A General and *Efficient* Multiple Kernel Learning Algorithm

Sören Sonnenburg
Fraunhofer Institute FIRST
Kekuléstr. 7
12489 Berlin
Germany
http://ida.first.fhg.de/~sonne

Gunnar Rätsch
Friedrich Miescher Laboratory
Max Planck Society
Spemannstr. 39
Tübingen, Germany
www.fml.tuebingen.mpg.de/raetsch/projects/mkl_silp

Christin Schäfer
Fraunhofer Institute FIRST
Kekuléstr. 7
12489 Berlin
Germany
http://ida.first.fhg.de/~christin

Abstract

While classical kernel-based learning algorithms are based on a single kernel, in practice it is often desirable to use multiple kernels. Lankriet et al. (2004) considered conic combinations of kernel matrices for classification, leading to a convex quadratically constraint quadratic program. We show that it can be rewritten as a semi-infinite linear program that can be efficiently solved by recycling the standard SVM implementations. Moreover, we generalize the formulation and our method to a larger class of problems, including regression and one-class classification. Experimental results show that the proposed algorithm helps for automatic model selection, improving the interpretability of the learning result and works for hundred thousands of examples or hundreds of kernels to be combined.

This file contains the appendix of our submission to NIPS 2005, which had to be removed due to space constraints.

A Derivation of the MKL Dual for Generic Loss Functions

Conic Primal From the MKL primal problem (P), one can derive the following *equivalent* second order cone problem, where $\mathcal{K}_D = \{(\mathbf{x}, c) \in \mathbb{R}^D \times \mathbb{R},\ \|\mathbf{x}\|_2 \leq c\}$ is the second-order cone of order D.

$$
\begin{array}{llll}
 & \min_{\mathbf{w},u,t} & \frac{1}{2}u^2 + \sum_{i=1}^{N} L(f(\mathbf{x}_i), y_i) \\
(P_{\text{cone}}) & \text{w.r.t.}: & u \in \mathbb{R},\ t \in \mathbb{R}, (\mathbf{w}_k, t) \in \mathcal{K}_{k_k},\ \forall k = 1 \dots K \\
 & \text{s.t.}: & f(\mathbf{x}_i) = \sum_{k=1}^{K} \Phi_k(\mathbf{x}_i)\mathbf{w}_k + b,\ \forall i = 1 \dots N \\
 & & \sum_{k=1}^{K} t_k \leq u
\end{array}
$$

Introducing Lagrange multipliers $\gamma \in \mathbb{R}^+$, $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $(\boldsymbol{\lambda}_k, \mu_k) \in (\mathcal{K}_{\|})^* = \mathcal{K}_d$ living on the self dual cone $\mathcal{K}$, the conic Lagrangian is given as

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}, \gamma, \boldsymbol{\lambda}, \boldsymbol{\mu}) \quad = \quad & \frac{1}{2}u^2 + \sum_{i=1}^{N} L(f(\mathbf{x}_i), y_i) - \sum_{i=1}^{N} \alpha_i f(\mathbf{x}_i) + \sum_{i=1}^{N} \alpha_i \sum_{k=1}^{K} \Phi_k(\mathbf{x}_i)\mathbf{w}_k + \\
& + \gamma \left(\sum_{k=1}^{K} t_k - u \right) - \sum_{k=1}^{K} \left(\Phi_k(\boldsymbol{\lambda})^{\mathrm{T}} \mathbf{w}_k + \mu_k t_k \right).
\end{aligned}
$$

The derivatives of the Lagrangian w.r.t. the primal variables, $u, \mathbf{w}, t$ have to vanish. We therefore get the following constraints

$$
\begin{aligned}
\partial_u L &= u - \gamma \Rightarrow \gamma = u \\
\partial_{\mathbf{w}_k} L &= \sum_{i=1}^{N} \alpha_i \Phi_k(\mathbf{x}_i) - \Phi_k(\boldsymbol{\lambda}) \ \Rightarrow\ \sum_{i=1}^{N} \alpha_i \Phi_k(\mathbf{x}_i) = \Phi_k(\boldsymbol{\lambda}) \\
\partial_{t_k} L &= \gamma 1 - \mu_k \ \Rightarrow\ \gamma = \mu_k \\
\partial_{f(\mathbf{x}_i)} L &= L'(f(\mathbf{x}_i), y_i) - \alpha_i \Rightarrow f(\mathbf{x}_i) = L'^{-1}(\alpha_i, y_i).
\end{aligned}
$$

There L' is the derivative of the loss function and L'^{-1} is the inverse of the L' for which L is required to be strictly convex and differentiable. We thus obtain the following dual function

$$
\begin{aligned}
D(\alpha, \gamma, \boldsymbol{\lambda}, \mu) \quad = \quad & -\frac{1}{2}\gamma^2 + \sum_{i=1}^{N} L(L'^{-1}(\alpha_i, y_i), y_i) - \sum_{i=1}^{N} \alpha_i L'^{-1}(\alpha_i, y_i) + \\
& + \sum_{i=1}^{N} \alpha_i \sum_{k=1}^{K} \Phi_k(\mathbf{x}_i)^{\mathrm{T}} \mathbf{w}_k - \sum_{i=1}^{N} \alpha_i \sum_{k=1}^{K} \Phi_k(\mathbf{x}_i)^{\mathrm{T}} \mathbf{w}_k \\
= \quad & -\frac{1}{2}\gamma^2 + \sum_{i=1}^{N} L(L'^{-1}(\alpha_i, y_i), y_i) - \sum_{i=1}^{N} \alpha_i L'^{-1}(\alpha_i, y_i)
\end{aligned}
$$

subject to the constraints $\gamma \geq 0$ and $\left\| \sum_{i=1}^{N} \alpha_i \Phi_k(\mathbf{x}_i) \right\|_2 \leq \gamma,\ \forall k = 1 \dots K$.

This leads to:

$$
\begin{aligned}
\max_{\gamma,\boldsymbol{\alpha}} \quad & -\frac{1}{2}\gamma^2 + \sum_{i=1}^{N} L(L'^{-1}(\alpha_i, y_i), y_i) - \sum_{i=1}^{N} \alpha_i L'^{-1}(\alpha_i, y_i) \\
\text{w.r.t. :} \quad & \gamma \in \mathbb{R},\ \boldsymbol{\alpha} \in R^N \\
\text{s.t. :} \quad & \gamma \geq 0 \\
& \left\| \sum_{i=1}^{N} \alpha_i \Phi_k(\mathbf{x}_i) \right\|_2 \leq \gamma,\ \forall k = 1 \ldots K
\end{aligned}
$$

Applying $(.)^2$ to the latter constraint, multiplying by $\frac{1}{2}$, relabeling $\frac{1}{2}\gamma^2 \mapsto \gamma$ we obtain the MKL dual for arbitrary strictly convex loss functions ($\gamma \geq 0$ is fulfilled implicitly).

$$
\begin{aligned}
\min_{\gamma,\boldsymbol{\alpha}} \quad & \gamma - \sum_{i=1}^{N} L(L'^{-1}(\alpha_i, y_i), y_i) + \sum_{i=1}^{N} \alpha_i L'^{-1}(\alpha_i, y_i) \\
(D_{\text{cone}})\ \text{w.r.t. :} \quad & \gamma \in \mathbb{R},\ \boldsymbol{\alpha} \in R^N \\
\text{s.t. :} \quad & \frac{1}{2} \left\| \sum_{i=1}^{N} \alpha_i \Phi_k(\mathbf{x}_i) \right\|_2^2 \leq \gamma,\ \forall k = 1 \ldots K.
\end{aligned}
$$