Implicit Learning of Simper Output Kernels for Multi-label Prediction

Hanchen Xiong, Sandor Szedmak, Justus Piater
UNIVERSITY OF INNSBRUCK, AUSTRIA

OVERVIEW

Research in Multi-label Prediction:

- exploiting and utilizing inter-label dependencies; ☺
- increasingly more sophisticated dependency are used; ?
- "overfit" output structural dependencies when the desired ones are simpler ☹
- a regularization should be added on output structural dependencies ☑

Our contributions: Joint SVM

- joint SVM $\Longleftrightarrow$ structural SVM with linearly decomposable score functions;
- in joint SVM, (non-)linear kernels can be assigned on outputs to capture inter-label dependencies;
- joint SVM shares the same computational complexity as a single SVM;
- when linear output kernels are used, a output-kernel regularization is implicitly added.
- yield promising results on 3 image annotation databases.

FROM STRUCTURAL SVM TO JOINT SVM

Structural SVM is an extension of SVM for structured-outputs, in which, however, the margin to be maximized is defined as the score gap between the desired output and the first runner-up.

$$\arg\min_{\mathbf{W}\in\mathbb{R}^{\Psi}} \frac{1}{2}||\mathbf{W}||^2 + C\sum_{i=1}^{m}\max_{\mathbf{y}'\in\mathcal{Y}}\left\{d(\mathbf{y}^{(i)},\mathbf{y}') - \Delta_F(\mathbf{y}^{(i)},\mathbf{y}')\right\} \quad (1)$$

where $\Delta_F(\mathbf{y}^{(i)},\mathbf{y}') = F(\mathbf{x}^{(i)},\mathbf{y}^{(i)};\mathbf{W}) - F(\mathbf{x}^{(i)},\mathbf{y}';\mathbf{W})$, the score function is $F\left(\mathbf{x}^{(i)},\mathbf{y}^{(i)};\mathbf{W}\right) = \langle\mathbf{W},\phi(\mathbf{x}^{(i)})\otimes\mathbf{y}^{(i)}\rangle$, and *Hamming distance* $d(\mathbf{y}^{(i)},\mathbf{y}')$ is used on outputs. Because of linear decomposability, (1) can be rewritten as:

$$\begin{aligned}
&\arg\min_{\mathbf{W}\in\mathbb{R}^{\mathcal{H}_\phi\times\mathbb{R}^T}} \tfrac{1}{2}||\mathbf{W}||_F^2 + C\textstyle\sum_{i=1}^m\sum_{t=1}^T \max_{y'_t=\{-1,+1\}}\left\{d(y_t^{(i)},y'_t) - \Delta_F(y_t^{(i)},y'_t)\right\}\\
&\Downarrow\\
&\arg\min_{\mathbf{w}_1,\cdots,\mathbf{w}_T\in\mathbb{R}^{\mathcal{H}_\phi}} \textstyle\sum_{t=1}^T\left\{\tfrac{1}{2}||\mathbf{w}_t||^2 + C\sum_{i=1}^m \max\left\{0, d(y_t^{(i)},-y_t^{(i)}) - \Delta_F(y_t^{(i)},-y_t^{(i)})\right\}\right\} \quad (2)\\
&\Downarrow\\
&\arg\min_{\mathbf{w}_1,\cdots,\mathbf{w}_T\in\mathbb{R}^{\mathcal{H}_\phi}} \textstyle\sum_{t=1}^T\left\{\tfrac{1}{2}||\mathbf{w}_t||^2 + 2C\sum_{i=1}^m \max\left\{0, 1 - y_t^{(i)}\mathbf{w}_t^\top\phi(\mathbf{x}^{(i)})\right\}\right\}
\end{aligned}$$

where $\langle\cdot,\cdot\rangle_F$ denotes Frobenius product and $||\mathbf{W}||_F$ is the Frobenius norm of matrix $\mathbf{W}$.

It can be seen in (2) that,with linearly decomposable score functions and output distances, structural SVM on multi-label learning is equivalent to learning T SVMs jointly: Joint SVM

$$\begin{aligned}
&\arg\min_{\mathbf{W}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} && \tfrac{1}{2}||\mathbf{W}||_F^2 + C\textstyle\sum_{i=1}^m\bar{\xi}^{(i)}\\
&\text{s.t.} && \left\langle\psi(\mathbf{y}^{(i)}),\mathbf{W}\phi(x^{(i)})\right\rangle \ge 1-\bar{\xi}^{(i)}, \xi_i\ge 0, i\in\{1,\ldots,m\}
\end{aligned} \quad (3)$$

where $\mathbf{y}^{(i)} = [y_1^{(1)},\ldots,y_T^{(i)}]$, $\mathbf{W} = [\frac{\mathbf{w}_1^\top}{T};\ldots;\frac{\mathbf{w}_T^\top}{T}]^\top$. The corresponding dual form of Joint SVM is:

$$\begin{aligned}
&\arg\min_{\alpha_1,\cdots,\alpha_m} && \textstyle\sum_{i=1}^m\alpha_i - \sum_{i,j=1}^m\alpha_i\alpha_j K_\psi(\mathbf{y}^{(i)},\mathbf{y}^{(j)})K_\phi(\mathbf{x}^{(i)},\mathbf{x}^{(j)})\\
&\text{s.t} && \forall i, 0\le\alpha_i\le C
\end{aligned} \quad (4)$$

therefore, the same computational complexity as a single regular SVM.

IMPLICIT REGULARIZATION ON LINEAR OUTPUT KERNELS

When a linear output kernel is used to capture pairwise dependencies of labels, via which the output vectors can be linearly mapped as $\psi(\mathbf{y}) = \mathbf{P}\mathbf{y}$: $K_\psi^{Lin}(\mathbf{y}^{(i)},\mathbf{y}^{(j)}) = \mathbf{y}^{(i)\top}\mathbf{\Omega}\mathbf{y}^{(j)}$ where $\mathbf{\Omega} = \mathbf{P}^\top\mathbf{P}$. By denoting $\mathbf{U} = \mathbf{P}^\top\mathbf{W}$, we can have:

$$\begin{aligned}
&\arg\min_{\mathbf{W}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} && \tfrac{1}{2}||\mathbf{W}||_F^2 + C\textstyle\sum_{i=1}^m\bar{\xi}^{(i)}\\
&\text{s.t.} && \left\langle\mathbf{y}^{(i)},\mathbf{U}\phi(x^{(i)})\right\rangle \ge 1-\bar{\xi}^{(i)}, \xi_i\ge 0, i\in\{1,\ldots,m\}
\end{aligned} \quad (5)$$

we use a compact regularization for both $\mathbf{W}$ and $\mathbf{\Omega}$, $\frac{1}{2}||\mathbf{W}^\top\mathbf{\Omega}^\top\mathbf{W}||_F^2$, resulting in:

$$\begin{aligned}
&\arg\min_{\mathbf{U}\in\mathbb{R}^{H_\psi\times\mathcal{H}_\phi}} && \tfrac{1}{2}||\mathbf{U}||_F^2 + C\textstyle\sum_{i=1}^m\bar{\xi}^{(i)}\\
&\text{s.t.} && \left\langle\mathbf{y}^{(i)},\mathbf{U}\phi(x^{(i)})\right\rangle \ge 1-\bar{\xi}^{(i)}, \xi_i\ge 0, i\in\{1,\ldots,m\}
\end{aligned} \quad (6)$$

Remarkably, a linear output kernel is implicitly learned, and absorbed in $\mathbf{W}$, also a regularization on the output kernel is also implicitly added.

EXPERIMENTAL RESULTS AND COMPARISON

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

Table 1: Comparison between different versions of joint SVM and other related methods on three benchmark databases. P, R and F1 denote precision, recall and F1 measure respectively.