Machine Learning & Our Work

Haiqin Yang

Department of Computer Science & Engineering
The Chinese University of Hong Kong

Feb. 22, 2010

Outline

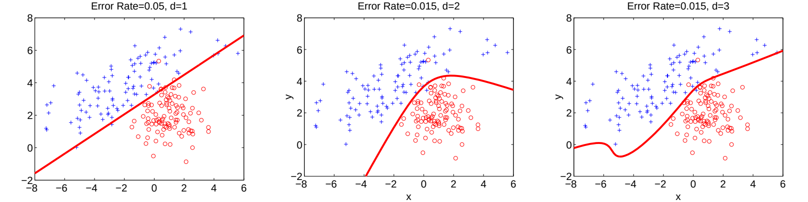
1. Introduction
   - Supervised learning
   - Support vector machines
   - L_1-norm regularization
2. Our Work
   - Summary
   - Sparse generalized multiple kernel learning

Classification

Setup

- $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$,
  $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_i \in \{-1, 1\}$
- **Objective:** seek $f_{\vartheta}(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$, $\vartheta = (\mathbf{w}, b)$,
  to classify $\mathbf{x}$ into -1 or $+1$

Illustration

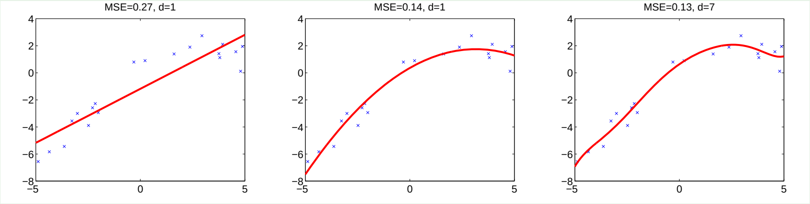


Regression

Setup

- $\mathcal{L} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{L}$,
  $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$, $y_i \in \mathbb{R}$
- **Objective:** seek $f_\vartheta(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$, $\boldsymbol{\vartheta} = (\mathbf{w}, b)$,
  to make $f_\vartheta(\mathbf{x}) \approx y_i$

Illustration



Tikhonov regularization–ridge regression

History and definition

- ✓ Developed by Andrey Tychonoff in 1940's
- ✓ The most commonly used method of regularization of ill-posed problems
- ✓ In statistics, named ridge regression

Definition:

$$\min_{\mathbf{w}} \underbrace{\|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2}_{\text{loss}} + \underbrace{\|\Gamma \mathbf{w}\|^2}_{\text{Regularizer}}$$

Γ is the Tikhonov matrix, usually $\Gamma = \mathbf{I}$.

Support vector classification

History and definition

✓ Theories mainly developed by Vapnik in 1970's

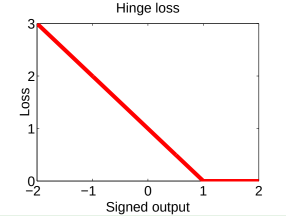
✓ First introduced in COLT 1992, by Boser, Guyon, Vapnik

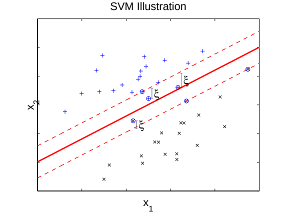
Definition:

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{L} H_1(y_i f(\mathbf{x}_i)) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$H_1(z) = \max\{0, 1 - z\} : \text{ hinge loss}$$

Illustration





Support vector regression

History and definition

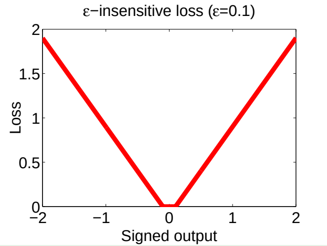
✓ First introduced in NIPS 1996, by H. Drucker, et al. (1997)

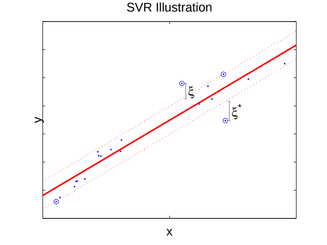
Definition:

$$\min_{\mathbf{w}} \quad \sum_{i=1}^{L} I_{\varepsilon}(y_i - f(\mathbf{x}_i)) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$I_{\varepsilon}(z) = \max\{0, |z| - \varepsilon\} : \ \varepsilon\text{-insensitive loss}$$

Illustrations





Lasso

History and definition

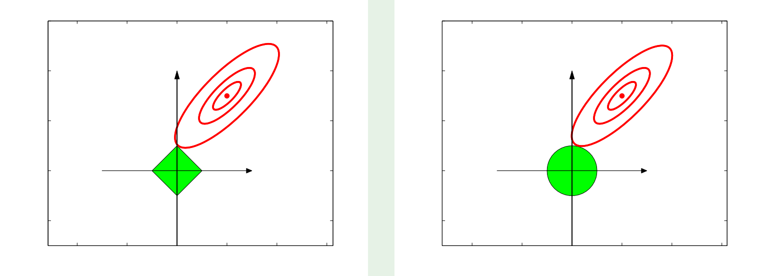
✓ Find a least-squares solution with the L_1-regularizer

✓ Mainly developed by R. Tibshirani (1996)

Definition:

$$\min_{\mathbf{w}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \lambda\|\mathbf{w}\|_1$$

Illustrations



Group lasso

History and definition

✓ Do variable selection in a group manner

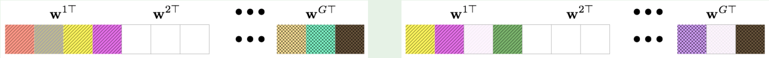
✓ First proposed by Yuan, M. and Lin, Y. (2006)

Definition:

Group Lasso: $$\min_{\mathbf{w}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \lambda \sum_{g=1}^{G} \sqrt{d_g} \|\mathbf{w}^g\|_2$$

Sparse Group Lasso: $$\min_{\mathbf{w}} \quad \|\mathbf{X}\mathbf{w} - \mathbf{Y}\|^2 + \lambda \sum_{g=1}^{G} (\sqrt{d_g} \|\mathbf{w}^g\|_2 + r_g \|\mathbf{w}^g\|_1)$$

Illustrations



(a) Group Lasso

(b) Sparse Group Lasso

Citations

SVM

V. Vapnik (1995) The nature of statistical learning theory.	16257
V. Vapnik (1998) Statistical learning theory.	253
N. Cristianini and J Shawe-Taylor (2000) An introduction to support vector machines...	6691
C. Burges (1998) A tutorial on support vector machines for pattern recognition	6618
A. Smola and B. Schölkopf (2004) A tutorial on support vector regression	1669
B. Schölkopf and A. Smola (2002) Learning with kernels	5429
C. Chang and C. Lin (2001) LIBSVM: a library for support vector machines.	2753
T. Joachims (1999) SVMLight: support vector machine library.	112

Lasso

R. Tibshirani (1996) Regression shrinkage and selection via the lasso.	2489
B. Efron, T. Hastie, I. Johnstone and R. Tibshirani (2004) Least angle regression.	1246
M. Yuan and Y. Lin (2006) Model selection and estimation in regression with grouped variables	307

Optimization

Y. Nesterov and A. Nemirovskii (1987) Interior-point polynomial algorithms in convex programming.	1313
L. Vandenberghe and S. P. Boyd (1996) Semidefinite programming	1726
S. P. Boyd and L. Vandenberghe (2004) Convex optimization	6173

Lists

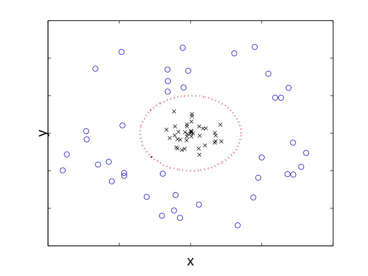
- Localized support vector regression
- Multi-task learning models
- Tri-class support vector machines
- Sparse generalized multiple kernel learning method
- Online learning models
    - Group Lasso
    - Multi-task learning models
    - . . .

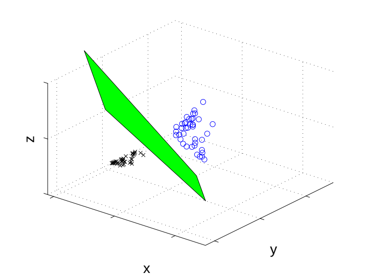
Background

SVM–nonlinear extension

Data: $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$
Decision: $f(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}) + b, \qquad \phi(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^f$

Illustration





Kernelized version

Objective: $\max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad \mathbf{1}_N^\top\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top\mathbf{K}(\boldsymbol{\alpha}\circ\mathbf{y})$

$\mathcal{A} = \{\boldsymbol{\alpha}\in\mathbb{R}_+^N,\ \boldsymbol{\alpha}^\top\mathbf{y}=0,\ \boldsymbol{\alpha}\le C\mathbf{1}_N\}$

Decision: $f(\mathbf{x}) = \sum_{i=1}^N \alpha_i^\star\mathbf{K}(\mathbf{x},\mathbf{x}_i) + b^\star,$

Kernels

Definition: $k(\mathbf{x}_1,\mathbf{x}_2) = \phi(\mathbf{x}_1)^\top\phi(\mathbf{x}_2)$

Polynomial $k(\mathbf{x}_1,\mathbf{x}_2) = (\mathbf{x}_1^\top\mathbf{x}_2+1)^d$

RBF $k(\mathbf{x}_1,\mathbf{x}_2) = \exp(-\gamma\|\mathbf{x}_1-\mathbf{x}_2\|^2)$

$\ldots \quad \ldots$

Construct Kernels

$$\begin{bmatrix} \cdot & \cdots & \cdot \\ \vdots & \mathbf{K}_1 & \vdots \\ \cdot & \cdots & \cdot \end{bmatrix} \quad \cdots \quad \cdots \begin{bmatrix} \cdot & \cdots & \cdot \\ \vdots & \mathbf{K}_q & \vdots \\ \cdot & \cdots & \cdot \end{bmatrix} \quad \cdots \quad \begin{bmatrix} \cdot & \cdots & \cdot \\ \vdots & \mathbf{K}_Q & \vdots \\ \cdot & \cdots & \cdot \end{bmatrix}$$

How to select optimal kernel?

Cross-validation or learn from data based on some criteria

L_1-norm MKL

Formulation

Objective: $\min_{\hat{\mathbf{w}},b,\boldsymbol{\theta}\geq\mathbf{0}} \; C\sum_{i=1}^{N} R(f_{\hat{\mathbf{w}},b,\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \frac{1}{2}\hat{\mathbf{w}}^\top\hat{\mathbf{w}} + \lambda\mathcal{J}(\boldsymbol{\theta})$,

Dual: $\min_{\boldsymbol{\theta}\in\Theta}\max_{\boldsymbol{\alpha}\in\mathcal{A}} \; \mathcal{D}(\boldsymbol{\theta},\boldsymbol{\alpha}) = \mathbf{1}_N^\top\boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top\left(\sum_{q=1}^{Q}\theta_q\mathbf{K}_q\right)(\boldsymbol{\alpha}\circ\mathbf{y})$

$\Theta = \{\boldsymbol{\theta}\in\mathbb{R}_+^Q : \|\boldsymbol{\theta}\|_1 \leq 1\}$

$\mathcal{A} = \{\boldsymbol{\alpha}\in\mathbb{R}_+^N, \; \boldsymbol{\alpha}^\top\mathbf{y} = 0, \; \boldsymbol{\alpha}\leq C\mathbf{1}_N\}$

Decision: $f(\mathbf{x}) = \sum_{i=1}^{N}\alpha_i^\star\left(\sum_{q=1}^{Q}\theta_q^\star\mathbf{K}_q(\mathbf{x},\mathbf{x}_i)\right) + b^\star$,

Research on this framework

Speed-up methods: Semi-definite programming
Semi-infinite linear programming
Gradient descent
Extended level method

Model extensions: L_2/L_p-norm MKL
Mixed norms

Our generalized MKL

Motivations

- ✓ L_1-norm MKL may discard useful information when kernels are orthogonal or with correlation characterizations
- ✓ L_p-norm MKL yields non-sparse solutions for $p > 1$

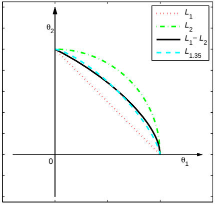
Formulation

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\boldsymbol{\alpha}\in\mathcal{A}} \quad \mathbf{1}_N^\top \boldsymbol{\alpha} - \frac{1}{2}(\boldsymbol{\alpha}\circ\mathbf{y})^\top \left(\sum_{q=1}^{Q} \theta_q \mathbf{K}_q\right)(\boldsymbol{\alpha}\circ\mathbf{y})$$

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}_+^Q : v\|\boldsymbol{\theta}\|_1 + (1-v)\|\boldsymbol{\theta}\|_p \leq 1\}$$

$$\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}_+^N,\ \boldsymbol{\alpha}^\top\mathbf{y} = 0,\ \boldsymbol{\alpha} \leq C\mathbf{1}_N\}$$

Here, we consider $p = 2$



Properties

- $v\|\boldsymbol{\theta}^\star\|_1 + (1-v)\|\boldsymbol{\theta}^\star\|_2^2 \Leftrightarrow 1$
- For $\mathbf{K}_i = \mathbf{K}_j$,

  $v \neq 1 \quad \theta_q^\star = \max\left\{0, \frac{1}{2(1-v)}\left(\frac{1}{\lambda}(\boldsymbol{\alpha} \circ \mathbf{y})^\top \mathbf{K}_q (\boldsymbol{\alpha} \circ \mathbf{y}) - v\right)\right\}$

  $v = 1 \quad \theta_i$ and θ_j are not unique
- $\frac{(\boldsymbol{\alpha}^\star \circ \mathbf{y})^\top \mathbf{K}_i (\boldsymbol{\alpha}^\star \circ \mathbf{y})}{(\boldsymbol{\alpha}^\star \circ \mathbf{y})^\top \mathbf{K}_j (\boldsymbol{\alpha}^\star \circ \mathbf{y})} \rightarrow 1 \Rightarrow \theta_i^\star \rightarrow \theta_j^\star$

Algorithm-level method

Given: predefined tolerant error $\delta > 0$
Initialization: Let $t = 0$ and $\boldsymbol{\theta}^0 = c\mathbf{1}_q$,
Repeat
Solve the dual problem of an SVM with $\sum_{q=1}^{Q} \theta_q^t \mathbf{K}_q$ to get $\boldsymbol{\alpha}$;
Construct the cutting plane model, $h^t(\boldsymbol{\theta}) = \max_{1 \leq i \leq t} \mathcal{D}(\boldsymbol{\theta}, \boldsymbol{\alpha}^i)$;
Calculate the lower bound and the upper bound of the cutting plane and the gap, Δ^t;
Project $\boldsymbol{\theta}^t$ onto the level set by solving a QCQP;
Update $t = t + 1$;
until $\Delta^t \leq \delta$.

- The convergence rate of the level method is $\mathcal{O}(\delta^{-2})$
- `DualGap`

$$=\mathcal{D}(\boldsymbol{\theta}^t,\boldsymbol{\alpha}^t)-\mathbf{1}_N^\top\boldsymbol{\alpha}^t+\max_q \varpi_q$$

$$\varpi_q=\left(\tfrac{1}{2}(\boldsymbol{\alpha}^t\circ\mathbf{y})^\top\mathbf{K}_q(\boldsymbol{\alpha}^t\circ\mathbf{y})\varsigma\right)$$

$$\varsigma=\frac{(1-v)\sum_{q=1}^{Q}(\theta_q^t)^2+1}{v+2(1-v)\theta_q^t}$$

Experiments I

Algorithms

- SimpleMKL for L_1-norm MKL
- L_2-norm MKL
- GMKL

Platform

- Mosek to solve the QCQP
- Matlab on a a PC with Intel Core 2 Duo 2.13GHz CPU and 3GB memory.

Experiments II

Datasets

Dataset	# Classes	# Training (N)	# Test	# Dim	# Kernel (Q)
Toy1	2	150	150	20	273
Toy2	2	150	150	20	273
Breast	2	341	342	10	143
Heart	2	135	135	13	182
Ionosphere	2	175	176	33	442
Liver	2	172	173	6	91
Pima	2	384	384	8	117
Sonar	2	104	104	60	793
Wdbc	2	284	285	30	403
Wpbc	2	99	99	33	442
Colon	2	31	31	2,000	2,000
Lymphoma	2	48	48	4,026	4,026
Plant	4	470	470		69
Psort+	4	270	271		69
Psort-	5	722	722		69

Experiments III

Schemes on generating toy data

- **Toy1**

$$Y_i = \text{sign}\left(\sum_{j=1}^{3} f_1(x_{ij}) + \epsilon_i\right)$$

- **Toy2**

$$Y_i = \text{sign}\left(\sum_{j=1}^{3} f_1(x_{ij}) + \sum_{j=4}^{6} f_2(x_{ij}) + \sum_{j=7}^{9} f_3(x_{ij}) + \sum_{j=10}^{12} f_4(x_{ij}) + \epsilon_i\right)$$

- The outputs (labels) are dominated by only some features
- Each mapping acts on three features equally, implicitly incorporating grouping effect
- Each mapping is with zero mean on the corresponding feature, which yields zero mean on the output

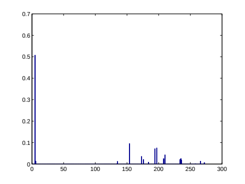
Experimental results I

Toy data results

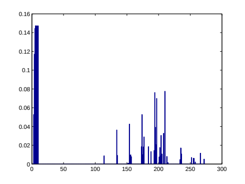
Dataset	Method	Accuracy	# Kernel	Times (s)
Toy 1	GMKL	**71.6**±1.2	43.0±3.3	2.8±0.7
	L_1-MKL	67.3±1.1	20.5±2.1	4.2±0.9
	L_2-MKL	69.2±1.0	273	2.6±1.0
Toy 2	GMKL	**76.5**±1.2	48.5±3.3	3.6±0.2
	L_1-MKL	73.1±2.4	25.3±2.5	6.7±2.4
	L_2-MKL	74.2±1.8	273	3.3±0.3

Experimental results II

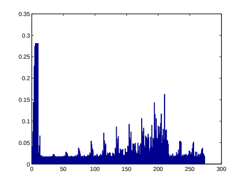
Selected kernels on toy data



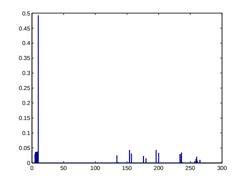
L_1-MKL on Toy 1



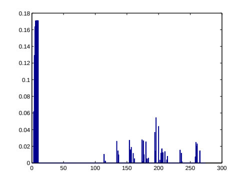
GMKL on Toy 1



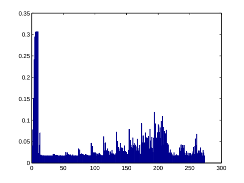
L_2-MKL on Toy 1



L_1-MKL on Toy 2



GMKL on Toy 2



L_2-MKL on Toy 2

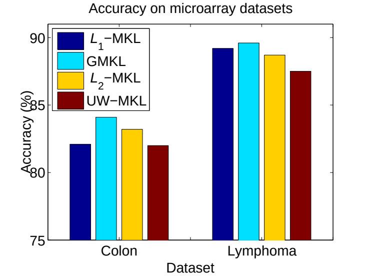
Experimental results III

Results on UCI data

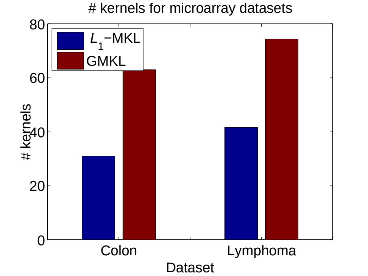
Dataset	Method	Accuracy	# Kernel	Times (s)
Breast	GMKL	†**97.3**±0.3	49.7±2.1	5.1±0.3
	L_1-MKL	96.8±0.8	14.3±3.5	36.1±4.1
	L_2-MKL	97.0±0.6	143	8.7±0.5
Heart	GMKL	**84.6**±0.6	40.5±3.5	1.6±0.4
	L_1-MKL	**84.6**±1.2	28.0±4.8	3.4±0.3
	L_2-MKL	**84.6**±0.7	182	2.9±0.2
Ionosphere	GMKL	92.4±1.1	64.7±2.5	7.3±1.0
	L_1-MKL	92.0±2.6	35.0±3.6	14.0±2.3
	L_2-MKL	**93.3**±1.0	442	6.6±0.5
Liver	GMKL	†**68.6**±2.0	30.3±2.2	1.2±0.2
	L_1-MKL	65.4±4.9	11.0±2.6	2.7±0.7
	L_2-MKL	†**68.6**±2.5	91	2.3±0.2
Pima	GMKL	†**79.4**±0.5	80.5±7.8	3.1±0.4
	L_1-MKL	77.5±0.9	17.7±1.2	47.0±7.9
	L_2-MKL	77.3±0.7	117	11.8±0.7
Sonar	GMKL	†**84.3**±2.8	80.0±7	19.3±0.8
	L_1-MKL	79.6±7.6	64.3±9.1	9.7±2.3
	L_2-MKL	81.1±5.7	793	6.0±0.2
Wdbc	GMKL	96.6±0.2	76.5±4.5	10.8±0.7
	L_1-MKL	96.5±1.2	18±1.0	54.5±0.4
	L_2-MKL	**96.7**±0.7	403	17.7±1.8
Wpbc	GMKL	**77.7**±2.0	379.0±60.1	1.7±0.4
	L_1-MKL	77.1±2.1	45.0±8.2	4.2±0.9
	L_2-MKL	**77.7**±2.3	442	2.5±0.7

Experimental results IV

Results on microarray data



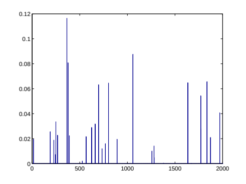
Accuracy



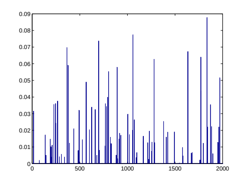
No. of selected kernels

Experimental results V

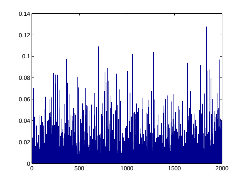
Selected kernels on microarray data



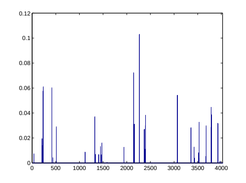
L_1-MKL on Colon



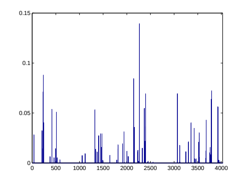
GMKL on Colon



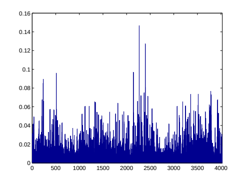
L_2-MKL on Colon



L_1-MKL on Lymphoma



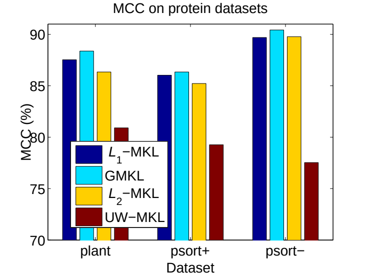
GMKL on Lymphoma



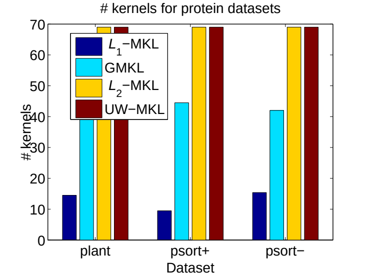
L_2-MKL on Lymphoma

Experimental results VI

Results on protein subcellular localization data



Accuracy



No. of selected kernels

Summary

- A generalized multiple kernel learning (GMKL) model by imposing L_1-norm and L_2-norm regularization on the kernel weights
- Properties, e.g., sparse solutions, are discussed
- Model is solved by the level method, convergence rate and optimal conditions are provided.
- Experiments on both synthetic and real-world datasets are provided.

Questions ?