi}{E(y)} \right) = \min_{\tilde{s}, \theta} \frac{E(y)}{\tilde{s}, \theta} \cdot 2^\theta (\xi, \tilde{s})
\]
Approximating Slack inference

Now approximate the inference problem as:

\[
\max_y s(y) \frac{\overrightarrow{\mu}}{\mathbb{E}(y)} = \max_y \min_{\phi \geq 0} s(y) + \mathbb{E}(y) + 2^{p \overrightarrow{\mu}} \cdot \min_{\phi \geq 0} \max_y s(y) + \mathbb{E}(y) + 2^{p \overrightarrow{\mu}}
\]

Same tractable MAP as in Margin Scaling
Approximating slack inference

- Now approximate the inference problem as:

\[
\max_y \frac{s(y)}{E(y)} = \max \min \left( s(y) + E(y) - 2p \right) \quad \text{subject to} \quad s(y) \geq 0.
\]

Same tractable MAP as in margin scaling

Convex in \( \lambda \) \( \Rightarrow \) minimize using line search,

Bounded interval \([\lambda_l, \lambda_u]\) exists since only want violating \( y \).
ApproxSlack gives the accuracy gains of Slack scaling while requiring same the MAP inference same as Margin scaling.
Limitation of ApproxSlack

- Cannot ensure that a violating \( y \) will be found even if it exists
  - No \( \lambda \) can ensure that.

Proof:
- \( s(y_1) = -1/2 \) \hspace{1cm} \( E(y_1) = 1 \)
- \( s(y_2) = -13/18 \) \hspace{1cm} \( E(y_2) = 2 \)
- \( s(y_3) = -5/6 \) \hspace{1cm} \( E(y_3) = 3 \)
- \( s(\text{correct}) = 0 \)
- \( \xi = 19/36 \)
- \( y_2 \) has highest \( s(y) - \xi/E(y) \) and is violating.
- No \( \lambda \) can score \( y_2 \) higher than both \( y_1 \) and \( y_2 \)
Max-margin formulations

- Margin scaling

\[
\min_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \\
\text{s.t. } w : f(x_i, y_i) \geq E_i(y) + w : f(x_i, y) \xi_i, \quad \xi_i \geq 0 \quad \forall i
\]

- Slack scaling

\[
\min_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \\
\text{s.t. } w : f(x_i, y_i) \geq 1 + w : f(x_i, y) \xi_i \frac{\xi_i}{E_i(y)}, \quad \xi_i \geq 0 \quad \forall i
\]
The pitfalls of a single shared slack variables

- Inadequate coverage for decomposable losses

Non-separable: \( y_2 = [0 \ 0 \ 1] \)
Separable: \( y_1 = [0 \ 1 \ 0] \)
Correct: \( y_0 = [0 \ 0 \ 0] \)

Marginal/Slack loss = 1.
Since \( y_2 \) non-separable from \( y_0 \), \( \xi = 1 \), Terminate.
Premature since different features may be involved.
A new loss function: PosLearn

- Ensure margin at each loss position

\[\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max(0, 1 + \mathbf{w} \cdot \mathbf{f}(\mathbf{x}_i; y) - \mathbf{y}_i \cdot \mathbf{f}(\mathbf{x}_i; y)) \]

- Compare with slack scaling.

\[\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max(0, 1 + \mathbf{w} \cdot \mathbf{f}(\mathbf{x}_i; y) - \mathbf{y}_i \cdot \mathbf{f}(\mathbf{x}_i; y)) \]

\[\mathbf{y}_i \cdot \mathbf{f}(\mathbf{x}_i; \mathbf{y}) \geq \sum_{c=1}^{C} \mathbf{1}_{\mathbf{y}_i, c} \mathbf{E}_{i,c}(\mathbf{y}_c) \geq \mathbf{y}_i \cdot \mathbf{y}_c, \quad 8y : y_c \notin \mathbf{y}_i, \mathbf{y}_c \]

\[\mathbf{y}_i \cdot \mathbf{f}(\mathbf{x}_i; \mathbf{y}) \geq \sum_{c=1}^{C} \mathbf{1}_{\mathbf{y}_i, c} \mathbf{E}_{i,c}(\mathbf{y}_c) \geq \mathbf{y}_i \cdot \mathbf{y}_c, \quad 8y : y_c \notin \mathbf{y}_i, \mathbf{y}_c \]

\[\mathbf{y}_i \cdot \mathbf{f}(\mathbf{x}_i; \mathbf{y}) \geq \sum_{c=1}^{C} \mathbf{1}_{\mathbf{y}_i, c} \mathbf{E}_{i,c}(\mathbf{y}_c) \geq \mathbf{y}_i \cdot \mathbf{y}_c, \quad 8y : y_c \notin \mathbf{y}_i, \mathbf{y}_c \]

\[\mathbf{y}_i \cdot \mathbf{f}(\mathbf{x}_i; \mathbf{y}) \geq \sum_{c=1}^{C} \mathbf{1}_{\mathbf{y}_i, c} \mathbf{E}_{i,c}(\mathbf{y}_c) \geq \mathbf{y}_i \cdot \mathbf{y}_c, \quad 8y : y_c \notin \mathbf{y}_i, \mathbf{y}_c \]
The pitfalls of a single shared slack variables

- Inadequate coverage for decomposable losses

Non-separable: \( y_2 = [0 \ 0 \ 1] \)
Separable: \( y_1 = [0 \ 1 \ 0] \)
Correct: \( y_0 = [0 \ 0 \ 0] \)

Margin/Slack loss = 1.
Since \( y_2 \) non-separable from \( y_0 \), \( \xi = 1 \), Terminate.
Premature since different features may be involved.

PosLearn loss = 2
Will continue to optimize for \( y_1 \) even after slack \( \xi = 1 \) becomes 1
Comparing loss functions

Sequence labeling

<table>
<thead>
<tr>
<th></th>
<th>Margin</th>
<th>Slack</th>
<th>ApproxSlack</th>
<th>PosLearn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoNLL03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Segmentation

<table>
<thead>
<tr>
<th></th>
<th>Margin</th>
<th>Slack</th>
<th>ApproxSlack</th>
<th>PosLearn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cora</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PosLearn: same or better than Slack and ApproxSlack
Inference for PosLearn QP

- **Cutting plane inference**
  - For each position $c$, find best $y$ that is wrong at $c$

\[
\max_{y: y \notin y_i ; c} s_i(y) \frac{\max_{y_c \in y_i ; c} E_{i;c}(y)}{\max_{y_c \notin y_i ; c} E_{i;c}(y)} = \max_{y_c \notin y_i ; c} s_i(y) \frac{\max_{y} \max_{y_c \in y_i ; c} E_{i;c}(y)}{\max_{y_c \notin y_i ; c} E_{i;c}(y)}
\]

- Solve simultaneously for all positions $c$
  - Markov models: Max-Marginals
  - Segmentation models: forward-backward passes
  - Parse trees

Small enumerable set

MAP with restriction, easy!
Margin scaling might take time with less data since good constraints may not be found early. PosLearn adds more constraints but needs fewer iterations.
Summary

1. Margin scaling popular due to computational reasons, but slack scaling more accurate
   - A variational approximation for slack inference

2. Single slack variable inadequate for structured models where errors are additive
   - A new loss function that ensures margin at each possible error position of $y$

Future work: theoretical analysis of generalizability of loss functions for structured models
Questions?
Slack scaling: which constraint?

\[ y^S = \text{argmax}_y (w : f(x_i ; y) i \ \frac{\|i}{E_i(y)}) \] Primal

Versus

\[ y^T = \text{argmax}_y E_i(y)(1 i \ w : \pm f(x_i ; y)) \] Original

- \( y^S \) better coordinate ascent direction than \( y^T \) for the dual
  - \( \Rightarrow \) faster convergence
Is slack scaling the best there is?

- \( Y = \text{Vector } y_1, y_2, \ldots, y_n \)
- \( E(y) = \sum_i E(y_i) \)

Two cases:

- \( Y_i \) s independent
  - Margin scaling is exactly what we want!
  - Slack scaling puts too little margin

- \( Y_i \) s dependent
  - Margin scaling asks for more margin than needed
  - Slack scaling better but
Max-margin loss surrogates

True error $E_i(\text{argmax}_y w: f(x_i; y))$

Let $w: \pm f(x_i; y) = w: f(x_i; y_i) \pm w: f(x_i; y)$

1. Margin Loss
   \[ \max_y [E_i(y) \pm f(x_i; y)]^+ \]

2. Slack Loss
   \[ \max_y E_i(y)[1 \pm f(x_i; y)]^+ \]
Approximating slack inference

- $-\xi/E(y)$ is concave in $E(y)$

- Variational method to rewrite as linear function

$$i \frac{\bar{\eta}}{E(y)} = \min_{\eta \geq 0} E(y) - 2^\eta (\bar{\eta})$$
When is margin scaling bad

Margin scaling adequate when
• No useful edge features: each position independent of each other

PosLearn useful when
• Edge features are important \( \Rightarrow \) truly structured models

<table>
<thead>
<tr>
<th></th>
<th>Margin</th>
<th>PosLearn</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Address</strong></td>
<td>Edge</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>No Edge</td>
<td>38</td>
</tr>
<tr>
<td><strong>Cora</strong></td>
<td>Edge</td>
<td>25.1</td>
</tr>
<tr>
<td></td>
<td>No Edge</td>
<td>56</td>
</tr>
<tr>
<td><strong>ConLL</strong></td>
<td>Edge</td>
<td>15.3</td>
</tr>
<tr>
<td></td>
<td>No Edge</td>
<td>19</td>
</tr>
</tbody>
</table>

F1 Span error