Supplementary Material

Paper: On p-norm Path Following in Multiple Kernel Learning for Non-linear Feature Selection

Paper ID: 128

Abstract

Here we present: i) the proof of Theorem 3 stated in the main paper, ii) the proof of Theorem 4 stated in the main paper, iii) a lemma, analogous to Lemma 1 stated in the main paper, that holds for the case of generalized KL-divergence, iv) Formulation for the second order derivative along the solution path, and v) Tables corresponding to Table 2 and Table 3 in the main paper, displaying mean accuracies with standard deviations.

We follow the notation described in the submission. This text may refer to equations/theorems/lemmas in the original submission using the appropriate numbers therein. To avoid any confusion, the equations/theorems/lemmas appearing in this text are numbered using a prefix '0.'. We begin with the proof of Theorem 3:

Proof of Theorem 3

Eliminating α's from (7) results in the following optimization problem

$$\min_{\eta \geq 0} \ \frac{1}{2}\mathbf{y}^\top Q_\eta^{-1}\mathbf{y} + \lambda_2 \sum_{i=1}^{r} \eta_i^p$$

The above problem is convex and KKT conditions for its optimal solution are

$$\mathbf{G}_i(\eta, p) \equiv -\frac{1}{2}\mathbf{y}^\top Q_\eta^{-1} K_i Q_\eta^{-1}\mathbf{y} + \lambda_2 p \eta_i^{p-1} = 0 \tag{0.1}$$

where $i = 1, \ldots, r$. Since the base kernels are orthogonal with unit rank and unit trace, the term Q_η^{-1} in the above KKT condition can be simplified.

Next, we compute the derivate $\frac{\mathbf{d}\eta_i}{\mathbf{d}p}$ along the solution path of optimal η's by employing: $\mathrm{d}\mathbf{G}_i(\eta, p) = \sum_j \frac{\partial \mathbf{G}_i}{\partial \eta_j}\mathrm{d}\eta_j + \frac{\partial \mathbf{G}_i}{\partial p}\mathrm{d}p = 0$. We get

$$\frac{\mathbf{d}\eta_i}{\mathbf{d}p} = -\frac{\eta_i \left(\eta_i + \frac{1}{2\lambda_1}\right)[1 + p\ln(\eta_i)]}{p\left[(p-1)\left(\eta_i + \frac{1}{2\lambda_1}\right) + 2\eta_i\right]} \tag{0.2}$$

The proof follows from observing the sign of the above derivative expression and from the fact that $e^{-\frac{1}{p}}$ is a monotonically increasing function of p.

Proof of Theorem 4

The KKT conditions for optimality for the convex optimization problem (7) are:

$$\mathbf{G}_i(\eta, p) \equiv -\frac{1}{2}\mathbf{y}^\top Q_\eta^{-1} K_i Q_\eta^{-1}\mathbf{y} + \lambda_2 p \eta_i^{p-1} = 0 \tag{0.3}$$

where $Q_\eta = \sum_i \eta_i K_i + \frac{I}{2\lambda_1}$ and $i = 1, \ldots, r$. In the following, we first obtain a particular $p' (> 1)$ where $\eta_1^*(p') = 0$. Next, from (0.3), we show that η_1^* is zero only at $p = p'$.

Let $\eta_1^* = 0$ at $p = p'$. From the KKT conditions (0.3) corresponding to $\mathbf{G}_1(\eta, p)$, we get $\mathbf{y}^\top Q_\eta^{-1} K_1 Q_\eta^{-1} \mathbf{y} = 0$ (as $\eta_1^* = 0$). Employing the Sherman-Morrison formula for computing Q_η^{-1} and simplifying the L.H.S. of the above expression yields

$$\eta_2^* = \frac{\mathbf{y}^\top K_1 \mathbf{y}}{2\lambda_1(\mathbf{y}^\top K_2 K_1 \mathbf{y} - \mathbf{y}^\top K_1 \mathbf{y})} \tag{0.4}$$

Since we have $\mathbf{y}^\top K_2 K_1 \mathbf{y} > \mathbf{y}^\top K_1 \mathbf{y} > 0$, it follows $\eta_2^* > 0$. Similarly, from the KKT conditions (0.3) corresponding to $\mathbf{G}_2(\eta, p)$, we obtain

$$p'(\eta_2^*)^{p'+1} = \frac{1}{2\lambda_2} \left(\frac{\mathbf{y}^\top K_1 \mathbf{y} \mathbf{y}^\top K_2 \mathbf{y}}{\mathbf{y}^\top K_2 K_1 K_2 \mathbf{y}} \right) \tag{0.5}$$

Note that $\mathbf{y}^\top K_2 K_1 \mathbf{y} > 0 \Rightarrow \mathbf{y}^\top K_2 K_1 K_2 \mathbf{y} > 0$. Now, consider the following values of parameters (λ_1, λ_2):

$$\lambda_1 = \frac{\mathbf{y}^\top K_1 \mathbf{y}}{2(\mathbf{y}^\top K_2 K_1 \mathbf{y} - \mathbf{y}^\top K_1 \mathbf{y})}, \;\; \lambda_2 = \frac{\mathbf{y}^\top K_1 \mathbf{y} \mathbf{y}^\top K_2 \mathbf{y}}{3\mathbf{y}^\top K_2 K_1 K_2 \mathbf{y}}$$

Note that both $\lambda_1, \lambda_2 > 0$ in the above case. Employing these in (0.4) and (0.5), we obtain the optimal kernel weight (η_1^*, η_2^*) at $p' = 1.5$ as $(0, 1)$. Moreover, with the above mentioned values of parameters (λ_1, λ_2), it also follows that at optimality, $(\eta_1^* = 0) \Rightarrow (p' = 1.5)$ and $(\eta_1^* = 0) \Rightarrow (\eta_2^* = 1)$. Hence, it follows that $\eta_1^* > 0$ at any $1 < p < p'$.

Lemma for generalized KL-divergence

In the case of generalized KL-divergence as the Bregman divergence in (1), using the optimality conditions for the non-trivial case (5) and Theorem 2, the following lemma is immediate:

Lemma 1. *For any p, the deviation in the objective value of (1) obtained using the approximate path following algorithm from the true optimal objective is upper bounded by* $r(\lambda_1 \epsilon + \lambda_2 (p-1) \epsilon^p)$.

Second order derivative along the solution path

Here, we derive the second order derivative along the optimal solution path of (1), which can be employed in Algorithm 1, assuming that the function F is thrice differentiable. From (6), we have the following form: $\frac{\mathbf{d}\eta_i^*(p)}{\mathbf{d}p} = f(\eta_i^*(p), p)$ where f is a function corresponding to the R.H.S. of (6). Hence, employing the total derivative formula in the above equation, we get the following formulation for the second order derivative:

$$\begin{aligned}
\frac{\mathbf{d}^2 \eta_i^*(p)}{\mathbf{d}p^2} = -\frac{\lambda_2}{D} \Bigg[& \frac{\lambda_1}{\lambda_2} F'''(\eta_i^*(p)) + p(p-1)(p-2)\eta_i^*(p)^{p-3} \left(\frac{\mathbf{d}\eta_i^*(p)}{\mathbf{d}p} \right)^2 \\
& + 2\eta_i^*(p)^{p-2} \left(2p - 1 + p^2 \ln(\eta_i^*(p)) - p \ln(\eta_i^*(p)) \right) \frac{\mathbf{d}\eta_i^*(p)}{\mathbf{d}p} \\
& + \eta_i^*(p)^{p-1} \ln(\eta_i^*(p)) \left(2 + p \ln(\eta_i^*(p)) \right) \Bigg]
\end{aligned}$$

where $D = \lambda_1 F'''(\eta_i^*(p)) + \lambda_2 p(p-1)\eta_i^*(p)^{p-2}$ and the term $\mathbf{d}\eta_i^*(p)/\mathbf{d}p$ can be obtained from (6).

Experimental Results

In this section, we report the tables corresponding to Table 2 and 3 of the main paper, with mean accuracies and standard deviations. Table 1 corresponds to Table 2 of the main paper while Table 2 corresponds to Table 3 of the main paper.

Table 1: The maximum classification accuracy achieved (mean and standard deviations) along the feature selection path. Generalized l_p-KTA achieves significantly higher accuracies as compared to state-of-the-art KTA and $l_{p\geq 1}$-MKL formulations as well as leading feature selection techniques such as BAHSIC. The table reports mean and standard deviations results averaged over 5-fold cross validation. '-' denote results where the data set was too large for the feature selection algorithm to generate results.

	Arcene	Madelon	Relathe	Pcmac	Basehock	Dorothea
Gen l_p-KTA	$\mathbf{92.00} \pm 5.70$	$\mathbf{65.70} \pm 0.99$	$\mathbf{92.57} \pm 0.30$	$\mathbf{93.62} \pm 1.67$	$\mathbf{98.59} \pm 0.58$	$\mathbf{94.75} \pm 1.30$
Centered-KTA	75.00 ± 9.35	62.45 ± 2.07	90.40 ± 0.21	93.05 ± 1.44	97.29 ± 1.05	-
SMO-MKL	82.00 ± 5.70	62.05 ± 0.54	-	-	-	-
BAHSIC	69.00 ± 6.52	53.90 ± 2.97	85.07 ± 1.42	89.55 ± 0.22	93.58 ± 1.38	90.63 ± 0.77
PF-l_1-MKL	81.00 ± 6.52	62.76 ± 2.40	85.67 ± 1.90	-	-	-
PF-l_1-SVM	77.00 ± 12.04	61.25 ± 1.08	89.00 ± 2.16	90.68 ± 0.59	97.24 ± 0.89	93.88 ± 1.65
Uniform	81.00 ± 6.52	59.85 ± 0.84	90.96 ± 0.77	92.49 ± 0.64	97.99 ± 0.59	91.38 ± 1.42

Table 2: The maximum classification accuracy achieved on the ASU data sets (with the corresponding number of selected features) along the feature selection path. In keeping with the ASU experimental protocol, all algorithms are restricted to selecting at most 200 features and are allowed to train on only half the data. Generalized l_p-KTA (RBF) outperforms all the linear techniques and this demonstrates the advantages of non-linear feature selection. Amongst the linear methods, our proposed method with linear features is the best in general.

	Arcene	Madelon	Relathe	Pcmac	Basehock	Dorothea
Gen l_p-KTA (RBF)	76.80 ± 9.25	64.50 ± 1.14	89.40 ± 0.93	89.76 ± 0.87	95.46 ± 1.02	93.75 ± 1.18
Gen l_p-KTA (Linear)	$\mathbf{73.40} \pm 7.06$	62.04 ± 0.97	$\mathbf{88.39} \pm 0.87$	88.88 ± 2.61	94.76 ± 0.96	$\mathbf{93.60} \pm 1.30$
Inf. Gain	72.00 ± 5.89	61.63 ± 0.95	84.39 ± 0.94	$\mathbf{88.99} \pm 1.22$	95.26 ± 1.29	93.33 ± 0.97
Chi-Square	71.20 ± 7.90	61.69 ± 1.28	83.48 ± 0.80	88.24 ± 1.39	$\mathbf{95.28} \pm 1.26$	93.33 ± 1.34
Fisher Score	66.20 ± 9.68	61.47 ± 1.04	83.35 ± 1.05	88.02 ± 1.60	94.61 ± 1.47	93.30 ± 1.32
mRMR	68.20 ± 7.33	61.87 ± 1.17	75.01 ± 1.01	83.34 ± 1.18	88.88 ± 1.00	93.18 ± 1.35
ReliefF	68.40 ± 7.71	$\mathbf{62.06} \pm 1.21$	77.08 ± 2.73	80.76 ± 1.79	86.05 ± 3.55	93.33 ± 0.93
Spectrum	64.00 ± 6.60	60.19 ± 0.76	69.99 ± 1.90	66.74 ± 1.52	69.79 ± 1.31	90.28 ± 1.25
Gini Index	64.60 ± 5.50	59.43 ± 2.38	69.50 ± 2.09	66.60 ± 1.37	69.49 ± 1.35	90.28 ± 1.25
K.-Wallis	60.20 ± 10.26	55.04 ± 1.23	70.97 ± 2.02	65.20 ± 1.31	70.37 ± 1.05	90.08 ± 1.07