# Multiclass Multiple Kernel Learning

Alexander Zien and Cheng Soon Ong

June. 12th, 2007

# Review of Multiclass SVM

Existing works focus only on binary classification problem. This work propose an intuitive formulation of the multi-class SVM.

## mSVM

$$
\begin{aligned}
\textit{function}: &\quad f_{\mathbf{w},b} =< \mathbf{w}, \phi(\mathbf{x}, y) > + b_y \\
\textit{prediction}: &\quad x = \arg\max_{y \in \mathcal{Y}} f_{w,b}(\mathbf{x}, y) \\
\textit{objective}: &\quad \min_{\mathbf{w},\mathbf{b}} \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{n} \max_{u \neq y_i} \{\ell(f_{\mathbf{w},b}(\mathbf{x_i}, y_i) - f_{\mathbf{w},\mathbf{b}}(\mathbf{x_i}, u))\} \\
\textit{hinge loss}: &\quad \ell(t) := C \max(0, 1 - t)
\end{aligned}
$$

# Multiple Kernel Learning (MKL) Primal

Extend typical multiclass SVM to $p$ feature maps $\phi_k(\mathbf{x_i}, y_i)$:

$$f_{w,b,\beta}(\mathbf{x}, y) = \sum_{k=1}^{p} \beta_k < w_k, \phi_k(x, y) > + b_y$$

$$\underbrace{\sum_{k=1}^{p} \beta_k = 1, \beta_k \geq 0}_{\text{feature mapping (kernel) weight}}$$

**Intuitive Formulation**

$$\min_{\beta,w,b,\xi} \quad \frac{1}{2} \sum_{k=1}^{p} \beta_k ||w_k||^2 + \sum_{i=1}^{n} \xi_i$$

$$s.t. \quad \forall i : \xi_i = \ell(f_{w,b,\beta}(x_i, y_i) - f_{w,b,\beta}(x_i, u))$$

# MKL Primal 2

Interpretable and intuitive, but in general not being convex :(

Let $\mathbf{v}_k := \beta_k \mathbf{w}_k$, then

**MKL Primal**

$$\min_{\beta, v, b, \xi} \quad \frac{1}{2} \sum_{k=1}^{p} \frac{||v_k||^2}{\beta_k} + \sum_{i=1}^{n} \xi_i$$

$$s.t. \quad \forall i : \xi_i = \ell(< \mathbf{v}, \psi_{iu} > + b_{y_i} - b_u)$$

$$\psi_{kiu} = \phi_k(x_i, y_i) - \phi_k(x_i, u)$$

$$\psi_{iu} = (\psi_{kiu})_{k=1,\cdots,p}$$

(combine into just one vector)

This is Convex!

# MKL Primal 2

Interpretable and intuitive, but in general not being convex :(

Let $\mathbf{v}_k := \beta_k \mathbf{w}_k$, then

## MKL Primal

$$
\begin{aligned}
\min_{\beta, v, b, \xi} \quad & \frac{1}{2} \sum_{k=1}^{p} \frac{||v_k||^2}{\beta_k} + \sum_{i=1}^{n} \xi_i \\
s.t. \quad & \forall i : \xi_i = \ell(< \mathbf{v}, \psi_{iu} > + b_{y_i} - b_u) \\
& \psi_{kiu} = \phi_k(x_i, y_i) - \phi_k(x_i, u) \\
& \psi_{iu} = (\psi_{kiu})_{k=1,\cdots,p} \\
& \text{(combine into just one vector)}
\end{aligned}
$$

This is Convex!

# MKL Primal 2

Interpretable and intuitive, but in general not being convex :(

Let $\mathbf{v}_k := \beta_k \mathbf{w}_k$, then

## MKL Primal

$$
\begin{aligned}
\min_{\beta, v, b, \xi} \quad & \frac{1}{2} \sum_{k=1}^{p} \frac{||v_k||^2}{\beta_k} + \sum_{i=1}^{n} \xi_i \\
s.t. \quad & \forall i : \xi_i = \ell(< \mathbf{v}, \psi_{iu} > + b_{y_i} - b_u) \\
& \psi_{kiu} = \phi_k(x_i, y_i) - \phi_k(x_i, u) \\
& \psi_{iu} = (\psi_{kiu})_{k=1,\cdots,p} \\
& \text{(combine into just one vector)}
\end{aligned}
$$

This is Convex!

# Dual formulation for general loss function

$$\begin{aligned}
min_{\eta,\widetilde{\alpha},\gamma} \quad & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\
s.t. \quad & \forall k : \quad \gamma \geq \frac{1}{2} \sum_{i,j} \sum_{u \neq y_i} \sum_{v \neq y_j} \widetilde{\alpha}_{iu} \widetilde{\alpha}_{jv} < \psi_{kiu}, \psi_{kjv} > \\
& \forall i : \quad \forall u \neq y_i : 0 \leq \eta_{iu}, \\
& \forall i : \quad 1 = \sum_{u \neq y_i} \eta_{iu} \\
& \forall v : \quad 0 = \sum_i (1 - \delta_{y_i}, v) \widetilde{\alpha}_{iv} - \sum_i \delta_{y_i,v} \sum_{u \neq y_i} \widetilde{\alpha}_{iu}
\end{aligned}$$

Here $\ell^*$ is the conjugate function of the loss function $\ell$.

# Langrange

$$t_{iu} = \sum_k < v_k, \psi_{kiu} > + b_{y_i} - b_u \quad \text{score difference}$$

$$\xi_i \geq \ell(t_{iu}) \quad \text{Error upper bound}$$

So the Lagrangian is

$$\begin{aligned} L \quad = \quad & \frac{1}{2}\sum_k \frac{1}{\beta_k}||v_k||^2 + \sum_i \xi_i + \gamma\left(\sum_k \beta_k - 1\right) \\ & - \sum_k \epsilon_k \beta_k + \sum_i \sum_{u \neq y_i} \eta_{iu}(\ell(t_{iu}) - \xi_i) \\ & + \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu}\left(t_{iu} - \sum_k < \mathbf{v}_k, \psi_{kiu} > - b_{y_i} + b_u\right) \end{aligned}$$

$$\frac{\partial L}{\partial \mathbf{v}_k} = \frac{1}{\beta_k}\mathbf{v}_k - \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu}\psi_{kiu} = 0$$

$$\Longrightarrow \qquad \mathbf{w}_k = \frac{1}{\beta_k}\mathbf{v}_k = \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu}\psi_{kiu}$$

$$\max_{\eta,\widetilde{\alpha},\gamma} \min_{\mathbf{t}} \qquad \sum_i \sum_{u \neq y_i} (\eta_{iu}\ell(t_i u) + \widetilde{\alpha}_{iu} t_{iu}) - \gamma$$

$$s.t. \qquad \forall k : \gamma \geq \frac{1}{2}||\mathbf{w}_k||^2 \ \ \text{(Obtained from } \frac{\partial L}{\partial \beta_k} = 0)$$

$$\forall i : 1 = \sum_{u \neq y_i} \eta_{iu} \quad \text{(Obtained from } \frac{\partial L}{\partial \xi_i} = 0)$$

$$\forall v : 0 = \sum_i (1 - \delta_{y_i}, v)\widetilde{\alpha}_{iv} - \sum_i \delta_{y_i,v} \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \quad (\frac{\partial L}{\partial b_v} = 0)$$

$$
\begin{aligned}
& \max_{\eta,\widetilde{\alpha},\gamma} \min_{\mathbf{t}} && \sum_i \sum_{u\neq y_i} (\eta_{iu}\ell(t_i u) + \widetilde{\alpha}_{iu} t_{iu}) - \gamma \\
= \quad & \max_{\eta,\widetilde{\alpha},\gamma} \min_{\mathbf{t}} && \sum_i \sum_{u\neq y_i} \eta_{iu} \left( \ell(t_i u) + \frac{\widetilde{\alpha}_{iu} t_{iu}}{\eta_{iu}} \right) - \gamma \\
= \quad & \max_{\eta,\widetilde{\alpha},\gamma} \min_{\mathbf{t}} && \sum_i \sum_{u\neq y_i} -\eta_{iu} \left( -\ell(t_i u) - \frac{\widetilde{\alpha}_{iu} t_{iu}}{\eta_{iu}} \right) - \gamma \\
= \quad & \max_{\eta,\widetilde{\alpha},\gamma} && \sum_i \sum_{u\neq y_i} -\eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) - \gamma \\
\Longleftrightarrow \quad & \min_{\eta,\widetilde{\alpha},\gamma} && \gamma + \sum_i \sum_{u\neq y_i} \eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right)
\end{aligned}
$$

The definition of conjugate function of $f$ is defined as

$$
f^*(y) = sup_{x \in dom(f)}(y^T x - f(x))
$$

# Dual Formulation for General Loss

$$
\begin{aligned}
min_{\gamma,\widetilde{\alpha},\gamma} \quad & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\
s.t. \quad & \forall k: \quad \gamma \geq \frac{1}{2} \sum_{i,j} \sum_{u \neq y_i} \sum_{v \neq y_j} \widetilde{\alpha}_{iu} \widetilde{\alpha}_{jv} < \psi_{kiu}, \psi_{kjv} > \\
& \forall i: \quad \forall u \neq y_i : 0 \leq \eta_{iu}, \\
& \forall i: \quad 1 = \sum_{u \neq y_i} \eta_{iu} \\
& \forall v: \quad 0 = \sum_i (1 - \delta_{y_i,v}) \widetilde{\alpha}_{iv} - \sum_i \delta_{y_i,v} \sum_{u \neq y_i} \widetilde{\alpha}_{iu}
\end{aligned}
$$

# Dual Formulation for Hinge Loss

When the loss $\ell(t) = C \cdot max(0, 1-t)$,

$$\ell^*(\nu) = \begin{cases} \nu & -C \leq \nu \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

So in order to make the dual solvable, we must have

$$-C \leq -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \leq 0$$

$$\begin{aligned} As \quad & \forall i: \quad 1 = \sum_{u \neq y_i} \eta_{iu} \\ and \quad & \forall i: \quad \forall u \neq y_i : 0 \leq \eta_{iu} \\ \text{We have} \quad & \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \leq C \\ & \forall u \neq y_i : 0 \leq \widetilde{\alpha}_{iu} \end{aligned}$$

$$\begin{aligned} & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\ = \quad & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\ = \quad & \gamma - \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \end{aligned}$$

# Dual Formulation for Hinge Loss

When the loss $\ell(t) = C \cdot max(0, 1 - t)$,

$$\ell^*(\nu) = \begin{cases} \nu & -C \leq \nu \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

So in order to make the dual solvable, we must have

$$-C \leq -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \leq 0$$

$$\begin{aligned} \textit{As} \quad & \forall i: \quad 1 = \sum_{u \neq y_i} \eta_{iu} \\ \textit{and} \quad & \forall i: \quad \forall u \neq y_i : 0 \leq \eta_{iu} \\ \text{We have} \quad & \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \leq C \\ & \forall u \neq y_i : 0 \leq \widetilde{\alpha}_{iu} \end{aligned}$$

$$\begin{aligned} & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\ = \; & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\ = \; & \gamma - \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \end{aligned}$$

# Dual Formulation for Hinge Loss

When the loss $\ell(t) = C \cdot max(0, 1-t)$,

$$\ell^*(\nu) = \begin{cases} \nu & -C \leq \nu \leq 0 \\ \infty & \text{otherwise} \end{cases}$$

So in order to make the dual solvable, we must have

$$-C \leq -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \leq 0$$

$$\begin{aligned} As \quad & \forall i: \quad 1 = \sum_{u \neq y_i} \eta_{iu} \\ and \quad & \forall i: \quad \forall u \neq y_i : 0 \leq \eta_{iu} \\ \text{We have} \quad & \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \leq C \\ & \forall u \neq y_i : 0 \leq \widetilde{\alpha}_{iu} \end{aligned}$$

$$\begin{aligned} & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \ell^* \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\ = \; & \gamma + \sum_i \sum_{u \neq y_i} \eta_{iu} \left( -\frac{\widetilde{\alpha}_{iu}}{\eta_{iu}} \right) \\ = \; & \gamma - \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \end{aligned}$$

## Dual for Hinge Loss (Folded formulation)

$$
\begin{aligned}
\min_{\widetilde{\alpha},\gamma} \quad & \gamma - \sum_i \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \\
s.t. \quad & \forall k : \gamma \geq \frac{1}{2} ||w_k(\widetilde{\alpha})||^2 \\
& \sum_{u \neq y_i} \widetilde{\alpha}_{iu} \leq C \\
& \forall u \neq y_i : 0 \leq \widetilde{\alpha}_{iu} \\
& \forall v : \quad 0 = \sum_i (1 - \delta_{y_i}, v) \widetilde{\alpha}_{iv} - \sum_i \delta_{y_i, v} \sum_{u \neq y_i} \widetilde{\alpha}_{iu}
\end{aligned}
$$

$$\alpha_{iu} = \begin{cases} -\widetilde{\alpha}_{iu} & \text{if } u \neq y_i \\ \sum_{v \neq y_i} \widetilde{\alpha}_{iv} & \text{if } u = y_i \end{cases}$$

## Unfolded Formulation

$$\min_{\alpha,\gamma} \quad \gamma - \sum_i \alpha_{i,y_i}$$

$$s.t. \quad \forall k : \gamma \geq \frac{1}{2}||w_k(\alpha)||^2$$

$$\alpha \in \mathcal{S} := \left\{ \alpha \;\middle|\; \begin{array}{l} \forall i : 0 \leq \alpha_{i,y_i} \leq C \\ \forall i : \forall u \neq y_i : \alpha_{iu} \leq 0 \\ \forall i : \sum_{u \in \mathcal{Y}} \alpha_{iu} = 0 \\ \forall u \in \mathcal{Y} : \sum_i \alpha_{iu} = 0 \end{array} \right\}$$

## SILP formulation

$$\max_{\beta,\theta} \quad \theta$$

$$s.t. \quad \forall \alpha \in S : \theta \leq \frac{1}{2} \sum_k \beta_k ||w_k(\alpha)||^2 - \sum_i \alpha_{i,y_i}$$

- Both version of formulation are QCQPs.
- When there's only one single kernel, it reduces to the dual of mSVM.
- When there are only 2 classes, the formulation reduces to Gert's formulation of kernel selection
- Unfolded formulation can be easily transformed into semi-infinite linear programming.

$$\min_{\beta,\mathbf{w},\mathbf{b},\xi,s} \quad \frac{1}{2}\sum_{k=1}^{p}\beta_k\|\mathbf{w}_k\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \quad \xi_i = \max_{u\neq y_i} s_{iu}, \qquad s_{iu} \geqslant 0$$
$$\sum_{k=1}^{p}\beta_k\langle\mathbf{w}_k,\Psi_{kiu}\rangle + b_{y_i} - b_u \geqslant 1 - s_{iu}$$

Intuitive MKL: Equation (3); Non-Convex

substitute $\mathbf{v}_k := \beta_k\mathbf{w}_k$

$$\min_{\beta,\mathbf{v},\mathbf{b},\xi,s} \quad \frac{1}{2}\sum_{k=1}^{p}\frac{1}{\beta_k}\|\mathbf{v}_k\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \quad \xi_i \geqslant s_{iu}, \qquad s_{iu} \geqslant 0$$
$$\sum_{k=1}^{p}\langle\mathbf{v}_k,\Psi_{kiu}\rangle + b_{y_i} - b_u \geqslant 1 - s_{iu}$$

Equation (4); Convex

$$\min_{\beta,\mathbf{w},\mathbf{b},\xi,s} \quad \frac{1}{2}\left(\sum_{k=1}^{p}\beta_k\|\mathbf{w}_k\|\right)^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \quad \xi_i = \max_{u\neq y_i} s_{iu}, \qquad s_{iu} \geqslant 0$$
$$\sum_{k=1}^{p}\beta_k\langle\mathbf{w}_k,\Psi_{kiu}\rangle + b_{y_i} - b_u \geqslant 1 - s_{iu}$$

Generalized from Sonnenburg et. al. [20]; Non-Convex

substitute $\mathbf{v}_k := \beta_k\mathbf{w}_k$

$$\min_{\mathbf{v},\mathbf{b},\xi,s} \quad \frac{1}{2}\left(\sum_{k=1}^{p}\|\mathbf{v}_k\|\right)^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \quad \xi_i = \max_{u\neq y_i} s_{iu}, \qquad s_{iu} \geqslant 0$$
$$\sum_{k=1}^{p}\langle\mathbf{v}_k,\Psi_{kiu}\rangle + b_{y_i} - b_u \geqslant 1 - s_{iu}$$

Generalized from Bach et. al. [1]; Convex

Theorem 1 [1]

$$\min_{\alpha} \quad \gamma - \sum_i \alpha_{iy_i}$$
$$\text{s.t.} \quad \forall i: 0 \leqslant \alpha_{iy_i} \leqslant C, \qquad \forall i: \forall u \neq y_i: \alpha_{iu} \leqslant 0$$
$$\forall i: \sum_{u\in\mathcal{Y}}\alpha_{iu} = 0 \qquad \forall u: \sum_i \alpha_{iu} = 0$$
$$\forall k: \gamma \geqslant \frac{1}{2}\sum_{i,j,u,v}\alpha_{iu}\alpha_{jv}\langle\Phi_k(\mathbf{x}_i,u),\Phi_k(\mathbf{x}_j,v)\rangle$$

Common Lagrange Dual; Equation (9)

# Time Complexity Comparison

| | Examples | Classes | Kernels |
|---|---|---|---|
| QP, unfolded | 2.5 | 1.8 | – |
| QCQP, unfolded (9) | 3.0 | 2.0 | 2.3 |
| SILP, unfolded (11) | 2.4 | 1.7 | 1.1 |
| SILP, compact | 2.6 | 2.2 | 1.0 |

# Performance

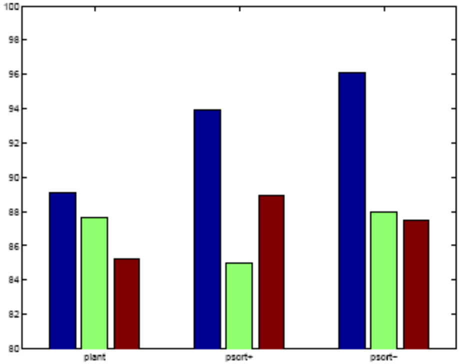


Figure 3: Protein Subcellular Localization results. The bars within each group correspond to different methods: (left, blue) MKL; (center, green) unweighted sum of kernels; (right, red) current state of the art.

# Comments

1. Contribution: Generalize the kernel selection problem to any kinds of convex loss functions and multiple classes.
2. If just focus on multi-class SVM, deriving from the dual formulation is much easier.
3. Possible interesting direction: kernel selection in $y$ space, kernel selection for structured output.
4. Whether or not all the classes should share the same feature space (Mentioned in the paper, but has been done by us).
5. Can we remove the weight $\beta$ in the regularizer? Different formulation and variants might need more understanding.