Multiple Kernel Learning (*MKL*)

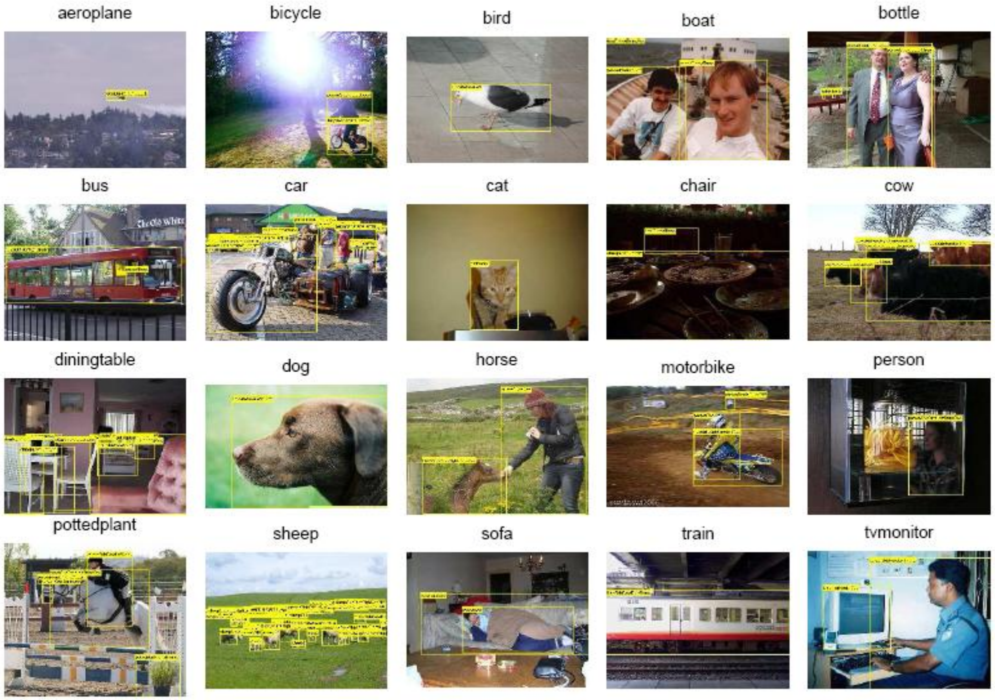
Lecture

Marius Kloft
Klaus-Robert Müller

Motivating example

Multi-label image categorization

- given: a set of images



Goal: assign novel images to correct category

- Specific characteristic of the application: images can be described by various **groups** of features: ***colors, textures, shape information, etc.***

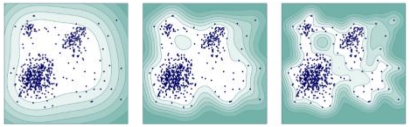
Formal problem setting

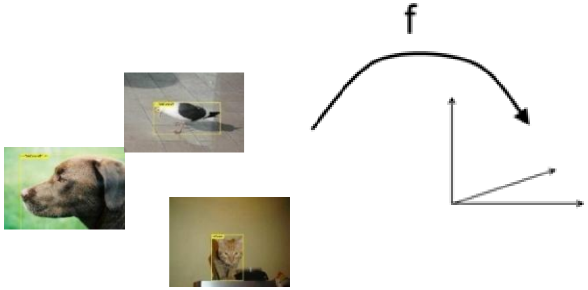
Classification

here: categorization of novel images.

Given: data/label pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ and multiple feature maps,

each giving rise to a kernel $k_j, \ j = 1, ..., m$





Example (image classification):

Color-kernel, texture-kernel, shape-kernel

(in this example, x has a block structure)

Goal:

- learning a classification model $f : \mathcal{X} \rightarrow \mathcal{Y} \stackrel{\text{e.g.}}{=} \{0, 1\}$
- and an optimal kernel mixing $K = \sum_{j=1}^{m} \beta_j K_j, \quad \boldsymbol{\beta} \geq \mathbf{0}$

What is the optimal kernel mixture?

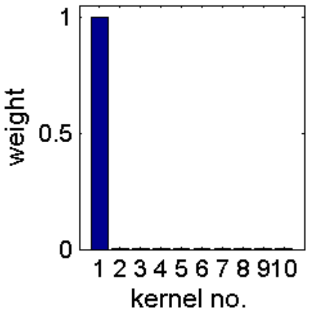
We focus on linear kernel mixtures $K = \sum_{j=1}^{m} \beta_j K_j$ and SVMs.

Heuristic 1: use a single kernel

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

Disadvantage:

useful information discarded.

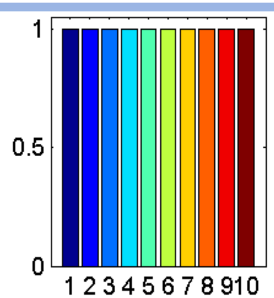


Heuristic 2: the uniform kernel mixture $\beta_1 = \cdots = \beta_m = \frac{1}{m}$

("average kernel")

Disadvantage:

arbitrary choice → *irrelevant kernels considered.*



What is the optimal kernel mixture?

Heuristic 3: Brute-Force: try out all possible mixtures (e.g. grid search)

Infeasible: *computationally too demanding (if #kernels large).*

Can we do better?

➔ Truly integrate feature/kernel selection into the learning machine, e.g., Support Vector Machine (SVM)

(i.e., really learning *the optimal weights!)*

Problem formalization

Problem setting:

Optimally we would like to minimize the expected loss

$$E_{(x,y)\sim P}\Big[\big(l\big(f_K(x), y\big)\Big]$$

with respect to the optimal kernel $K = \sum_j \beta_j K_j$

where $f_K(x) = \langle w, \phi_K(x)\rangle$ and ϕ_K is the feature map corresponding to K

Unfortunately, the underlying probability distribution is unknown

➔ *i.e., no access to the expected risk*

Remedy: instead minimize *empirical* risk

$$\widehat{E}\Big[\big(l\big(f_K(x), y\big)\Big] = \frac{1}{n}\sum_{i=1}^{n}\Big[\big(l\big(f_K(x_i), y_i\big)\Big]$$

Problem relaxation

Key observation:

empirical risk is upper bounded by SVM objective:

$$\widehat{E}\Big[(l\big(f_K(x),y\big)\Big] = \frac{1}{n}\sum_{i=1}^{n}\Big[(l\big(f_K(x_i),y_i\big)\Big] \leq \underbrace{\frac{C}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[(l\big(f_K(x_i),y_i\big)\Big]}_{\text{SVM(K,w)}}$$

MKL idea:

➔ minimize SVM objective also over the kernel mixing β:

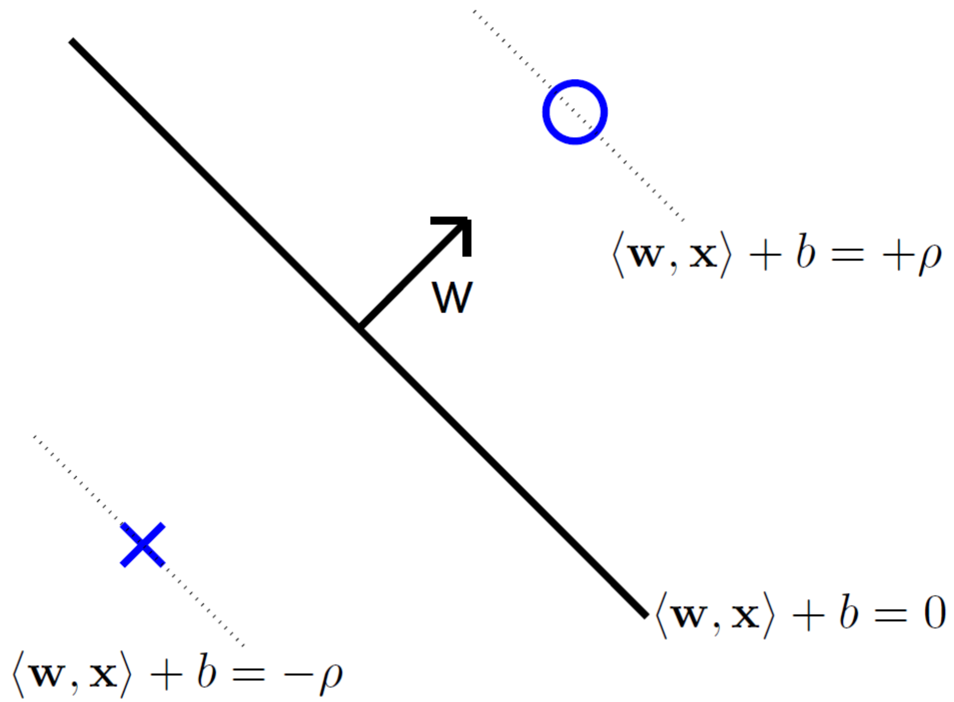
$$\min_{w,\boldsymbol{\beta}} \quad \mathrm{SVM}\Big(\sum_j \beta_j K_j, w\Big)$$

Wait a minute….

[Lanckriet et al., 2004]

Mini-review SVMs

SVM:
maximum margin
classifier



$$\max_{\mathbf{w},b,\rho} \quad \underbrace{\rho}_{\text{margin}} \quad s.t. \quad \underbrace{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq \rho,}_{\text{data fitting}} \quad \underbrace{\|\mathbf{w}\| = 1}_{\text{normalization}}$$

Let's jump back to the previous slide: Problem relaxation

Key observation:

empirical risk is upper bounded by SVM objective:

$$\widehat{E}\Big[\big(l\big(f_K(x),y\big)\Big] = \frac{1}{n}\sum_{i=1}^{n}\Big[\big(l\big(f_K(x_i),y_i\big)\Big] \le \underbrace{\frac{C}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[\big(l\big(f_K(x_i),y_i\big)\Big]}_{\mathrm{SVM(K,w)}}$$

MKL idea:

➔ minimize SVM objective instead:

[Lanckriet et al., 2004]

$$\min_{w,\boldsymbol{\beta}} \quad \mathrm{SVM}\Big(\sum_j \beta_j K_j, w\Big)$$

Problem:

Overfits! ➔ need regularizer on β

(for the same reason that we need one on ***w*** *)*

MKL overfitting issue

Example: consider MKL problem

$$\min_{w,\boldsymbol{\beta}} \quad \frac{C}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[(l\big(f_K(x_i), y_i\big)\Big]$$

$$\text{s.t.} \quad K = \sum_{j=1}^{m} \beta_j K_j \,, \quad f_K(x) = \langle w, \phi_K(x)\rangle$$

and the special case: $m=1$, i.e., $K=\beta_1 K_1$

Then, we can decrease the objective by transferring weight from w to to β_1

➔ Would lead to an unbounded, i.e., infinite β_1

➔ **We do need regularization on β**

Classical MKL approach

➔ ***Impose a 1-norm constraint on β***

SVM objective

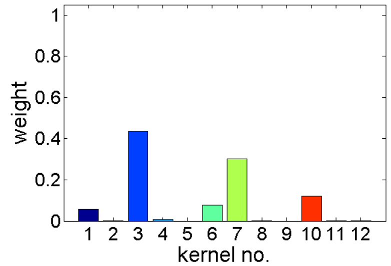
Classical MKL problem: [Lanckriet et al., 2004]

MKL **1-norm** constraint

$$\min_{w,\boldsymbol{\beta}} \quad \frac{C}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[\big(l\big(f_K(x_i), y_i\big)\Big]$$

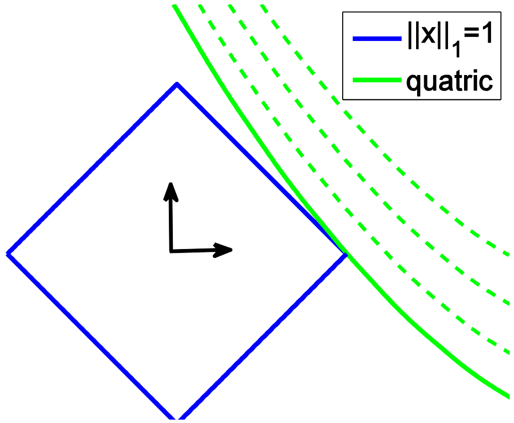
$$\text{s.t.} \quad f_K(x) = \langle w, \phi_K(x)\rangle\,, \quad \underbrace{\sum_{j=1}^{m}\beta_j K_j, \quad \boldsymbol{\beta} \geq \mathbf{0}, \quad \boxed{||\boldsymbol{\beta}||_1 = 1}}_{\text{MKL constraints}}$$

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



Why 1-norm yields sparsity

Illustration: Level sets of a quadratic function (e.g., SVM) intersecting with an 1-norm equality constraint



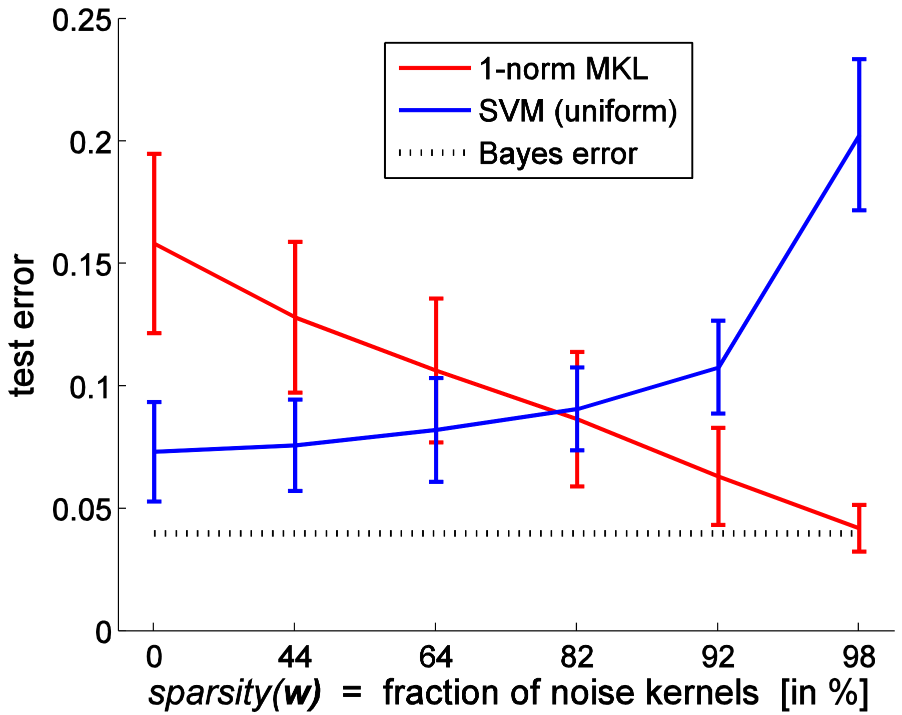
Optimal solution *x* is attained when the functions "hit"

➔ In the above example, $x=(1,0)$

➔ thus *x* is sparse (means: "*x has some zero entries*")

Sparsity leads to MKL failing frequently

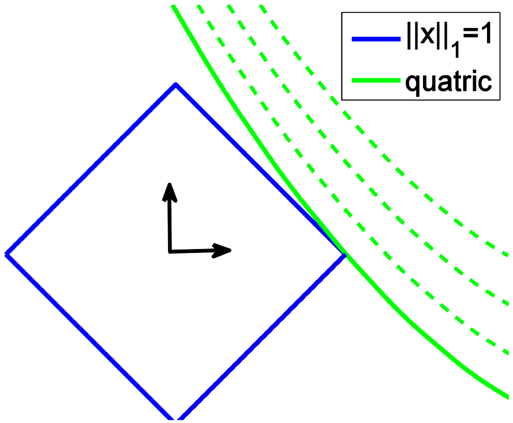
Results of a toy experiment (described later):

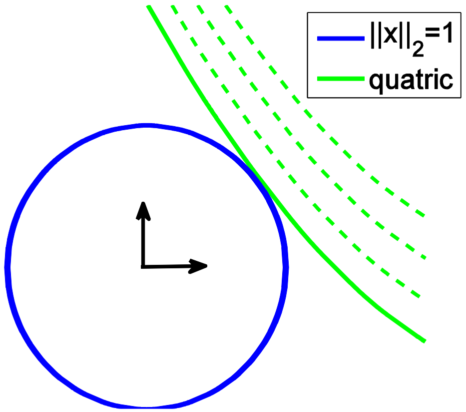


Reason: kernels often encode complementary properties of the data, in contrast to redundant ones

Geometry of the L_p-norm

Illustration: Comparison of the level sets of a quadratic function (SVM) intersecting a 1-norm (*left*) and a 2-norm constraint (*right*)





Optimal solution is attained when the functions "hit"

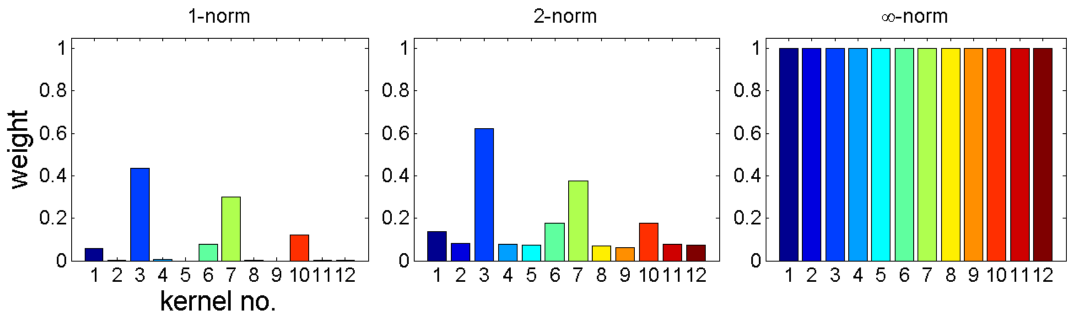
➔ Leads to sparse (*left*) or non-sparse solutions (*right*)

„Desparsification“ of MKL

➔ **take a general p>1-norm constraint $||\boldsymbol{\beta}||_{\boldsymbol{p}} \leq \mathbf{1}$ instead of the sparsity-inducing 1-norm constraint**

Note: the p-norm is defined as $||\beta||_p = \sqrt[p]{\sum_j |\beta_j|^p}$

➔ Exemplary kernel mixtures output by MKL:



➔ Contains the important special cases *classical MKL* *(p=1)* and *uniform-kernel SVM* $(p = \infty)$

So-called „L_p-norm multiple kernel learning“

We finally obtain:

L_p-norm MKL optimization problem:

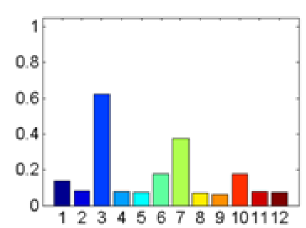
SVM objective

MKL **p-norm** constraint

$$\min_{w,\boldsymbol{\beta}} \quad \frac{C}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[\big(l\big(f_K(x_i), y_i\big)\Big]$$

$$\text{s.t.} \quad f_K(x) = \langle w, \phi_K(x)\rangle\,, \quad \underbrace{K = \sum_{j=1}^{m}\beta_j K_j, \quad \boldsymbol{\beta} \geq \mathbf{0}, \boxed{||\boldsymbol{\beta}||_p \leq 1}}_{\text{MKL constraints}}$$

➔ yields non-sparse $\boldsymbol{\beta} > 0$



[Kloft et al., 2009/2011]

MKL optimization algorithm

L_p-norm MKL Optimization Problem

$$\min_{\boxed{w,\boldsymbol{\beta}}} \frac{C}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[\big(l\big(f_K(x_i), y_i\big)\Big]$$

$$\text{s.t.} \quad K = \sum_{j=1}^{m} \beta_j K_j, \quad \boldsymbol{\beta} \geq \mathbf{0}, \ ||\boldsymbol{\beta}||_p \leq 1$$

***Alternating* optimization:**

repeat

w-step: consider β as constant and optimize with respect to w

(boils down to solving an SVM)

β-step: consider w as constant and optimize with respect to β

(Can be done analytically: $\beta_j^* = \dfrac{\|\boldsymbol{w}_j\|^{\frac{2}{p+1}}}{\sqrt[p]{\sum_l \|\boldsymbol{w}_l\|^{\frac{2p}{p+1}}}}$ *)*

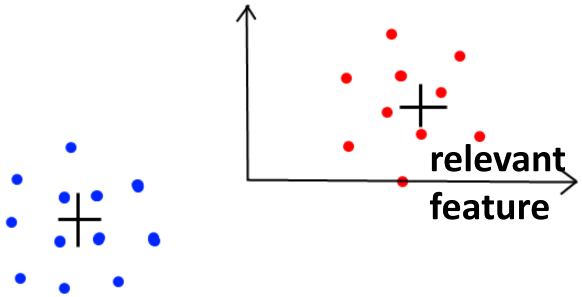
until converged

proof is left as an excercise at the acutal worksheet

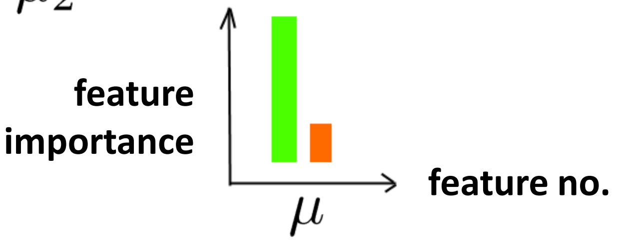
Experiment 1: Toy experiment

Data set

- sampled binary labeled data
- from two isotrop gaussian distributions with opposing means μ_1 and $\mu_2 = -\mu_1$



- alignment of mean vectors $w_{\text{Bayes}} = \mu_1 - \mu_2$ controls feature importance:
- Generated 6 data scenarios with varying alignments of mean vectors



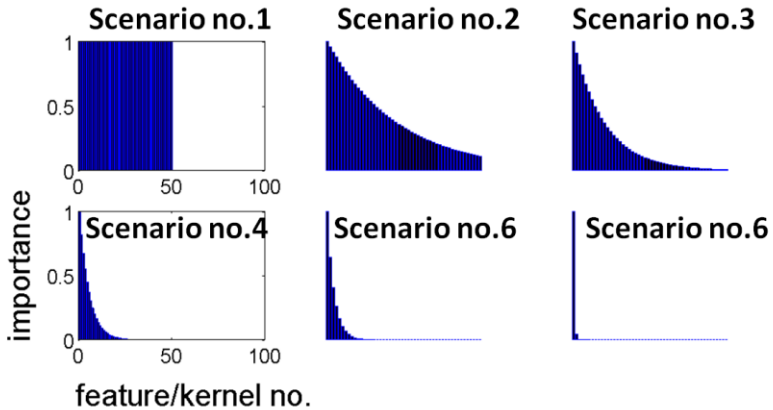
Experimental setup

- 50-dimensional data; one feature per kernel ; disjoint
- randomly drawn distinct training, tuning, and test sets (n=50/5000/10000)
- 250 repetitions, model selection

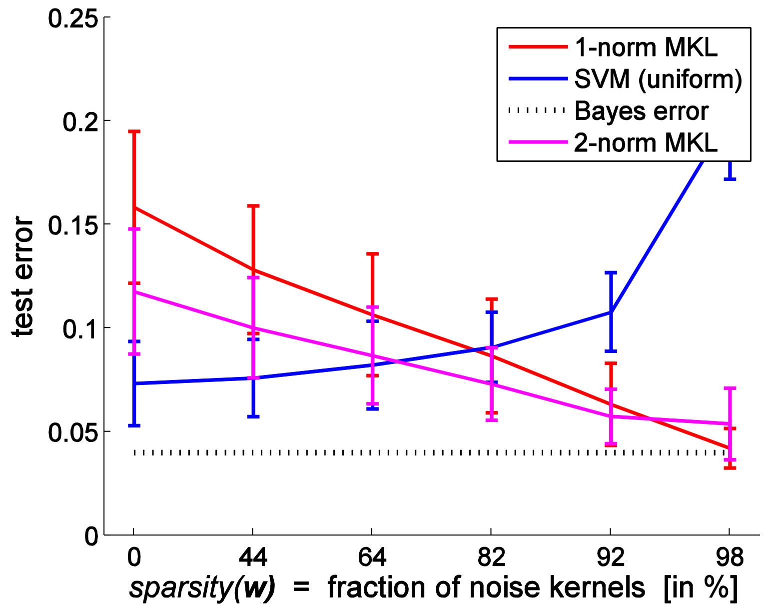
Experiment 1 (Toy): Empirical results

Scenarios:

Scatter plot of Bayes-w



Test error:



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

ℓ_2-MKL is outperforms ℓ_1 -MKL in almost all scenarios

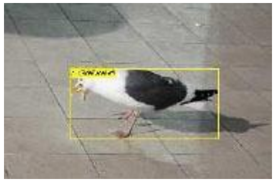
Experiment 2: Multi-label image categorization

[Binder et al.]

Data set

- VOC 2008 challenge data set:
  - 8780 images
  - 20 categories (*aeroplane, bird*, *dog*, …).

bird



dog

Feature extraction

- employed 12 domain-specific kernels (variation: 30 kernels)
  - based on several combination of basic features:
    e.g. histogram of visual words, color sets, pyramid level tilings
- all kernels are normalized.

Experimental setup

- train a binary model for each category (one-vs.-rest)
- randomly drawn distinct training, tuning, and test sets
- 10 repetitions, model selection.

Image categorization experiment: empirical results (AP)

	average	aeroplane	bicycle	bird	boat	bottle	bus	car
1-norm	40.8±0.9	**66.9±6.8**	36.4±6.6	44.1±5.7	56.8±5.0	**19.2±3.8**	39.3±10.6	49.0±2.8
∞-norm	40.8±1.0	66.4±6.6	39.1±5.9	43.3±5.8	**57.5±5.0**	18.4±3.6	**42.3±9.1**	48.9±3.3
p-single	**42.3±0.9**	**67.1±6.3**	**40.7±6.6**	44.7±5.4	**57.8±5.4**	**19.5±3.6**	**41.7±9.5**	**50.3±3.4**
p-joint	**42.6±0.7**	**67.6±6.0**	**41.6±6.8**	**45.3±5.7**	**58.4±5.6**	**19.4±3.8**	**44.5±9.9**	**50.6±2.9**
p selected		1.1562	1.1250	1.2500	1.2500	1.2500	2.0000	1.2500

		cat	chair	cow	diningtable	dog	horse	motorbike
1-norm		47.7±3.8	**44.1±4.9**	**10.8±3.5**	27.1±7.0	**34.4±4.4**	39.6±5.8	41.7±4.5
∞-norm		46.1±3.2	43.0±4.7	**8.2±2.9**	**29.5±9.1**	**33.2±2.7**	**42.5±6.5**	42.8±2.9
p-single		**48.9±3.7**	**44.9±3.8**	**10.3±3.1**	**30.1±6.2**	**34.0±3.4**	**42.0±6.6**	**44.7±4.2**
p-joint		**49.6±3.0**	**45.3±3.9**	**9.8±2.8**	**30.7±8.3**	**34.0±3.8**	**43.2±7.1**	**44.3±3.5**
p selected		1.1250	1.0312	1.0938	1.3125	1.3125	1.3750	1.2500

		person	pottedplant	sheep	sofa	train	tvmonitor	
1-norm		84.1±1.3	**14.7±4.5**	**26.3±7.7**	**33.0±7.2**	50.9±9.8	50.8±5.5	
∞-norm		83.9±1.2	**15.5±4.3**	22.9±7.0	31.3±6.5	50.9±9.2	**51.0±5.7**	
p-single		**84.5±1.2**	**16.1±4.7**	**27.5±7.6**	**33.9±6.9**	**53.7±9.9**	**52.9±5.7**	
p-joint		**84.5±1.3**	**15.7±5.0**	**27.6±7.5**	**34.0±8.5**	**54.1±10.1**	**52.2±5.2**	
p selected		1.1250	1.2500	1.0312	1.3125	1.2500	1.1250	

p-single = *p* optimized class-wise
p-joint = one joint *p* for all classes

ℓ_p-MKL *outperforms* ℓ_1-MKL *and the* uniform *mixture (*ℓ_∞*-norm MKL)*

ℓ_p-MKL *best prediction model for* ***all*** *classes*

Conclusion

Multiple kernel learning

simultaneously learning a model and kernel weightings

incorporate optimization over the weights into the (SVM) optimization problem

use p-norm regularizer to avoid overfitting

1-norm yields sparsity, p>1 yields non-sparsity

classical MKL uses 1-norm

empirically, p-norm MKL often better than classical 1-norm MKL

Free implementation…:

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

References

- Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R., and Zien, A.: **Efficient and accurate lp-norm multiple kernel learning.** In: *Advances in Neural Information Processing Systems, 2009.*

- Kloft, M., Brefeld, U., Sonnenburg, S., Laskov, P., Müller, K.-R., and Zien, A.: **lp-norm multiple kernel learning.** In: *JMLR, 2011.*

- Lanckriet, G., Christianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M.: **Learning the kernel matrix with demidefinite programming.** In: *Journal of Machine Learning Research,* 5(Jan):27-72, 2004.

- Nakajima, S., Binder, A., Müller, C., Wojcikiewicz, S., Kloft, M., Brefeld, U., Müller, K.-R., and Kawanabe, M.: **Classifying visual objects with multiple kernels.** In: *IBIS 2009.*

- Sonnenburg, S., Rätsch, G., Schäfer, C., and Schölkopf, B.: **Large scale multiple kernel learning.** In: *Journal of Machine Learning Research,* 7(Jul):1531-1565, 2006.