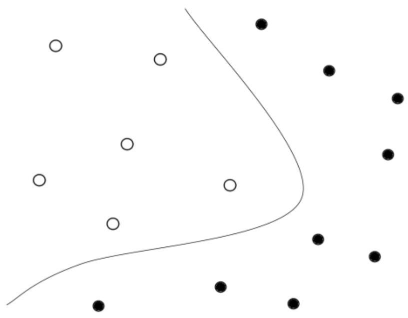
Non-Sparse Multiple Kernel Learning

Marius Kloft
Pavel Laskov
Ulf Brefeld
Sören Sonnenburg

TU berlin
Fraunhofer FIRST Institut Rechnerarchitektur und Softwaretechnik
Eberhard Karls Universität Tübingen

Problem Setting

Binary classification



Given: labels y_i

data x_i

p views on the data, each encoded by a kernel $K_i,\ i = 1, ..., p$.

Some Baseline Approaches

Train a classifier on…

(1) the uniform kernel mixture $K = \sum_{j=1}^{p} \beta_j K_j, \;\; \beta_1 = ... = \beta_p = \frac{1}{p}$

Problems:

arbitrary choice

irrelevant (noise) kernels are considered

(2) a single kernel $K_i, \;\; i \in \{1, ..., p\}$

which is optimal in model selection (e.g. cross-validation)

Problems:

useful information discarded

training time consuming (*p* nested loops)

Multiple Kernel Learning (MKL) Approach

Simultaneously learning a convex combination $K = \sum_{j=1}^{p} \beta_j K_j$,

and a model $f(K)$, such that the expected test error $R[f(K)]$ is minimal in *K*.

[Lanckriet et al., 2004; Bach et al., 2004, Sonnenburg et al., 2006]

Optimization Problem

$$\min_{\beta} \quad \mathrm{svm}\left(\sum_{j=1}^{p} \beta_j K_j\right), \quad \text{s.t.} \quad \boldsymbol{\beta} \geq 0, \boxed{\|\boldsymbol{\beta}\|_1 = 1}$$

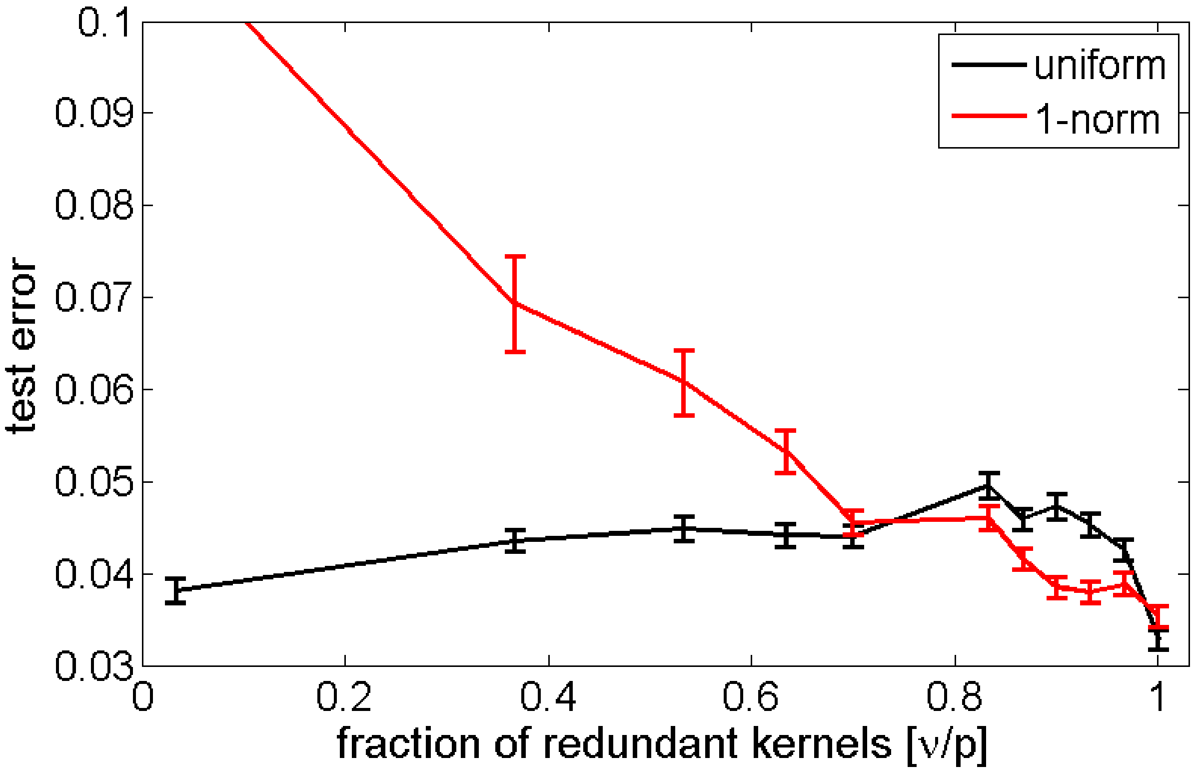
where $$\mathrm{svm}(K) = \max_{\boldsymbol{\alpha}} \quad \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}' D(\mathbf{y}) K D(\mathbf{y}) \boldsymbol{\alpha}$$

$$\text{s.t.} \quad 0 \leq \boldsymbol{\alpha} \leq \eta \,; \quad \mathbf{y}'\boldsymbol{\alpha} = 0$$

$\beta_i = 0$ for most *i*: regular MKL finds a sparse combination of kernels

Problem: kernels often encode complementary properties of the data

Multiple Kernel Learning (MKL) Approach



Problem: kernels often encode complementary properties of the data

Non-Sparse MKL

We have seen: a sparse MKL may be inappropriate.

Remedy: we substitute the $\|\beta\|_1 = 1$ constraint by $\|\boldsymbol{\beta}\|_2 = 1$.

Optimization Problem

$$\min_{\beta} \quad \text{svm}\left(\sum_{j=1}^{p} \beta_j K_j\right), \quad \text{s.t.} \quad \boldsymbol{\beta} \geq 0, \; \|\boldsymbol{\beta}\|_2 = 1$$

$$\text{where} \quad \text{svm}(K) = \max_{\boldsymbol{\alpha}} \quad \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}' D(\mathbf{y}) K D(\mathbf{y}) \boldsymbol{\alpha}$$

$$\text{s.t.} \quad 0 \leq \boldsymbol{\alpha} \leq \eta \; ; \quad \mathbf{y}'\boldsymbol{\alpha} = 0$$

Problem: ℓ_2-norm ruins convexity.

Convex Relaxation

Remedy: we relax the ℓ_2 -norm equality constraint $\|\boldsymbol{\beta}\|_2 = 1$ to $\|\boldsymbol{\beta}\|_2 \leq 1$.

We show:

Theorem *Let* $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ *be optimal points of the relaxed* ℓ_2*-regularized MKL problem and* $K_1, \ldots, K_p$ *be positive definite. Then we have* $\|\boldsymbol{\beta}^*\|_2 = 1$.

Approximation is tight.

Min-Max Problem

Hence we have:

Min-Max problem. Given kernel matrices $K_1, ..., K_p$.

$$\min_{\beta} \quad \text{svm}\left(\sum_{j=1}^{p} \beta_j K_j\right), \quad \text{s.t.} \quad \boldsymbol{\beta} \geq 0, \boxed{\|\boldsymbol{\beta}\|_2 \leq 1}$$

where $\quad \text{svm}(K) = \max_{\boldsymbol{\alpha}} \quad \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}' D(\mathbf{y}) K D(\mathbf{y}) \boldsymbol{\alpha}$

$\text{s.t.} \quad 0 \leq \boldsymbol{\alpha} \leq \eta\,; \quad \mathbf{y}'\boldsymbol{\alpha} = 0$

Optimization of Min-Max Problem by

→ Translation into semi-infinite program (SIP) [Sonnenburg et al., 2006]

SIP

Hence we arrive at:

Optimization problem (SIP). Given kernel matrices $K_1, ..., K_p$.

$$\begin{aligned} \min_{\Theta, \boldsymbol{\beta}} \quad & \Theta \\ s.t. \quad & \Theta \geq \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}' D(\mathbf{y}) \sum_{j=1}^{p} \beta_j K_j D(\mathbf{y}) \boldsymbol{\alpha} \\ & \forall \boldsymbol{\alpha} \in \mathbb{R}^n \quad \textit{with} \quad \boldsymbol{y}'\boldsymbol{\alpha} = 0, \; \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{1} \\ & \|\boldsymbol{\beta}\|_2 \leq 1; \quad \boldsymbol{\beta} \geq \mathbf{0} \; . \end{aligned}$$

Optimization by column generation:

- Step 1: solve SVM(α)
- Step 2: optimize for β: quadratically constrained program (QCP)

Experiment 1: Toy Experiment

Data set

Goal: generation of $p=30$ kernel matrices $K_1,\ldots,K_p$ for different “levels of kernel redundancy“

Process:

generated two $d=120$ dimensional multivariate gaussians

for some values of $1 \leq m \leq 30$, mod$(m,d)=0$,

for $i=1:p$

K_i = random linear transformation of a randomly drawn m-elemental feature subset

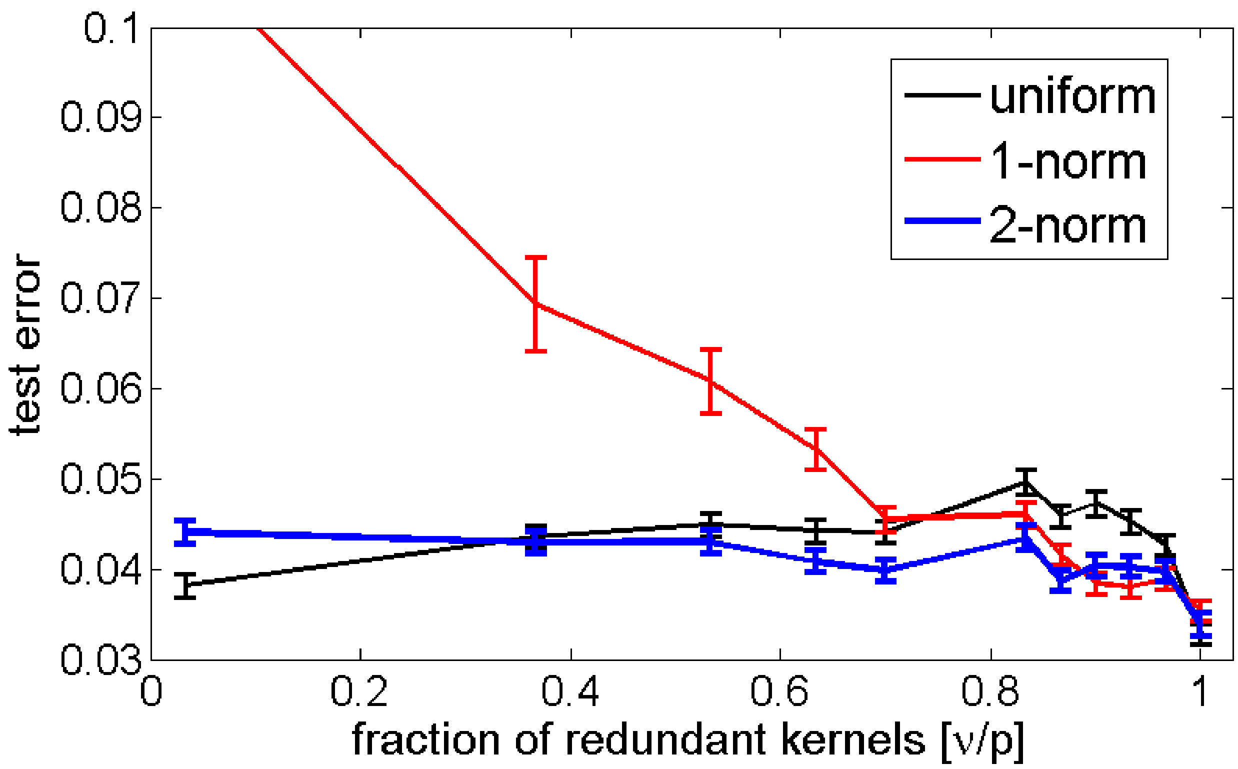
Experimental setup

kernel matrices normed $K_{ij} \to K_{ij}/\sqrt{K_{ii}K_{jj}}$

parameter tuning by grid search on a validation set

100 repetitions

Experiment 1: Results (Toy)



ℓ_2-MKL (blue line) achieves low test errors for most levels of redundancy.

ℓ_2-MKL is never significantly worse than ℓ_1-MKL

Experiment 2: DNA

Prediction of transcription start sites in DNA sequences

[Data available at http://www.fml.tuebingen.mpg.de/raetsch/projects/arts/]

5 domain-specific kernels:

TSS signal:	weighted degree shift kernel on TSS *signal*
promoter:	spectrum kernel on *TSS upstream*
1st exon:	spectrum kernel on *TSS downstream*
energy:	linear kernel on binding stacking *energies*
angles:	linear kernel on *angle* of dinucleotides

Experimental setup:

50K-elemental independent test set

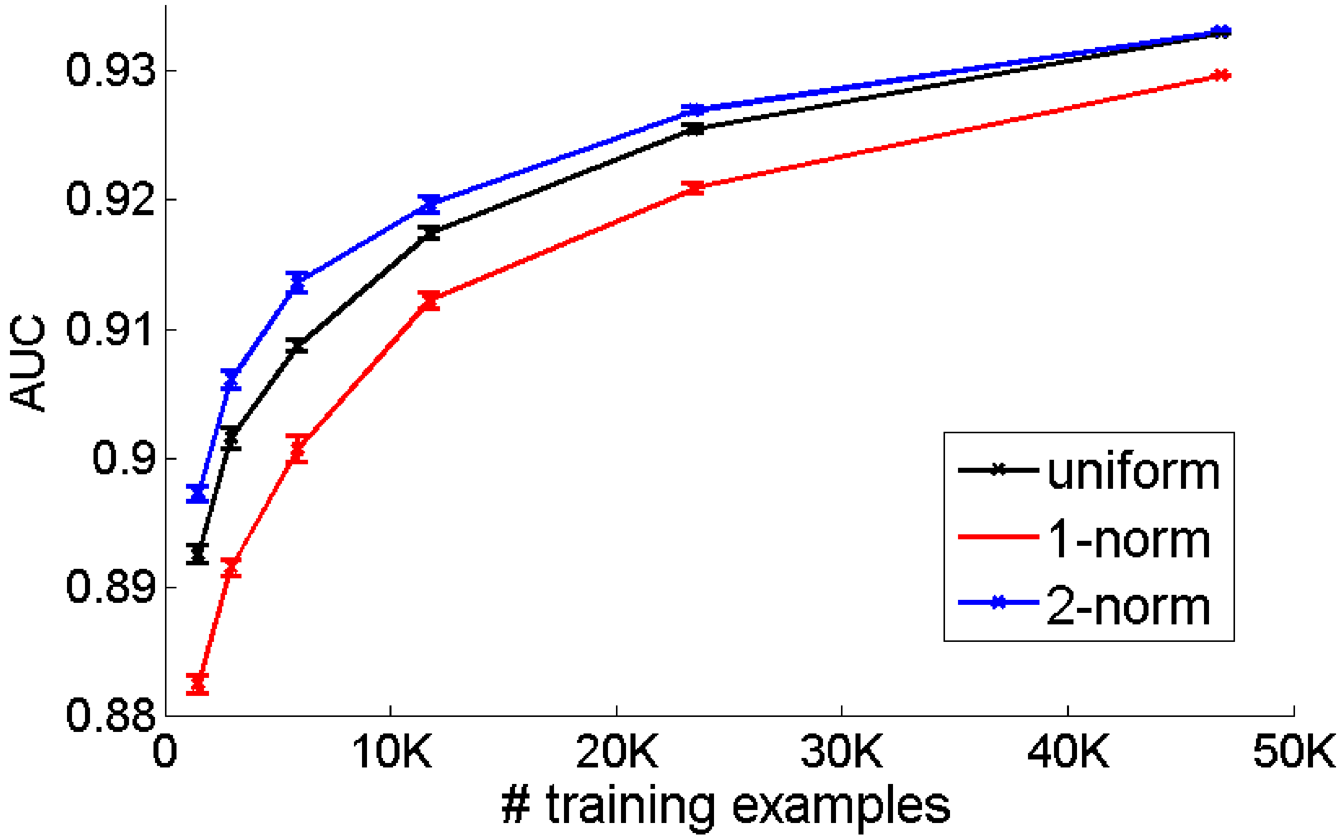
Kernel matrices normalized $K_{ij} \to K_{ij}/\sqrt{K_{ii}K_{jj}}$

SVM soft margin parameter tuning by grid search on a validation set

held out test set

100 repetitions

Experiment 2: Results (DNA)



ℓ_2 -MKL *outperforms* ℓ_1-MKL *and the* uniform *mixture at small and large scales*

Conclusion

Non-sparse multiple kernel learning

- ℓ_2-penalty on the kernel mixture
- problem not convex
- but: tight approximation was shown

Empirical evaluation:

- ℓ_1-MKL was often outperformed by uniform mixture
- ℓ_2-MKL best prediction model in our experiments

If you like to try out yourself…:

- http://www.shogun-toolbox.org/

The End

Thank you!

References

- Bach, F., Lanckriet, G., Jordan, M.: **Multiple Kernel Learning, Conic Duality, and the SMO algorithm.** In: Proceedings of the Twenty-first International Conference (ICML 2004).
- Lanckriet, G., Christianini, N., Bartlett, P., El Ghaoui, L., Jordan, M.: **Learning the Kernel Matrix with Semidefinite Programming.** In: *Journal of Machine Learning Research,* 5(Jan):27--72, 2004.
- Rakotomamonjy, A., Bach, F. R., Canu, S., Grandvalet, Y.: **SimpleMKL.** In: *Journal of Machine Learning Research,* 9(Nov):2491--2521, 2008.
- Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: **Large Scale Multiple Kernel Learning.** In: *Journal of Machine Learning Research,* 7(Jul):1531--1565, 2006.
- Sonnenburg, S., Zien, A., Rätsch, G.: **ARTS: accurate recognition of transcription starts in human.** In: *Bioinformatics,* 22(14):472--e480, 2006.
- Xu, Z., Jin, R., King, I., Lyu, M.: **An Extended Level Set Method for Efficient Multiple Kernel Learning.** *to appear (NIPS).*