A Binary Classification Framework for Two-Stage Multiple Kernel Learning
Supplementary Material

1. Theoretical Results

This is the expanded version of Section 3 from the main paper that includes complete proofs.

In this section we make the connection between the performance a K-classifier in the K-space and the performance on the original problem precise. Thus justifying the approach taken in this paper not only intuitively, but also from a theoretical standpoint. Specifically, we bound the generalization error of an SVM that uses the kernel induced by a K-classifier in terms of the expected hinge loss and the margin of the K-classifier in the K-space:

Theorem 1.1 *Let P be a distribution on $\mathcal{X} \times \{\pm 1\}$, $z_{xx'}$ and $t_{yy'}$ be as in Equation **??**, h be a K-classifier, and R be a constant s.t. $h(z_{xx}) \leq R^2 \ \forall x \in \mathcal{X}$. Let*

$$HL_{h,\gamma} = E_{((x,y),(x',y')) \in P \times P} \left[\!\left[\left[1 - \frac{t_{yy'} h(z_{xx'})}{\gamma} \right]_+ \right]\!\right]$$

be the expected K-space hinge loss relative to margin γ of the K-classifier h. Then, with probability $1-\delta$, a classifier $\hat{f}$ with generalization error

$$P_{(x,y)} \left[\!\left[y\hat{f}(x) \leq 0 \right]\!\right] \leq HL_{h,\gamma} + \mathcal{O}\left(\sqrt{\frac{R^4 \ln(1/\delta)}{\gamma^2 n}} \right)$$

can be learned efficiently from a training sample of n instances drawn IID from P.

The theorem follows from the two lemmas stated below. The first lemma shows that a K-classifier that has low expected hinge loss in the K-space will induce a "good" kernel. The second lemma shows that a good kernel allows for a classifier with low generalization error to be efficiently learned from a finite training sample. The following definition states formally what we mean by a good kernel (Srebro, 2007).[1]

Definition A kernel K is an (ϵ, γ) good kernel in hinge loss with respect to a distribution P on $\mathcal{X} \times \{\pm 1\}$ if there exist a classifier $w \in \mathcal{H}_K$ with $\|w\|_{\mathcal{H}_K} = 1$ s.t.

$$E_{(x,y)} \left[\!\left[\left[1 - \frac{y\langle w, \phi(x) \rangle}{\gamma} \right]_+ \right]\!\right] \leq \epsilon$$

where $\mathcal{H}_K$ is the Hilbert space and $\phi(\cdot)$ is the feature mapping corresponding to K.

Lemma 1.2 *Let P, h, $HL_{h,\gamma}$, R be as in Theorem 1.1. Then the $\tilde{K}_h$ is a $(HL_{h,\gamma}, \gamma/R)$ good kernel in hinge loss with respect to P.*

Proof Let $w = E_{(x',y')}(y'\tilde{\phi}(x')) \in \mathcal{H}_{\tilde{K}_h}$. We have:

$$\begin{aligned}
\epsilon &= E_{(x,y),(x',y')} \left[\!\left[\left[1 - \frac{t_{yy'} h_{xx'}}{\gamma} \right]_+ \right]\!\right] \\
&= E_{(x,y),(x',y')} \left[\!\left[\left[1 - \frac{yy'\tilde{K}(x,x')}{\gamma} \right]_+ \right]\!\right] \\
&= E_{(x,y)} \left[E_{(x',y')} \left[\!\left[\left[1 - \frac{yy'\tilde{K}(x,x')}{\gamma} \right]_+ \Big| (x,y) \right]\!\right] \right] \\
&\qquad \text{(Jensen's inequality)} \\
&\geq E_{(x,y)} \left[\!\left[\left[1 - \frac{E_{(x',y')} \left[\!\left[yy'\langle \tilde{\phi}(x'), \tilde{\phi}(x) \rangle | (x,y) \right]\!\right]}{\gamma} \right]_+ \right]\!\right] \\
&= E_{(x,y)} \left[\!\left[\left[1 - \frac{y\langle w/||w||_{\mathcal{H}}, \tilde{\phi}(x) \rangle}{\gamma/||w||_{\mathcal{H}}} \right]_+ \right]\!\right]
\end{aligned}$$

To conclude the proof, we bound $\|w\|_{\mathcal{H}}$ by R:

$$\begin{aligned}
\|w\|^2_{\mathcal{H}} &= E_{(x,y)} \left[\!\left[y\tilde{\phi}(x) \right]\!\right] \cdot E_{(x',y')} \left[\!\left[y'\tilde{\phi}(x') \right]\!\right] \\
&= E_{(x,y),(x',y')} \left[\!\left[yy'\tilde{K}(x,x') \right]\!\right] \\
&\leq \sqrt{E_{(x,y),(x',y')} \left[\!\left[y^2 y'^2 \right]\!\right] \cdot E_{(x,y),(x',y')} \left[\!\left[\tilde{K}^2(x,x') \right]\!\right]} \\
&= \sqrt{E_{(x,y),(x',y')} \left[\!\left[\tilde{K}^2(x,x') \right]\!\right]} \leq R^2
\end{aligned}$$

Lemma 1.3 *Let K be an (ϵ, γ) good kernel in hinge loss, with $K(x,x) \leq R^2 \ \forall x \in \mathcal{X}$. Let $(x_i, y_i)_{i=1}^n$ be an IID training sample, and $\hat{f}(x) = \hat{w} \cdot \phi(x)$ with*

$$\hat{w} = \underset{||w||_{\mathcal{H}_K} \leq 1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left[1 - \frac{y_i w \cdot \phi(x_i)}{\gamma} \right]_+$$

[1] A kernel that does not satisfy this definition is not necessarily a "bad" kernel. We just can not make any formal statements with respect to its performance.

be a kernel classifier that minimizes the average hinge loss relative to γ on the training sample. Then, with probability at least $1-\delta$, we have:

$$P_{(x,y)}\left[\!\left[y\hat{f}(x)\leq 0\right]\!\right] \leq \epsilon + \mathcal{O}\left(\sqrt{\frac{R^2\ln(1/\delta)}{\gamma^2 n}}\right)$$

Lemma 1.3 follows directly from Theorem 21 in (Bartlett & Mendelson, 2002).

Thus, in the case of learning a linear combination of kernels, with $K_i(x,x)\leq 1$, the following generalization bounds applies:

Corollary 1.4 *Let $h_\mu(z_{xx'}) = \mu\cdot z_{xx'}$ be a K-classifier with $\|\mu\|_2 = 1$. Then, with probability at least $1-\delta$, a classifier $\hat{f}$ with generalization error*

$$P_{(x,y)}\left[\!\left[y\hat{f}(x)\leq 0\right]\!\right] \leq HL_{h_\mu,\gamma} + \mathcal{O}\left(\sqrt{\frac{p\ln(1/\delta)}{\gamma^2 n}}\right)$$

can be learned efficiently from a training sample of n instances drawn IID from P.

Corollary 1.5 *Let $h_\mu(z_{xx'}) = \mu\cdot z_{xx'}$ be a K-classifier with $\|\mu\|_1 = 1$. Then, with probability at least $1-\delta$, a classifier $\hat{f}$ with generalization error*

$$P_{(x,y)}\left[\!\left[y\hat{f}(x)\leq 0\right]\!\right] \leq HL_{h_\mu,\gamma} + \mathcal{O}\left(\sqrt{\frac{\ln(1/\delta)}{\gamma^2 n}}\right)$$

can be learned efficiently from a training sample of n instances drawn IID from P.

Note that, unlike in the one-stage kernel learning case, the generalization bound in Theorem 1.1 is in terms of the expected hinge loss of the K-classifier not the training hinge loss. While we are hopeful a generalization bound for the classification problem in the K-space can be obtained, as of now it remains an open problem.

We can, however, prove a concentration bound for the expected hinge loss of a K-classifier. This is the analog of the concentration bounds for target alignment in (Cortes et al., 2010; Cristianini et al., 2001).[2]

Theorem 1.6 *Let P, h, $HL_{h,\gamma}$, R be as in Theorem 1.1. Let $(x_i,y_i)_{i=1}^n$ be an IID sample distributed according to P. Then the following inequality holds with probability at least $1-\delta$*

$$HL_{h,\gamma} \leq \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\left[1-\frac{t_{ij}h(z_{ij})}{\gamma}\right]_+ + \sqrt{\frac{2\left(1+\frac{R^2}{\gamma}\right)^2\ln 1/\delta}{n}}$$

Proof We will prove the concentration bound using McDiarmid's inequality (**?**). Let

$$f((x_1,y_1),\cdots,(x_n,y_n)) = \\ = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\left[1-\frac{y_iy_j\tilde{K}(x_i,x_j)}{\gamma}\right]_+$$

Let (x_l',y_l') be a new sample drawn at random from P. We have

$$\begin{aligned}
&|f((x_1,y_1),...,(x_l,y_l),...,(x_n,y_n))- \\
&\quad - f((x_1,y_1),...,(x_l',y_l'),...,(x_n,y_n))| \leq \\
&\leq \frac{2}{n(n-1)}\left(\sum_{i=1}^{l-1}\left|\left[1-\frac{y_iy_l\tilde{K}(x_i,x_l)}{\gamma}\right]_+ - \right.\right. \\
&\qquad \left.\left. - \left[1-\frac{y_iy'l\tilde{K}(x_i,x_l')}{\gamma}\right]_+\right|\right) + \\
&\quad + \frac{2}{n(n-1)}\left(\sum_{i=l+1}^{n}\left|\left[1-\frac{y_ly_i\tilde{K}(x_l,x_i)}{\gamma}\right]_+ \right.\right. \\
&\qquad \left.\left. - \left[1-\frac{y_l'y_i\tilde{K}(x_l',x_i)}{\gamma}\right]_+\right|\right) \\
&\leq \frac{2}{n}\left(1+\frac{R^2}{\gamma}\right)
\end{aligned}$$

Where the last inequality comes from the fact that for any (x,y) and (x',y')

$$0 \leq \left[1-\frac{yy'\tilde{K}(x,x')}{\gamma}\right]_+ \leq 1+\frac{R^2}{\gamma}$$

Applying McDiarmid's inequality gives

$$P\left[\!\left[E\left[\!\left[f((x_1,y_1),\cdots,(x_n,y_n))\right]\!\right] - \right.\right. \\ \left.\left. -f((x_1,y_1),\cdots,(x_n,y_n)) \geq \epsilon_1\right]\!\right] \leq \\ \leq \exp\left(\frac{-n\epsilon_1^2}{2\left(1+\frac{R^2}{\gamma}\right)^2}\right)$$

[2] This is not a regular generalization bound as the K-classifier is not allowed to depend on the training sample.

The statement of the theorem is obtained by equating the right side with δ, and observing that for any $i \neq j$

$$E_{(x_i,y_i),(x_j,y_j)} \left[\!\!\left[\left[1 - \frac{t_{ij} h(z_{ij})}{\gamma} \right]_+ \right]\!\!\right] = \\ = E_{(x,y),(x',y')} \left[\!\!\left[\left[1 - \frac{t_{yy'} h(z_{x,x'})}{\gamma} \right]_+ \right]\!\!\right]$$

which implies

$$E\left[\!\left[f((x_1,y_1),\cdots,(x_n,y_n)) \right]\!\right] = \\ = E_{(x,y),(x',y')} \left[\!\!\left[\left[1 - \frac{t_{yy'} h(z_{x,x'})}{\gamma} \right]_+ \right]\!\!\right]$$

References

Bartlett, P. and Mendelson, S. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3, 2002.

Cortes, C., Mohri, M., and Rostamizadeh, A. Two-Stage Learning Kernel Algorithms. In *International Conference on Machine Learning*, 2010.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. S. On Kernel-Target Alignment. In *NIPS*, 2001.

Srebro, N. How Good is a Kernel When Used as a Similarity Measure. In *COLT*, 2007.