Implicit Learning of Simpler Output Kernels for Multi-label Prediction

Hanchen Xiong, Sandor Szedmak, Justus Piater
University of Innsbruck, Austria

Montreal, Canada. 2014.12.13
Support Vector Machines

\[
\begin{align*}
\min_{\mathbf{w} \in \mathbb{R}^H} & \quad \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \xi(i) \\
\text{s.t.} & \quad y(i) \left( \mathbf{w}^\top \phi(\mathbf{x}(i)) \right) \geq 1 - \xi(i), \xi(i) \geq 0, i \in \{1, \ldots, m\}
\end{align*}
\]

define a score function

\[
F(\mathbf{x}(i), y(i); \mathbf{w}) = y(i) \left( \mathbf{w}^\top \phi(\mathbf{x}(i)) \right)
\]

and output distance

\[
d(y(i), -y(i)) = |y(i) - (-y(i))| = 2
\]

\[
\min_{\mathbf{w} \in \mathbb{R}^H} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \max\{0, d(y(i), -y(i)) - \Delta_F(y(i), -y(i))\}
\]

\[
F \left( \mathbf{x}(i), y(i); \mathbf{w} \right) - F \left( \mathbf{x}(i), -y(i); \mathbf{w} \right) \geq d(y(i), -y(i)) - \xi(i)
\]

\[
\Delta_F(y(i), -y(i)) \geq 0
\]

\[
F(\mathbf{x}(i), y; \mathbf{w})
\]
Inter-Label Dependencies

Structural Outputs: $\mathcal{Y} = \{-1, +1\}^T$

$$y^{(i)} = [y_1^{(i)}, y_2^{(i)}, y_3^{(i)}, \ldots, y_d^{(i)}]^T$$

<table>
<thead>
<tr>
<th>tree</th>
<th>mountain</th>
<th>beach</th>
<th>sea</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1</td>
<td>+1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>-1</td>
<td>-1</td>
<td>+1</td>
<td>+1</td>
</tr>
</tbody>
</table>

Labels are inter-dependent: e.g. Markov network
Structural Support Vector Machine

\[ F(\mathbf{x}^{(i)}, y^{(i)}; \mathbf{w}) - F(\mathbf{x}^{(i)}, -y^{(i)}; \mathbf{w}) \geq d(y^{(i)}, -y^{(i)}) - \xi^{(i)} \]

\[ \Delta_F(y^{(i)}, -y^{(i)}) \]

generalized to structured outputs

\[ F(\mathbf{x}^{(i)}, y; \mathbf{w}) \]

\[ \Delta F(y^{(i)} \uparrow y', y') = F(\mathbf{x}^{(i)}, y^{(i)}; \mathbf{w}) - F(\mathbf{x}^{(i)}, y'; \mathbf{w}) \]

Structural SVM:

\[ \min_{\mathbf{W} \in \mathbb{R}^n} \frac{1}{2}\|\mathbf{W}\|^2 + C \sum_{i=1}^{m} \max_{y' \in \mathcal{Y}} \left\{ d(y^{(i)}, y') - \Delta_F(y^{(i)}, y') \right\} \]
SSVM for Multi-Label Prediction

Structural SVM

We define function as \( F(x^{(i)}, y^{(i)}; W) = \langle W, \phi(x^{(i)}) \otimes y^{(i)} \rangle_F \) and use Hamming distance on outputs:

because of linear decomposibility

\[
\min_{W \in \mathbb{R}^H} \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^{m} \max_{y' \in \mathcal{Y}} \left\{ d(y^{(i)}, y') - \Delta_F(y^{(i)}, y') \right\}
\]
Linear Score Function in Structural SVM

Structural SVM:
\[
\min_{\mathbf{w} \in \mathbb{R}^{n \times T}} \frac{1}{2} \| \mathbf{W} \|^2_F + C \sum_{i=1}^{m} \sum_{t=1}^{T} \max_{y_t' \in \{-1, +1\}} \left\{ d(y_t^{(i)}, y_t') - \Delta_F(y_t^{(i)}, y_t') \right\}
\]

\[
\min_{\mathbf{w}_1, \ldots, \mathbf{w}_T \in \mathbb{R}^{n \times T}} \sum_{t=1}^{T} \left\{ \frac{\| \mathbf{w}_t \|^2}{2} + C \sum_{i=1}^{m} \max \left\{ 0, d(y_t^{(i)}, -y_t^{(i)}) - \Delta_F(y_t^{(i)}, -y_t^{(i)}) \right\} \right\}
\]

Regular SVM:
\[
\min_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{m} \max \{ 0, d(y^{(i)}, -y^{(i)}) - \Delta_F(y^{(i)}, -y^{(i)}) \}
\]

linearly decomposable

Structural SVM → training $T$ SVMs jointly.
Joint SVM

training $T$ SVMs jointly:

$$\begin{align*}
\min_{w_1, w_2, \ldots, w_T} & \quad \frac{1}{2} \sum_{t=1}^{T} \|w_t\|^2 + C \sum_{t=1}^{T} \sum_{i=1}^{m} \xi_t^{(i)} \\
\text{w.r.t.} & \quad w_1, w_2, \ldots, w_T \in \mathbb{R}^{H_{\phi} \times 1} \\
\text{s.t.} & \quad \sum_{t=1}^{T} y_t^{(i)} (w_t^\top \phi(x^{(i)})) \geq T - \sum_{t=1}^{T} \xi_t^{(i)} \\
y^{(i)} &= [y_1^{(1)}, \ldots, y_T^{(i)}], \text{ and } W = [\frac{w_1^\top}{T}; \ldots; \frac{w_T^\top}{T}]^\top, \xi^{(i)} = \frac{\sum_{t=1}^{T} \xi_t^{(i)}}{T}
\end{align*}$$

Joint SVM:

$$\begin{align*}
\min_{W \in \mathbb{R}^{T \times H_{\phi}}} & \quad \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^{m} \bar{\xi}^{(i)} \\
\text{s.t.} & \quad \langle y^{(i)}, W\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}^{(i)}, \bar{\xi}^{(i)} \geq 0, i \in \{1, \ldots, m\}
\end{align*}$$

define linear kernels on structured outputs:

$$K_\psi(y^{(i)}, y^{(j)}) = \langle \psi(y^{(i)}), \psi(y^{(j)}) \rangle$$

$$\begin{align*}
\min_{W \in \mathbb{R}^{H_{\psi} \times H_{\phi}}} & \quad \frac{1}{2} \|W\|_F^2 + C \sum_{i=1}^{m} \bar{\xi}^{(i)} \\
\text{s.t.} & \quad \langle \psi(y^{(i)}), W\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}^{(i)}, \xi_i \geq 0, i \in \{1, \ldots, m\}
\end{align*}$$
Joint SVM: Complexity

Low Training Complexity:

Dual Joint SVM:

$$\max_{\alpha_1, \ldots, \alpha_m} \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j K_\psi(y^{(i)}, y^{(j)}) K_\phi(x^{(i)}, x^{(j)})$$

s.t. \( \forall i, \ 0 \leq \alpha_i \leq C \)

the same computational complexity as a single SVM

Dual SVM:

$$\max_{\alpha \in \mathbb{R}^m} \sum_{i=1}^{m} \alpha_i - \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} K_\phi(x^{(i)}, x^{(j)})$$

s.t. \( \ 0 < \alpha_i < C, \ i \in \{1, \ldots, m\} \)
When linear output is used:

$$\min_{\mathbf{w} \in \mathbb{R}^{H_{\psi} \times \mathcal{H}_\phi}} \left\{ \frac{1}{2} \| \mathbf{W} \|_F^2 + C \sum_{i=1}^{m} \bar{\xi}(i) \right\}$$

s.t. \( \langle \psi(\mathbf{y}^{(i)}), \mathbf{W}_\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}(i), \xi_i \geq 0, i \in \{1, \ldots, m\} \)

$$\psi^{Lin}(\mathbf{y}^{(i)}) = \mathbf{P} \mathbf{y}^{(i)} \quad K^{Lin}_{\psi}(\mathbf{y}^{(i)}, \mathbf{y}^{(j)}) = \mathbf{y}^{(i)\top} \mathbf{P}^\top \mathbf{P} \mathbf{y}^{(j)}$$

$$\mathbf{U} = \mathbf{P}^\top \mathbf{W}$$

$$\min_{\mathbf{w} \in \mathbb{R}^{H_{\psi} \times \mathcal{H}_\phi}} \left\{ \frac{1}{2} \| \mathbf{W} \|_F^2 + C \sum_{i=1}^{m} \bar{\xi}(i) \right\}$$

s.t. \( \langle \mathbf{y}^{(i)}, \mathbf{U}_\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}(i), \xi_i \geq 0, i \in \{1, \ldots, m\} \)

a new regularization: \( \frac{1}{2} \text{tr}(\mathbf{W}^\top \mathbf{P} \mathbf{P}^\top \mathbf{W}) \)

$$\min_{\mathbf{u} \in \mathbb{R}^{H_{\psi} \times \mathcal{H}_\phi}} \left\{ \frac{1}{2} \| \mathbf{U} \|_F^2 + C \sum_{i=1}^{m} \bar{\xi}(i) \right\}$$

s.t. \( \langle \mathbf{y}^{(i)}, \mathbf{U}_\phi(x^{(i)}) \rangle \geq 1 - \bar{\xi}(i), \xi_i \geq 0, i \in \{1, \ldots, m\} \)
Results and Comparisons

Benchmark Databases for Image Annotation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>labels</th>
<th>Number of training instances</th>
<th>Number of test instances</th>
<th>Average labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corel5k</td>
<td>260</td>
<td>4500</td>
<td>500</td>
<td>3.3965</td>
</tr>
<tr>
<td>Espgame</td>
<td>268</td>
<td>18689</td>
<td>2081</td>
<td>4.6859</td>
</tr>
<tr>
<td>Iaprtc12</td>
<td>291</td>
<td>17665</td>
<td>1962</td>
<td>5.7187</td>
</tr>
</tbody>
</table>
## Comparable Results:

<table>
<thead>
<tr>
<th>Method</th>
<th>Corel5K</th>
<th></th>
<th></th>
<th>Espgame</th>
<th></th>
<th></th>
<th>Iaprtc12</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P(%)</td>
<td>R(%)</td>
<td>F1(%)</td>
<td>P(%)</td>
<td>R(%)</td>
<td>F1(%)</td>
<td>P(%)</td>
<td>R(%)</td>
<td>F1(%)</td>
</tr>
<tr>
<td>MBRM</td>
<td>24.0</td>
<td>25.0</td>
<td>24.0</td>
<td>18.0</td>
<td>19.0</td>
<td>18.0</td>
<td>24.0</td>
<td>23.0</td>
<td>23.0</td>
</tr>
<tr>
<td>JEC</td>
<td>27.0</td>
<td>32.0</td>
<td>29.0</td>
<td>24.0</td>
<td>19.0</td>
<td>21.0</td>
<td>29.0</td>
<td>19.0</td>
<td>23.0</td>
</tr>
<tr>
<td>TagProp</td>
<td>33.0</td>
<td>42.0</td>
<td>37.0</td>
<td>39.0</td>
<td>27.0</td>
<td>32.0</td>
<td>45.0</td>
<td>34.0</td>
<td>39.0</td>
</tr>
<tr>
<td>FastTag</td>
<td>32.0</td>
<td>43.0</td>
<td>37.0</td>
<td>46.0</td>
<td>22.0</td>
<td>30.0</td>
<td>47.0</td>
<td>26.0</td>
<td>34.0</td>
</tr>
<tr>
<td>JSVM</td>
<td>48.5</td>
<td>38.0</td>
<td>42.6</td>
<td>32.7</td>
<td>31.6</td>
<td>32.2</td>
<td>42.2</td>
<td>29.4</td>
<td>34.6</td>
</tr>
<tr>
<td>JSVM+Pol(2)</td>
<td>46.6</td>
<td>37.0</td>
<td>41.3</td>
<td>32.6</td>
<td>24.4</td>
<td>27.9</td>
<td>37.9</td>
<td>26.6</td>
<td>31.2</td>
</tr>
<tr>
<td>JSVM+Pol(3)</td>
<td>41.5</td>
<td>31.3</td>
<td>35.7</td>
<td>28.5</td>
<td>21.3</td>
<td>24.4</td>
<td>38.0</td>
<td>26.1</td>
<td>31.0</td>
</tr>
</tbody>
</table>


Preference to Simpler Label-Dependence
A take-home message: A simpler output kernel is desirable to avoid overfitting in output structural dependencies.

Thanks for your attention!

Questions and Answers?
Joint SVM : Pre-Image

\[
\hat{y}^* = \arg \max_{y \in \{+1, -1\}} y^T \langle \psi(y), W \phi(\hat{x}) \rangle = \arg \max_{y \in \{+1, -1\}} \sum_{i=1}^{m} \alpha_i K_{\phi}(x^{(i)}, \hat{x}) K_{\psi}(y^{(i)}, y) 
\]

\[
\hat{y}^* = \left( \sum_{k=1}^{K} y^{(k)} w_k \right) / \sum_{k=1}^{K} w_k \quad w_j = \sum_{i=1}^{m} \alpha_i \beta_i K_{\psi}(y^{(i)}, y^{(j)})
\]
\[ \frac{1}{2} \text{tr}(W^\top PP^\top W) = \frac{1}{2} \text{tr}(PP^\top WW^\top) \]

**lemma:** for positive (semi-)definite matrices A and B:

\[ \text{tr}(AB)^m \leq \left\{ \text{tr}(A)^{2m} \text{tr}(B)^{2m} \right\}^{1/2} \]

where \( m \) is a positive integer.

\[ \frac{1}{2} \text{tr}(PP^\top WW^\top) \leq \frac{1}{2} \text{tr}(PP^\top) \text{tr}(WW^\top) = \frac{1}{2} \|P\|_F^2 \|W\|_F^2 \]

## Joint SVM: Pre-Image

### Multiple SVMs: train $T$ SVMs independently

- too expensive ($T$ can be very large)
- ignore inter-label dependencies 😞

<table>
<thead>
<tr>
<th></th>
<th>Training Time (sec)</th>
<th>Testing Time (sec)</th>
<th>Testing Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Precision (%)</td>
</tr>
<tr>
<td>Independent SVMs (Gau)</td>
<td>6285.11</td>
<td>117.20</td>
<td>15.3</td>
</tr>
<tr>
<td>Independent SVMs (Pol)</td>
<td>4612.23</td>
<td>147.9</td>
<td>15.1</td>
</tr>
<tr>
<td>Joint SVM (Gau)</td>
<td>80.68</td>
<td>6.92</td>
<td>40.8</td>
</tr>
<tr>
<td>Joint SVM (Pol)</td>
<td>76.48</td>
<td>9.11</td>
<td>48.5</td>
</tr>
</tbody>
</table>